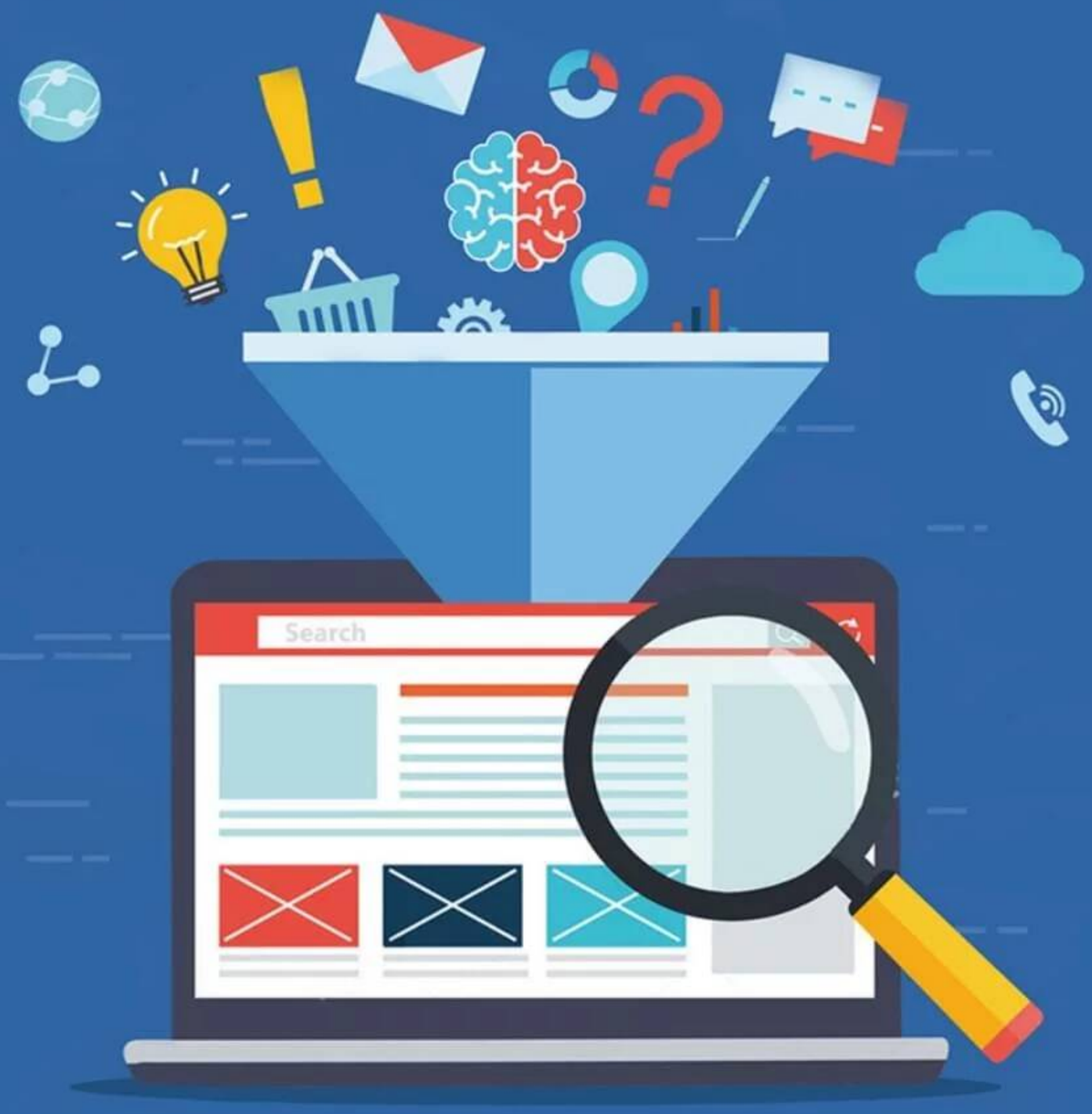
NTICS

tecnología

INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS

UNA VISIÓN GLOBAL DE BUSINESS INTELLIGENCE & ANALYTICS

Luis Joyanes Aguilar



Alfaomega

Descargado en: eybooks.com

INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS

Una visión global de Business Intelligence & Analytics

Luis Joyanes Aguilar

INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS

Una visión global de Business Intelligence & Analytics

Luis Joyanes Aguilar

Alfaomega

Buenos Aires • Bogotá • Ciudad de México • Santiago de Chile

ALFAOMEGA

Empresas del Grupo

Colombia: Alfaomega Colombiana S.A.
Calle 62 No.20-46 esquina, Bogotá
Teléfono (57-1) 746 0102 Fax: (57-1) 210 0122
cliente@alfaomegacolombiana.com

México: Alfaomega Grupo Editor S.A. de C.V.
Calle Doctor Olvera No. 74, Colonia Doctores,
Delegación Cuauhtemoc, Ciudad de México
C.P. 06720 • teléfono (52-55) 5089 7740
Fax (52-55) 5575 2420
Sin costo 01-800-020-4396
libreriapitagoras@alfaomega.com.mx

Argentina: Alfaomega Grupo Editor Argentino S.A.
Av. Córdoba 1215, Piso 10
Capital Federal, Buenos Aires
Teléfono/Fax: (54-11) 4811 7183 / 8352 / 0887
ventas@alfaomegaeditor.com.ar

Chile: Alfaomega Grupo Editor S.A.
Av. Providencia 1443. Oficina 24, Santiago
Teléfonos (56-2) 2235 4248 / 2947 9351 / 2235 5786
agechile@alfaomega.cl

www.alfaomega.com.co

Inteligencia de negocios y analítica de datos. Una visión global de Business Intelligence & Analytics
Bogotá, 2019

© Luis Joyanes Aguilar
© Alfaomega Grupo Editor, México

Todos los derechos son reservados. Esta publicación no puede ser reproducida total ni parcialmente. No puede ser registrada por un sistema de recuperación de información, en ninguna forma ni por ningún medio, sea mecánico, fotoquímico, electrónico, magnético, electroóptico, fotocopia o cualquier otro, sin el permiso previo y por escrito de la editorial.
d e s c a r g a do e n: e y b o oks. c o m .

Edición: Damián Fernández
Revisión de estilo: Silvia Mellino
Diseño de la cubierta: Melina Daffunchio

ISBN: 978-958-778-541-8

Hecho en Colombia
Printed and made in Colombia

A mis queridas nietas, “mis niñas”, Olivia e Inés con el inmenso cariño que les profeso y su recuerdo que me acompaña en todo momento.

Y a mi hermana Juana Mary y mis sobrinos Raquel y Roberto que siempre están a mi lado y siempre cuento con su apoyo.

Contenido

PARTE I

VISIÓN MODERNA DE INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS

CAPÍTULO 1

INTELIGENCIA DE NEGOCIOS. Una panorámica global 1
1.1 Introducción 2
1.2 Inteligencia de negocios: Historia, definiciones y conceptos 3
1.3 Business Intelligence, Business Analytics y Big Data: Los tres pilares de la inteligencia empresarial 7
1.4 Arquitectura de un sistema de inteligencia de negocios 8
1.5 Introducción a Big Data y su impacto en la inteligencia de negocios 18
1.6 Arquitectura de inteligencia de negocios con integración de Big Data 22
1.7 Visión gerencial de inteligencia de negocios 28
1.8 Analítica de negocios (business analytics) 31
1.9 Inteligencia de negocios en Big Data .. 35
1.10 Inteligencia de negocios móvil 38
1.11 Inteligencia de negocios en la nube.. 38
1.12 Proveedores de inteligencia de negocios: Cuadrante mágico de Gartner de BI & Analytics 42
1.13 Inteligencia de negocios futura: Integración de Big Data, Internet de las Cosas e Inteligencia Artificial 46
1.14 La evolución hacia la Inteligencia de negocios en la nube (Cloud BI) 47
1.15 RESUMEN 49

CAPÍTULO 2

ANALÍTICA DE NEGOCIOS (BUSINESS ANALYTICS): UNA VISIÓN GLOBAL 53
2.1 Introducción 54
2.2 Conceptos básicos de analítica de negocios (business analytics) 55
2.3 Business Analytics versus Data Analytics 57
2.4 Analítica avanzada (AA) 62
2.5 Caso de estudio: Cuadrante mágico de Gartner de BI & Analytics 64
2.6 Organización, tipos y fuentes de datos 69
2.7 Ciclo de vida de los datos 72
2.8 Analítica de datos: conceptos y tipos 77
2.9 Big Data Analytics 84
2.10 Ciencia de datos: Evolución de la analítica de negocios y el análisis de datos 86
2.11 Tendencias de Analytics 91
2.12 RESUMEN 93

CAPÍTULO 3

TRANSFORMACIÓN DIGITAL EN ORGANIZACIONES Y EMPRESAS (ECONOMÍA COLABORATIVA, EXPERIENCIA DE CLIENTE Y BLOCKCHAIN) 97
3.1 Introducción 98
3.2 ¿Qué es Transformación Digital? 99
3.3 Tecnologías facilitadoras de la Transformación Digital 101
3.4 La empresa digital 105
3.5 La Transformación Digital en la industria y en la empresa 107
3.6 El proceso de Transformación Digital . 113
3.7 Fábrica inteligente: la Transformación Digital en la Industria 4.0 114

3.8 Economía Colaborativa 116
3.9 Experiencia de Cliente 121
3.10 Blockchain (cadena de bloques) 124
3.11 Blockchain en Inteligencia de Negocios 128
3.12 RESUMEN 130

PARTE II
INFRAESTRUCTURAS Y ARQUITECTURA DE INTELIGENCIA DE NEGOCIOS

CAPÍTULO 4
ALMACENES DE DATOS: DATA WAREHOUSE, OLAP Y DATA LAKE 135
4.1 Introducción 136
4.2 Datos: gestión, gobierno, calidad e integridad 136
4.3 Administración de archivos 143
4.4 Bases de datos 145
4.5 Data Warehouse 147
4.6 Data Mart 151
4.7 Marco de trabajo (framework) de un sistema de almacenamiento de datos 153
4.8 Metadatos, calidad y gobierno de un Data Warehouse 160
4.9 Herramientas ETL 162
4.10 Desarrollo de un sistema de Data Warehouse 164
4.11 Enfoques de desarrollo (modelos) de un sistema de Data Warehouse 165
4.12 OLAP (Procesamiento analítico en línea) 168
413 Data Lakes (Lagos de Datos): Los nuevos depósitos de almacenamiento de datos 173
4.14 *Data Lake* versus Data Warehouse 177
4.15 Proveedores de soluciones de Data Warehouse 180
4.16 RESUMEN 186

Contenido disponible online

CAPÍTULO 5
BIG DATA: ARQUITECTURA, ECOSISTEMA HADOOP Y OPEN DATA) 187

CAPÍTULO 6
BASES DE DATOS NOSQL Y "EN MEMORIA" 189

CAPÍTULO 7
VISUALIZACIÓN DE DATOS: INFORMES Y CONSULTAS, CUADROS DE MANDO (DASHBOARDS) Y CUADRO DE MANDO INTEGRAL (CMI) 191
7.1 Introducción 192
7.2 Conceptos generales de visualización de datos 193
7.3 Gráficos 194
7.4 Tipos de gráficos 196
7.5 Mapas 196
7.6 Infografías 201
7.7 Informes (reporting) y consultas (query) 204
7.8 Cuadros de mando (dashboards) 207
7.9 Narrativa de Datos (Data Storytelling) 216
7.10 Cuadro de Mando Integral (CMI) o Balanced Scorecard 219
7.11 Herramientas de visualización de datos 220
7.12 RESUMEN 221

PARTE III
ANALÍTICA DE NEGOCIOS Y ANALÍTICA DE DATOS

CAPÍTULO 8
MINERÍA DE DATOS 225
8.1 Introducción 226
8.2 Minería de Datos: conceptos, definiciones y aplicaciones 227
8.3 Aplicaciones de la Minería de Datos 228
8.4 Proceso de descubrimiento del conocimiento: KDD 232
8.5 Proceso de Minería de Datos: metodología CRISP-DM 237
8.6 Proceso de Minería de Datos: metodología SEMMA 245

8.7 Modelos, algoritmos y técnicas de Minería de Datos 247
8.8 Relaciones de la Minería de Datos con otras disciplinas: de Big Data a Data Science 248
8.9 Herramientas de software de Minería de Datos 250
8.10 RESUMEN 256

CAPÍTULO 9
MINERÍA WEB Y MINERÍA DE TEXTOS 269
9.1 Introducción 270
9.2 Minería de Textos 271
9.3 Herramientas de la Minería de Textos 272
9.4 Minería Web: conceptos, definiciones y categorías 276
9.5 Arquitectura de la Minería Web 278
9.6 Categorías de la Minería Web 281
9.7 Minería Web de Contenido 283
9.8 Minería Web de la Estructura 283
9.9 Minería Web de Uso 287
9.10 Herramientas de Minería Web 289
9.11 Motores de búsqueda (buscadores) 290
9.12 Posicionamiento SEO: Optimización de los motores de búsqueda 295
9.13 Posicionamiento SEM 299
9.14 RESUMEN 300

CAPÍTULO 10
ANALÍTICA DE DATOS (BIG DATA & ANALYTICS) 303
10.1 Introducción 304
10.2 ¿Qué es Analítica de Datos? (Data Analytics) 305
10.3 Analítica de Negocios (Business Analytics/Analytics) 307
10.4 Una visión global de Analítica de Big Data 308
10.5 Categorías prácticas de Analítica .. 310
10.6 Analítica de Big Data 311
10.7 Características de una plataforma de integración de Analítica de Big Data. ...314
10.8 Analítica Digital 315
10.9 Analítica Web 316
10.10 Proliferación de datos sociales 319
10.11 Analítica Social 321
10.12 Análisis de Sentimientos 322
10.13 Analítica Móvil 325
10.14 RESUMEN 329

CAPÍTULO 11
ANALÍTICA WEB Y ANALÍTICA SOCIAL 333
11.1 Introducción 334
11.2 Primeras consideraciones empresariales sobre analítica web 336
11.3 Breve historia de la Analítica Web 337
11.4 Métricas 338
11.5 Indicadores clave de rendimiento (KPI) 344
11.6 Informes (Google Analytics) 346
11.7 Herramientas de Analítica Web 348
11.8 Analítica Web Móvil (Mobile Analytics) 351
11.9 Analítica Social 353
11.10 Herramientas de Analítica Social 357
11.11 Herramientas de monitorización ... 361
11.12 Herramientas de reputación e influencia social 366
11.13 RESUMEN 374

Contenido disponible online

CAPÍTULO 12
GESTIÓN DEL CONOCIMIENTO Y HERRAMIENTAS COLABORATIVAS 377

PARTE IV

LA INTELIGENCENCIA DE NEGOCIOS EN LA CUARTA REVOLUCIÓN INDUSTRIAL

CAPÍTULO 13
INTELIGENCIA ARTIFICIAL APLICADA Y ALGORITMOS EN INTELIGENCIA DE NEGOCIOS 379
13.1 Introducción 380
13.2 Inteligencia Artificial: Definición, historia y evolución 381

13.3 Tecnologías de Inteligencia Artificial 385
13.4 Aprendizaje automático 388
13.5 Aprendizaje profundo (Deep learning) 389
13.6 Computación cognitiva 394
13.7 Bots y chatbots 397
13.8 Chatbots de empresa: el caso de la atención al cliente 402
13.9 El algoritmo en inteligencia artificial como modelo de negocio en la economía digital..... 406
13.10 RESUMEN..... 414

CAPÍTULO 14
CIENCIA DE DATOS Y CIENTÍFICOS DE DATOS EN INTELIGENCIA DE NEGOCIOS 417
14.1 Introducción 418
14.2 Definición de Ciencia de Datos 418
14.3 Disciplinas de Ciencias de Datos .. 423
14.4 El proceso de Ciencia de Datos..... 426
14.5 El científico de datos..... 428
14.6 El perfil del científico de datos..... 430
14.7 Herramientas de programación para Ciencia de Datos 432
14.8 Roles profesionales relacionados con datos 435
14.9 La Ciencia de Datos en la Inteligencia de Negocios 438
14.10 RESUMEN..... 441

CAPÍTULO 15.
TENDENCIAS DE FUTURO EN INTELIGENCIA DE NEGOCIOS. PRIVACIDAD, PROTECCIÓN Y SEGURIDAD DE LOS DATO (Parte1) 445
15.1 Introducción 446
15.2 Inteligencia de Negocios en la nube: tendencias 447
15.3 Medidas de seguridad en el ciclo de vida de los datos..... 448
15.4 Los riesgos a la privacidad en la Inteligencia de Negocios 449
15.5 Ética y responsabilidad social de las empresas 451
15.6 El nuevo reglamento de protección de datos y de privacidad de la Unión Europea..... 453
15.7 Revisión general de tendencias futuras en Inteligencia de Negocios 459

Contenido disponible online

CAPÍTULO 15
TENDENCIAS DE FUTURO EN INTELIGENCIA DE NEGOCIOS. PRIVACIDAD, PROTECCIÓN Y SEGURIDAD DE LOS DATO (Parte2)

BIBLIOGRAFÍA Y RECURSOS

Prólogo

INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS
Una visión global de *Business Intelligence & Analytics*

Inteligencia de negocios (*Business Intelligence*) es una disciplina muy antigua y que ha ido evolucionando con el tiempo y adaptándose a la evolución de las tecnologías de la información y las comunicaciones más disruptivas, y su implantación en la empresa, con los años, así como las tendencias empresariales más innovadoras. **Analítica de negocios** (BA, *Business Analytics* o *Analytics*, términos cada día más utilizados en consultoría y en estrategias de negocios) es una disciplina complementaria y subconjunto de inteligencia de negocios que se apoya en técnicas de **análisis de datos.**

El término *Business Intelligence,* fue acuñado en 1958 por el investigador de IBM Hans Peter Luhn que publicó el artículo "*A Business Intelligence System*" y que lo definía como: "La habilidad de aprender las relaciones de hechos presentados de forma que guíen las acciones hacia una meta deseada". Inteligencia de negocios se vio potenciada en el año 1962 con la aparición del concepto de OLAP (procesamiento analítico en línea) acuñado por el canadiense Kenneth Iverson y que supuso un importante avance en la analítica de datos. Otro hito importante en la administración de datos fue la creación del concepto de bases de datos en 1969 y que se asentó en la década de los setenta y el desarrollo teórico y práctico de tan importante disciplina. En los años 80 apreció otro concepto soporte del almacenamiento de datos junto con las bases de datos, "*Data Warehouse*" (almacenes de datos).

Fue en 1989 cuando Howard Dresden, investigador de la consultora Gartner, hizo una de las primeras definiciones y más conocida de inteligencia de negocios: "Conceptos y métodos para mejorar las decisiones de negocio mediante el uso de sistemas de soporte basadas en hechos".

Analítica de negocios (BA) es una evolución de la inteligencia de negocios con la que se encuentra estrechamente relacionada y que consideraremos como una disciplina integrada en ella. En 2009, Michael J. Beller en su publicación "*Next Generation Business Analytics*" definía analítica de negocios como "los conocimientos, tecnologías y prácticas para la investigación y exploración continuamente interactiva del rendimiento del negocio para ganar visión y capacidad de dirección en la planificación del negocio". Business Intelligence (conocida en los últimos años, simplemente como Analytics). Analítica de negocios es un proceso asistido por tecnologías mediante el cual, el software analiza los datos para predecir lo que sucederá (análisis predictivo) o lo que podría suceder tomando un cierto enfoque (analítica prescriptiva). El análisis de datos se completa con otros dos tipos de análisis: descriptivo y de diagnóstico, ambos asociados directamente a la inteligencia de negocios tradicional.

Las herramientas de inteligencia de negocios acceden y analizan conjuntos de datos y presentan hallazgos analíticos en informes (reportes), resúmenes,

consultas (queries), gráficos, mapas, infografías, cuadros de mando (dashboards)... para proporcionar a los usuarios información detallada sobre el estado del negocio.

En los últimos años se han desplegado las metodologías y tecnologías de *Big Data*, por el crecimiento exponencial de datos presentes en las organizaciones y empresas. La era de los grandes volúmenes de datos *(Big Data*), su tratamiento, su explotación y la conversión de datos en conocimiento para una toma de decisiones efectiva. Las empresas han de obtener valor de la información. Así han aparecido las nuevas tendencias de analítica de *Big Data* como un proceso de examen de los grandes volúmenes de datos para descubrir patrones ocultos, correlaciones desconocidas y otra información de interés que se pueden utilizar para tomar mejores decisiones.

El mercado de inteligencia de negocios y analítica de datos pasaran a ser la tendencia principal del sector tecnológico, creciendo más rápidamente que cualquier otro ámbito del ecosistema de tecnologías de la información, aunque su elevado coste terminará por limitar su velocidad de expansión. Las modernas plataformas de inteligencia de negocios y analítica de datos han surgido para satisfacer los nuevos requerimientos organizacionales de accesibilidad, agilidad y una visión analítica más profunda. Estas plataformas de BI modernas se apoyan —esencialmente— en tecnologías de inteligencia artificial, aprendizaje automático y aprendizaje profundo, ciencia de datos, procesamiento del lenguaje natural y tecnologías conversacionales de voz (como bots, chatbots...) junto al análisis de los grandes volúmenes de datos (*Big Data*).

La citada consultora Gartner distingue en la actualidad dos tipos de inteligencia de negocios: 1. *Inteligencia de Negocios tradicional* o "*clásica*", donde los profesionales de BI utilizan datos transaccionales internos para generar informes; 2. *Inteligencia de Negocios moderna*, donde los usuarios empresariales interactúan con sistemas ágiles e intuitivos para analizar datos con mayor rapidez. Las organizaciones suelen utilizar las herramientas modernas de inteligencia de negocios cuando los usuarios de negocio necesitan tener una visión global de las dinámicas que cambian rápidamente en los que se valora obtener los datos con gran precisión y exactitud.

En nuestra obra pretendemos analizar las tecnologías y técnicas de inteligencia de negocios, analítica de negocios o analítica (*analytics*) y analítica de datos, tanto tradicionales como modernas,

¿A QUIEN VA DIRIGIDA ESTA OBRA?

La experiencia de muchos años impartiendo la asignatura de ***Inteligencia de Negocios*** en carreras de **Ingeniería de Organización Industrial** e **Ingeniería Informática**, así como numerosos cursos profesionales, seminarios, conferencias, talleres... unida al estudio continuo de las materias que componen un programa innovador y actualizado de la materia, nos llevó ya hace varias años a pensar en la redacción de un libro cuyo contenido pudiera contemplar, también, conocimientos incluidos en los programas clásicos de asignaturas similares a

Inteligencia de Negocios como *Sistemas de Información, Gestión del Conocimiento, Sistemas Informáticos, Administración de Empresas*, etc.

Dado que la inteligencia de negocios es una disciplina inmersa en la estrategia de las empresas y su infraestructura y arquitectura de sistemas de inteligencia de negocios están embebidas en toda la organización de, prácticamente, todas las organizaciones y empresas, hemos intentado, a la vez, escribir un libro profesional que pudiera ser empleado para la introducción en los conceptos fundamentales de inteligencia de negocios y analítica de negocios, tales como tecnologías de almacenamiento de datos —Data Warehouse, Data Mart, bases de datos NoSQL, "en memoria"... — , analítica de datos, minería de datos —herramientas clave para la toma de decisiones— , visualización de datos, analítica Web, etc. Así mismo hemos querido incluir las nuevas tendencias requeridas en las empresas como *Big Data,* analítica de *Big Data,* los nuevos sistemas de almacenamiento de datos como las lagunas de datos (*Data Lakes*), las tendencias de transformación digital y la evolución hacia la ciencia de datos; todas estas tendencias se soportan en las nuevas tecnologías de inteligencia artificial aplicada, como los chatbots o asistentes virtuales, analítica social, etc.

De igual modo hemos intentados llegar a profesionales y directivos de empresas interesados en las actuales y futuras materias que componen las diferentes materias de la inteligencia de negocios tradicional y la denominada inteligencia de negocios moderna como señalan los informes y estudios de las consultoras más prestigiosas como Gartner, Forrester, McKinsey o IDC y las consultoras y auditoras más reputadas como Accenture, PriceWaterhouseCooper, Deloitte, Indra, CapGemini, etc.

Como libro de texto que es, pretende incluir los programas de asignaturas clásicas de ***Inteligencia de Negocios y de Analítica de Datos*** en universidades, institutos tecnológicos, institutos politécnicos, institutos de formación profesional en carreras **de *Administración y Direcciones de Empresa, Económica, Mercadotecnia (Marketing*)...** y las diferentes ***Ingenierías (Sistemas, Informática, Industriales, Organización Industrial, Telecomunicaciones....)*** cuyos programas de estudio contemplan los conocimientos tecnológicos soporte de los diferentes componentes de los sistemas de inteligencia de negocios.

ORGANIZACIÓN DE LA OBRA

El contenido del libro se ha organizado considerando los conocimientos necesarios que consideramos necesarios que entendemos deben tener los técnicos consultores y directivos de inteligencia de negocios en las corporaciones, así como los profesionales y directivos empresariales que necesitan conocer y utilizar herramientas de software tradicionales y modernas de inteligencia de negocios empresariales.

Con el objetivo principal de conseguir alcanzar este amplio rango de conocimientos, el libro se ha organizados en cuatro partes y quince capítulos.

PARTE I. VISIÓN MODERNA DE INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS

El capítulo 1, **Inteligencia** *de negocios y analítica de datos: Una visión global,* se centra en la descripción de la arquitectura de inteligencia de negocios tradicional y un avance a la inteligencia de negocios moderna como así comienzan a denominar las grandes consultoras tecnológicas y de negocios, integradas en el ámbito de las tecnologías de *Big Data.* Así mismo se realiza una introducción a los diferentes sistemas de inteligencia de negocios: móvil, en la nube y de *Big Data.*

Los proveedores de soluciones de software de inteligencia de negocios tanto propietarias como de código abierto (*open source*) constituyen el soporte práctico en que se han de apoyar las corporaciones para implementar herramientas en las estrategias empresariales. En el capítulo se realiza una introducción al estudio "Cuadrante Mágico de Gartner de *Business Intelligence* y plataformas de *Analíticas* de 2017" donde se destacan las empresas comerciales proveedoras de las citadas soluciones más acreditadas y reconocidas por la citada consultora.

El capítulo 2, *Analítica de negocios (Business Analytics),* describe los conceptos fundamentales de la analítica de negocios (*Business Analytics*) centrada en las técnicas de análisis de datos. Se realiza una comparación entre *Business Analytics* (conocida simplemente como "*Analytics*") y analítica de datos, así como una introducción a *Big Data Analytics* (analítica de *Big Data*) y *Data Science* (Ciencia de Datos) componentes fundamentales de la Inteligencia de Negocios Moderna.

El capítulo 3, *Transformación digital en organizaciones y empresas: tendencias tecnológicas y de negocios (economía colaborativa, experiencia de usuario y blockchain),* es la estrategia fundamental de las empresas para su conversión en empresas digitales. El proceso de transformación digital es una necesidad vital que requiere la implantación de las tecnologías disruptiva de la tendencia *Industria 4.0* desencadenante de la cuarta revolución industrial. La economía digital ya implantada en numerosas corporaciones se apoya en una de las emergentes subdisciplinas, *economía colaborativa* que se describe en el capítulo, junto con la importante tendencia experiencia de cliente soporte de los sistemas de información CRM, ERP, GIS, etc.

PARTE II. INFRAESTRUCTURAS Y ARQUITECTURA DE INTELIGENCIA DE NEGOCIOS

La segunda parte se centra en describir las características fundamentales de la infraestructura y arquitectura de inteligencia de negocios.

En el capítulo 4, **Almacenes de datos***: Data Warehouse, OLAP y lagos de datos (Data Lake),* se describen los almacenes de datos o repositorios de datos, componente fundamental de los sistemas de inteligencia de negocios. Los almacenes de datos esenciales de un sistema de IN son los *Data Warehouses, Data Marts* y, en la actualidad como componentes emergentes, los lagos de datos (*Data Lakes*) y que juntos con los sistemas modernos de *Big Data*, constituyen los repositorios fundamentales para almacenar los datos. En el capítulo se describen

también las técnicas de procesamiento analítico de datos (**OLAP**) una de las herramientas más antiguas de analítica de datos y que todavía son de gran utilización en los sistemas de negocios actuales.

En el capítulo 5, *Introducción a Big Data: Arquitectura, Ecosistema Hadoop y Open Data,* se realiza una descripción de las técnicas fundamentales (imprescindibles) para manejar o gestionar los grandes volúmenes de datos existentes en organizaciones y empresas. En el capítulo se analizan los diferentes tipos de datos, fuentes de datos y características de *Big Data*, junto a la arquitectura de *Big Data* y sus herramientas de infraestructuras más populares como Hadoop o Spark.

En el capítulo 6, *Bases de datos NoSQL y "en memoria",* se examinan los componentes técnicos fundamentales de los repositorios de datos (estructurados, no estructurados y semiestructurados): bases de datos analíticas, NoSQL y "en memoria" (*in-memory*).

Una de las técnicas más necesarias e imprescindibles en los sistemas de inteligencia de negocios, son las de visualización. En el capítulo 7, *Visualización de datos: Informes y consultas, cuadros de mando (dashboards) y cuadro de mando integral (CMI),* se describen las herramientas y técnicas de visualización más empleadas: gráficos, tablas, mapas, infografías, cuadros de mando o tableros de control (*dashboards*), cuadros de mando integral (CMI) y una introducción a la técnica complementaria de descubrimiento de datos.

PARTE III. ANALÍTICA DE NEGOCIOS Y ANALÍTICA DE DATOS

El capítulo 8, *Minería de datos*, se centra en los fundamentos de minería de datos y sus aplicaciones más usuales. El proceso de descubrimiento de conocimiento de datos, **KDD** (Knowledge Data Discovery) es un sistema clave en inteligencia de negocios y la minería de datos es la etapa más importante del proceso KDD cuyos componentes fundamentales se describen en el capítulo. Se realiza una introducción de las herramientas más populares de minería de datos.

El capítulo 9, *Minería Web, minería de textos, minería de opinión y de sentimientos,* se centra en la minería web, una categoría de minería de datos centrada en datos de la Web y en la minería de textos. Se describen las tres categorías fundamentales de minería Web: contenido, estructura y uso. Una de las aplicaciones más importantes de la minería web y de textos son los motores de búsqueda (buscadores), su soporte y las técnicas de optimización de los buscadores SEO y SEM, son motivos de estudio del capítulo.

La analítica de *Big Data* como se introdujo en el capítulo 1 es una de las técnicas fundamentales que se deben implementar en las empresas. En el capítulo 10, *Analítica de Big Data (Big Data Analytics),* se hace una introducción a los diferentes tipos de analítica web, móvil, social y de sentimientos. Se describen también los conceptos fundamentales de métricas y KPI (indicadores clave de rendimiento o desempeño).

El capítulo 11, *Analítica Web y Analítica Social,* se centra en descubrir las técnicas clave de analítica web y analítica social, junto con la descripción de las

herramientas más utilizadas en las empresas en el análisis de datos junto con las herramientas más utilizadas puras de analítica junto con herramientas de monitorización, reputación e influencia social.

La gestión *del conocimiento y herramientas colaborativas* son conceptos y herramientas tradicionales de los sistemas de información, componentes esenciales integrados en los sistemas de inteligencia de negocios. En el capítulo 12 se describen los soportes teóricos y técnicos de los sistemas de gestión del conocimiento y sistemas colaborativos.

PARTE IV. LA INTELIGENCENCIA DE NEGOCIOS EN LA CUARTA REVOLUCIÓN INDUSTRIAL

La última parte del libro se centra en la inteligencia de negocios del futuro presente en la tendencia Industria 4.0 y su asociada cuarta revolución industrial. Las grandes consultoras de TI comienzan a denominar a esta tendencia futura de IN, la inteligencia de negocios moderna.

El capítulo 13, Inteligencia Artificial *y Algoritmos en la Inteligencia de Negocios,* se centra en una de las tendencias tecnológicas de mayor impacto en la actualidad y prevista para el futuro, inteligencia artificial y los algoritmos que son su espina dorsal. Las técnicas fundamentales de la inteligencia de negocios, aprendizaje automático y aprendizaje profundo, se describen en el capítulo. Una de las aplicaciones de inteligencia artificial que más se comienzan a utilizar y se utilizarán en el futuro, los asistentes virtuales (*chatbots*), en las organizaciones y empresas, se describen en el capítulo.

Ciencia de datos: la evolución de la minería de datos, es el soporte del capítulo 14. La ciencia de datos es la evolución más avanzada de las técnicas de minería de datos y otras tendencias descritas a lo largo de la obra. Se describe en el capítulo el proceso de ciencias de datos y las herramientas más sobresalientes. Otro concepto importante es la descripción del rol profesional del científico de datos, una de las profesionales más demandadas en la actualidad y en el futuro, por organizaciones y empresas de todo tipo.

El capítulo 15, *Tendencias de futuro de la inteligencia de negocios. Privacidad, protección y seguridad de los datos,* analiza las tendencias de futuro de la inteligencia de negocios y los riesgos y oportunidades de la privacidad, protección de datos y seguridad de los datos, así como reflexiones sobre el uso de la ética y de la responsabilidad social corporativa. Se realiza un análisis de los profesionales del futuro y relacionados con la inteligencia de negocios. Se termina describiendo las técnicas de la inteligencia de negocios moderna así como una introducción a las tendencias tecnológicas de impacto en los negocios para 2018 publicadas por un estudio de la consultora Gartner.

RECURSOS

Todos los capítulos contienen: Objetivos del aprendizaje, introducción, desarrollo teórico-práctico de cada capítulo, casos de estudio, resumen, bibliografía básica y de consulta, referencias web. Los casos de estudios y

herramientas de inteligencia de negocio tratan de contener enfoques prácticos, principalmente, apoyados en estudios e informes de consultoras internacionales prestigiosas como Gartner, Forrester, IDC y otras, así como estudios de organismos internacionales como el WEF (Foro Económico Mundial).

AGRADECIMIENTOS

En primer lugar quiero agradecer a todos mis alumnos de las asignaturas de Inteligencia de Negocios, Gestión del Conocimiento y Sistemas Informáticos de las carreras de Ingeniería de Organización Industrial y de Ingeniería Informática de la Facultad de Informática y posteriormente de la Escuela Superior de Ingeniería y Arquitectura de la Universidad Pontificia de Salamanca en el campus de Madrid. Mi experiencia en numerosos cursos y todo el conocimiento, recomendaciones, consultas, trabajos académicos y de investigación de mis alumnos ha sido el soporte fundamental del contenido de esta obra. También quiero agradecer a mis estudiantes de doctorado y a mis doctorandos —tanto españoles como portugueses, brasileños y latinoamericanos— a los que he dirigido sus tesis doctorales en líneas de investigación tales como gestión del conocimiento, inteligencia de negocios, analítica de datos, *Big Data*, etc.

Además la gran cantidad de ayuda y realimentaciones de los numerosos asistentes y colegas académicos y profesionales en mis conferencias, cursos, seminarios, talleres impartidos en universidades españolas y sobre todo latinoamericanas donde he impartido materias relacionadas con la inteligencia de negocios en estos últimos años. Así en estos tres últimos años he tenido la suerte de tener estancias académicas en países como Ecuador, México, Colombia, República Dominicana, Cuba, Perú, Panamá y Nicaragua, donde he impartido no solo conferencias y cursos específicos sobre Inteligencia de Negocios, Sistemas de Información o *Big Data*.

En último lugar, no puedo dejar de citar en primer lugar y a modo personal, a mi editor —y sin embargo gran amigo— Damián Fernández que como siempre en otras ocasiones similares, me ayuda a lo largo de todo el proyecto editorial y me asesora en cuanto así lo requiero o necesito. De igual forma a Marcelo Grillo, director editorial, con el que siempre cuento a la hora de orientaciones, consultas y referencias sobre mis obras y, en particular de esta obra de inteligencia de negocios. También mi agradecimiento, al resto del equipo editorial de Alfaomega, principalmente de Ciudad de México (CDMX) como de Bogotá (Colombia) de los que siempre recibo realimentación sobre mis obras.

En Carchelejo (Jaén), Sierra Mágina, Andalucía (España)
En Ciudad de México (CDMX), (México)
Enero, 2019

Lecturas complementarias en la Web
En la tabla de contenidos de la obra notará que algunos capítulos extra son para descargar desde nuestra página Web.
Por favor diríjase a http://libroweb.alfaomega.com.mx/ , complete el registro como usuario de nuestra plataforma. Una vez hecho esto busque este título en nuestro buscador y encontrará el link a los capítulos extra.
Estimado profesor: Si desea acceder a las presentaciones (PPTs) de cada capítulo, por favor contacte al representante de la editorial que lo suele visitar o directamente a la sede local de Alfaomega.

Acerca del autor

Luis Joyanes Aguilar

Presidente de la Fundación I+D del Software Libre (Fidesol), Granada (España). Dr. Ingeniero en Informática por la Universidad de Oviedo y Dr. en Sociología por la Universidad Pontificia de Salamanca. Dr. Honoris Causa por la Universidad Privada Antenor Orrego de Trujillo, UPAO, (Perú); por la Universidad San Martín de Porres, Lima (Perú) y por la Universidad Inca Garcilaso de la Vega, Lima (Perú). Líder Académico del TEC de Monterrey, México, campus Querétaro. Catedrático de Lenguajes y Sistemas Informáticos de la UPSA. Profesor de Inteligencia de Negocios y de Ciencia de Datos de la Universidad Católica de Ávila (UCAV) y de la Ávila Business School de UCAV. Profesor invitado y visitante de numerosas universidades de Latinoamérica y El Caribe. Conferenciante habitual en congresos, simposios, jornadas a nivel internacional. Ha dirigido más de 50 tesis doctorales de estudiantes españoles, portugueses y latinoamericanos. Ha escrito más de 40 libros de TIC y más de 100 artículos científicos y profesionales. Su último libro ha sido "Industria 4.0. La Cuarta Revolución Industrial". Investigador del Grupo de Investigación de "Ética en la Nube" de la Facultad de Filosofía de la Universidad Complutense de Madrid. Miembro del Instituto Universitario "Agustín Millares" de la Universidad Carlos III de Madrid. En abril de 2018 recibió la Mención Honorífica del Doctorado en Ingeniería de la Universidad Distrital Francisco José de Caldas, de Bogotá (Colombia).

PARTE I

VISIÓN MODERNA DE INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS

CAPÍTULO 1

INTELIGENCIA DE NEGOCIOS

UNA PANORÁMICA GLOBAL

CONTENIDO

1.1 Introducción
1.2 Inteligencia de negocios: Historia, definiciones y conceptos
1.3 Business Intelligence, Business Analytics y Big Data: Los tres pilares de la inteligencia empresarial
1.4 Arquitectura de un sistema de inteligencia de negocios
1.5 Introducción a Big Data y su impacto en la inteligencia de negocios
1.6 Arquitectura de inteligencia de negocios con integración de Big Data
1.7 Visión gerencial de inteligencia de negocios
1.8 Analítica de negocios (Business Analytics)
1.9 Inteligencia de negocios en Big Data
1.10 Inteligencia de negocios móvil
1.11 Inteligencia de negocios en la nube
1.12 Proveedores de inteligencia de negocios: Cuadrante mágico de Gartner de BI & Analytics
1.13 Inteligencia de negocios futura: Integración de Big Data, Internet de las Cosas e Inteligencia Artificial.
1.14 La evolución hacia la Inteligencia de negocios en la nube (Cloud BI)
1.15 RESUMEN

OBJETIVOS
• Conocer y comprender los conceptos fundamentales de inteligencia de negocios
• Conocer los objetivos de la inteligencia de negocios y su importante rol en la toma de decisiones de la gestión empresarial.
• Conocer la arquitectura de un sistema de inteligencia de negocios y sus componentes básicos.
• Conocer la infraestructura física de un sistema de inteligencia de negocios.
• Conocer los conceptos fundamentales de la analítica de negocio y su integración dentro de la inteligencia de negocios.
• Introducción a la analítica de negocios
• Introducción a Big Data
• Introducción a la analítica de big data.
• Conocer los diferentes tipos de inteligencia de negocios: móvil, en la nube y de big data.
• Introducción a la disciplina de ciencia de datos (data science) y el rol de científico de datos

1.1 INTRODUCCIÓN

Se realiza una introducción teórico-práctica a la Inteligencia de Negocios y a la Analítica de Negocios, extendida a la Analítica de Datos y sus diferentes categorías, así como a la Analítica de *Big Data*, dada la expansión de esta tendencia tecnológica en todo tipo de organizaciones y empresas, además de en los mundos académicos y de investigación. Se tratará de dar respuesta a las preguntas más utilizadas en la gestión empresarial y en los campos de la educación y de la investigación, tales como:

- ¿Qué es la Inteligencia de Negocios y la Analítica de Negocios, comparación y diferencias esenciales entre ambas?
- ¿Cuál es la infraestructura y la arquitectura de un sistema de Inteligencia de Negocios?
- ¿Cuáles son los diferentes modelos de Inteligencia de Negocios?
- ¿Qué es *Big Data*, una introducción a su concepto y a la Analítica de *Big Data*?
- ¿Qué es la Inteligencia de Negocios Móvil y la Inteligencia de Negocios en la nube?

En este capítulo se hará una primera introducción a los proveedores de soluciones de Inteligencia de Negocios tanto de *software* propietario como de *software* de código abierto (*open source*), para lo cual se hará un análisis del cuadrante mágico publicado por la consultora Gartner de *Inteligencia de Negocios y Plataformas Analíticas (Magic Quadrant for Analytics and Business*

Intelligence Platforms) ediciones 2017 y 2018, uno de los informes más acreditados en el mundo empresarial y, en particular, sobre Inteligencia de Negocios.

1.2 INTELIGENCIA DE NEGOCIOS: HISTORIA, DEFINICIONES Y CONCEPTOS

En la última década del siglo XX, los sistemas de apoyo a la decisión (***DSS***) eran el término dominante en la gestión empresarial y comenzaba a utilizarse una nueva disciplina conocida como Inteligencia de Negocios, la cual ha ido evolucionando y ganando fuerza y en la que se han integrado los DSS, dado que la idea central gira en torno a los datos de las empresas, su conversión en conocimiento para que, tras el correspondiente análisis, ayuden en la toma decisiones empresariales.

El término **Inteligencia de Negocios** (*Business Intelligence*) —con frecuencia, también se utiliza el término **Inteligencia de Negocio** (en singular) — fue acuñado por Gartner a mitad de la década de los 90, aunque el concepto tiene su origen en el comienzo de los sistemas de información gerenciales (MIS, *Management Information System*) de los años 70, cuando comenzaba la automatización de las tareas en las empresas. Hoy en día los sistemas de información son la espina dorsal de las empresas y su soporte diario y el eje sobre el que se vertebran los sistemas de Inteligencia de Negocios. Sistemas conocidos como **ERP, CRM, SCM, GIS,** etcétera, ya sea de modo independiente o integrados en paquetes de *software* "*suites*", son de uso diario en las organizaciones para la gestión de los datos corporativos. La necesidad de añadirle conocimientos (*insights*) adecuados para ayudar a la toma de decisiones ha ido asentando el concepto de Inteligencia de Negocios como un conjunto de componentes —infraestructura física, de *hardware* y *software*— que conforman una arquitectura para ayudar a una eficiente toma de decisiones. Un sistema de Inteligencia de Negocios incluye numerosas herramientas y técnicas que proporcionan grandes capacidades para la transformación de los datos en conocimiento que ayuden a la adecuada toma de decisiones con la realización de las acciones oportunas. Así, a lo largo del libro iremos desglosando un conjunto grande de técnicas y herramientas que constituyen el soporte de la inteligencia de negocio y la Analítica de Negocios asociadas (*Business Analytics* o *Analytics*), tales como:

- Bases de datos.
- Metadatos.
- *Data Warehouse y Data Marts.*
- *Data Lakes* (lagos de datos).
- Integración de datos (herramientas ETL y ELT).

- Hojas de cálculo (la herramienta más tradicional).
- Alertas y notificaciones.
- Herramientas de visualización: cuadros de mando o tableros de control (*dashboards y scorecards*).
- Informes y consultas (*reporting* y *query*).
- Cuadros de Mando Integral (CMI).
- Reglas de negocio.
- Analítica OLAP
- Analítica de Datos.
- Analítica predictiva y prescriptiva.
- Minería de Datos.
- Ciencia de Datos.

Inteligencia de Negocios comenzó a utilizarse por los proveedores de *software* y consultores de tecnologías de la información como un servicio de cómputo para describir la infraestructura de almacenamiento, integración, reportes y análisis de datos que vienen integrados en los entornos de datos (bases de datos transaccionales y almacenes de datos o "repositorios", incluyendo en la actualidad los grandes volúmenes de datos (*Big Data*) con las bases de datos NoSQL y "en memoria" (*in-memory*).

La infraestructura de Inteligencia de Negocios recolecta, almacena, limpia y pone la información relevante a disposición de los gerentes, apoyándose en bases de datos, repositorios de datos y últimamente Hadoop de *Big Data*, y las plataformas de Inteligencia de Negocios tanto de *software* propietario como de *software* de código abierto. Analítica de Negocios es otro término que es muy utilizado por los proveedores de soluciones de *software*, que se centra más en las herramientas y técnicas para analizar y comprender los datos mediante soluciones de Analítica con modelos estadísticos y de Minería de Datos.

1.2.1 EL ORIGEN DEL TÉRMINO IN

Inteligencia de Negocios (IN) —*Business Intelligence (*BI*)*— ha sido un término paraguas que Turban *et al* (2011) definen "como la combinación de arquitecturas, herramientas, bases de datos, herramientas analíticas, aplicaciones y metodologías"[1]. En realidad, se han mezclado diferentes términos, desde DSS hasta EIS y BPM. En nuestro caso, hemos integrado todas estas tendencias en el único término de Inteligencia de Negocios y Analítica de Negocios, de los que también definiremos y explicaremos las diferencias. El objetivo más importante de la IN (BI) es facilitar el acceso interactivo —hoy en día,

casi siempre en tiempo real— a datos para facilitar su manipulación y proporcionar a los gerentes, directivos y analistas, la capacidad de manejar análisis apropiados a la toma de decisiones.

Desde un punto de vista gerencial y empresarial, el análisis de los datos históricos y actuales, situaciones reales y el examen del desempeño/rendimiento proporciona a los administradores y restantes usuarios de los sistemas de IN la capacidad de adquirir conocimientos (*insights*) que les facilite tomar decisiones más informadas y mejores. El proceso de Inteligencia de Negocios se basa en la transformación de los datos a información, su conversión en conocimiento, para una mejor toma de decisiones y, por último, la realización de las acciones correspondientes y adecuadas.

En resumen, la Inteligencia de Negocios es relativa a las operaciones de captura, acceso, comprensión y conversión de los activos más valiosos de una empresa —los datos en bruto— en información accionable con el objetivo de mejorar su desempeño o rendimiento.

1.2.2 DEFINICIÓN DE INTELIGENCIA DE NEGOCIOS

La Inteligencia de Negocios y la Analítica de Negocios se han convertido en la piedra angular de la estrategia de negocios de las compañías.

La consultora Gartner —referencia mundial en tecnologías de la información y en consultoría estratégica— es considerada como una de las primeras organizaciones que definió el término. En su prestigioso *IT Glossary*, se define **Inteligencia de Negocios** (***Business Intelligence***) como "un término paraguas que incluye las aplicaciones, infraestructuras y herramientas, y las mejores prácticas que facilitan el acceso y análisis de información para mejorar y optimizar decisiones y rendimiento o desempeño (performance)".[2]

Este mismo glosario define **Analítica de Negocios** (***Business Analytics***) como "comprensión de las soluciones utilizadas para construir modelos de análisis y simulaciones para crear escenarios, comprender realidades y predecir estados futuros".[3]

¿Cuáles son las diferencias entre ambos términos? ¿Cómo afectan las tendencias móviles, la nube y *Big Data* al desarrollo de ambas materias? En este capítulo, a modo de introducción, intentaremos dar respuesta a éstas y otras preguntas que surgen en el desarrollo diario de la gestión empresarial, y que profundizaremos en capítulos siguientes.

Otras definiciones que contemplamos son las dadas por el prestigioso portal tecnológico **techtarget.com**, que, al igual que el *IT Glossary* de Gartner, también tiene su propio glosario, pero al que añade una enorme cantidad de artículos, informes y noticias que completan de modo muy amplio los términos que se han de definir: **Inteligencia de Negocios** (***Business Intelligence***) "es un proceso

controlado por tecnología para el análisis de datos y presentación de información accionable para ayudar a los directivos corporativos, gerentes de gestión y otros usuarios en la toma de decisiones de negocios mejor informadas"[4] y **Analítica de Negocios** *(Business Analytics) es* "la práctica de la exploración iterativa y metódica de los datos de una organización con énfasis en el análisis estadístico".[3]

Por último, el prestigioso *The Data Warehousing Institute (tdwi.org)* define Inteligencia de Negocios como "la combinación de tecnología, herramientas y procesos que me permiten transformar mis datos almacenados en información, esta información en conocimiento y este conocimiento dirigido a un plan o una estrategia comercial". La Inteligencia de Negocios debe ser parte de la estrategia empresarial, permitiendo optimizar la utilización de recursos, monitorear el cumplimiento de los objetivos de la empresa y la capacidad de tomar buenas decisiones para así obtener mejores resultados.

Inteligencia de Negocios se refiere al proceso de convertir datos en conocimiento y conocimiento en acciones para crear la ventaja competitiva del negocio (TDWI).

1.2.3 INTELIGENCIA DE NEGOCIOS *VERSUS* ANALÍTICA DE NEGOCIOS

Inteligencia de Negocios o inteligencia empresarial es un término muy utilizado por los proveedores de *hardware* y *software*, así como los consultores de TI (tecnologías de la información), para describir la infraestructura de generación, almacenamiento, integración, generación de informes (*reporting*), análisis y visualización de datos que proceden de los entornos de negocio, incluyendo en la actualidad *Big Data*. La infraestructura de BI captura, almacena, limpia y pone disponible información relevante a los directivos y gerentes, en bases de datos, almacenes de datos (*Data Warehouses* y *Data Marts*), sistemas de *Big Data* como Hadoop/Spark, bases de datos en memoria y plataformas analíticas, así como los novedosos repositorios de datos "*Data Lakes*".

Analítica de Negocios (*Business Analytics*) es también un término acuñado por proveedores y consultores de TI pero enfocado más en herramientas y técnicas para el análisis y comprensión de los datos. Las herramientas van desde el procesamiento analítico en línea (OLAP), estadísticas, modelos de datos, minería de datos y cada vez más herramientas de inteligencia artificial, como aprendizaje automático y aprendizaje profundo. Analítica de Negocios, aunque tiene soluciones propias, se suele integrar como subconjunto a Inteligencia de Negocios y así lo consideraremos en el libro, pese a la gran influencia en consultorías y asesorías de negocios del término *analytics*.

1.3 *BUSINESS INTELLIGENCE, BUSINESS ANALYTICS* Y *BIG DATA*: LOS TRES PILARES DE LA INTELIGENCIA EMPRESARIAL

En la prensa generalista y en la prensa económica o tecnológica especializada, se suelen utilizar los tres términos, bien de modo diferenciado o bien como sinónimos. La realidad es que los tres conceptos conviven en consultoras, medios de comunicaciones, proveedores de *software*, desarrolladores de aplicaciones, etc. Es difícil encontrar semejanzas y diferencias, pero trataremos de hacerlo en este caso desde el punto de vista de que las tres disciplinas sirven para dar soporte a la toma de decisiones. Prueba evidente de la dificultad de acotar bien los objetivos y características fundamentales de las tres tendencias empresariales es la gran cantidad de Master y Maestrías con nombres más variados: *Business Intelligence & Analytics*, *Big Data & Business Intelligence*, *Big Data & Analytics*, *Big Data &* Ciencia de Datos, *Big Data* y Analítica Visual. Para tratar de dar luz a la polémica de los términos, pero sobre todo entender que en la segunda mitad de la segunda década del siglo XXI aunque los tres términos se traten de forma independiente o conjunta, los objetivos de las tres disciplinas son de vital necesidad para implantar las estrategias de negocio de las empresas. Vamos a recurrir a diferentes autores y a estudiar sus diferentes opiniones.

En nuestra visión particular consideraremos que inteligencia de negocios es un superconjunto necesario para obtener el mayor rendimiento de *Big Data* y de Analítica de Negocios o *Analytics*, y por ello en la obra trataremos de analizar las tres disciplinas y cómo integrarlas en el beneficio de la toma de decisiones de éxito en las empresas.

Inteligencia de Negocios

Se entiende por *Business Intelligence* el conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización (Josep Curto)[5].

Como ya se mencionó, es un término paraguas que combina arquitecturas, herramientas, bases de datos, herramientas de analítica, aplicaciones y metodología (Efraim Turban *et al*). El objetivo principal de BI es facilitar el acceso interactivo (de ser posible en tiempo real) a datos, facilitar la manipulación de los mismos y dar a los gerentes de negocios y analistas la capacidad de conducir el análisis adecuadamente. El proceso de BI se basa en la transformación de los datos en información, a continuación decisiones y, por último, acciones.

Business Analytics

Se entiende por *Business Analytics* el conjunto de estrategias, tecnologías y sistemas que permiten analizar el rendimiento pasado de una organización para poder predecir comportamientos futuros, así como para detectar patrones ocultos en la información.

Es la aplicación directa de modelos a los datos del negocio. El análisis de Negocios implica el uso de herramientas de DSS, especialmente modelos, que asisten a las tomas de decisiones.

Big Data

Se entiende por *Big Data* el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos, que frecuentemente, pero no siempre, viene definida por volumen, velocidad y variedad. (Josep Curto)[6].

Es el acceso a grandes volúmenes de datos, pero el valor real no se encuentra en ellos, sino en lo que podemos hacer con ellos. No es la cantidad de información lo que marca la diferencia, sino que se trata de nuestra capacidad para analizar series extensas y complejas de datos que van más allá de todo lo que hubiéramos podido hacer anteriormente. Esto significa que todas las empresas, organismos gubernamentales o cualquier persona realmente pueden utilizar el *Big Data* para mejorar la toma de decisiones (Bernard Marr 2016).

1.4 ARQUITECTURA DE UN SISTEMA DE INTELIGENCIA DE NEGOCIOS

La Inteligencia de Negocios se centra en el modo de capturar, acceder, almacenar, procesar, analizar y visualizar los resultados, convirtiendo uno de los activos más valiosos de una empresa, los datos en bruto (*raw data*), en información accionable con el objeto de mejorar el rendimiento del negocio. BI busca hacer corresponder el almacenamiento de datos y su procesamiento con herramientas analíticas, para proporcionar a los tomadores de decisiones una información competitiva que los diferencie de modo eficiente en su entorno de negocio.

A medida que la organización comienza a adoptar la BI, una tarea muy importante por realizar es asegurarse de que la misma sigue un buen plan arquitectónico en su proceso de implementación, de modo que compense con éxito la inversión realizada en el proyecto. La arquitectura de BI es un marco de trabajo (*framework*) que detalla los diferentes componentes del sistema de Inteligencia de Negocios, tales como datos, personas, procesos, tecnologías y gestión/administración, y la forma en que estos componentes se han de combinar y coordinar para asegurar el correcto funcionamiento del sistema.

La información contenida en una arquitectura de BI es el conjunto de tipos de datos que necesitan ser recolectados, los métodos que se utilizan para analizar los datos y el modo en que se presenta la información necesaria. Se requiere una arquitectura de BI sólida; si la arquitectura no está diseñada adecuadamente, se producirán inconsistencias que afectarán a los diferentes componentes y puede conducir a problemas como, por ejemplo, la incapacidad para compartir información entre dichos componentes. Una mala arquitectura de BI puede

conducir a un escenario de entrega de información incorrecta, inadecuada y, en momentos, equivocada a las personas correspondientes.

La arquitectura de un sistema de Inteligencia de Negocio consta de una serie de componentes o capas que, conectados debidamente, realizan las tareas fundamentales para la ayuda en la toma de decisiones empresariales. Debe tomar en consideración la calidad de los datos, así como el flujo de información en el sistema de Inteligencia de Negocios.

Existen varias arquitecturas de inteligencia de negocios realizadas por diferentes autores, empresas de software y consultoras, las cuales han ido evolucionando a medida que se consolidaban nuevas tecnologías y tendencias estratégicas empresariales junto con el crecimiento exponencial de los datos (big data) manejados por organizaciones y empresas. Por estas razones hemos decidido considerar dos arquitecturas de inteligencia de negocios: tradicional y con soporte de big data (apartado 1.6).

La arquitectura tradicional seleccionada es la propuesta presentada por Lih Ong et al (2011)[7] que se sustenta en un análisis y estudio previo de las metodologías existentes a finales de la primera década del siglo XXI, tales como las metodologías de Baars y Kemper (2008) y Turban et al (2008); la metodología de Turban ha ido evolucionando también con el tiempo y han sido presentadas en sucesivas ediciones de su libro de referencia de inteligencia de negocios (2008, 8ª ed.; 2011, 9ª ed., 2014, 10ª ed.). El impacto de la metodología de Turban en el sector de inteligencia de negocios es considerable y por ello abordaremos sus componentes principales en el apartado 1.7 con un enfoque gerencial e integrado con big data.

ARQUITECTURA DE INTELIGENCIA DE NEGOCIOS DE CINCO CAPAS

La arquitectura de inteligencia de negocios tradicional propuesta por Ong et al (2011)[8] tiene en cuenta, entre otras consideraciones, el valor y calidad de los datos (proceso de calidad de los datos) así como el flujo de información del sistema (proceso de gobierno de los datos). La metodología se compone de cinco capas:

- Capa de fuentes de datos.
- Capa de proceso ETL (Extract, Transform, Load)
- Capa de almacenes de datos (Data Warehouse, Data Mart)
- Capa de metadatos
- Capa de usuario final (análisis y visualización de resultados)

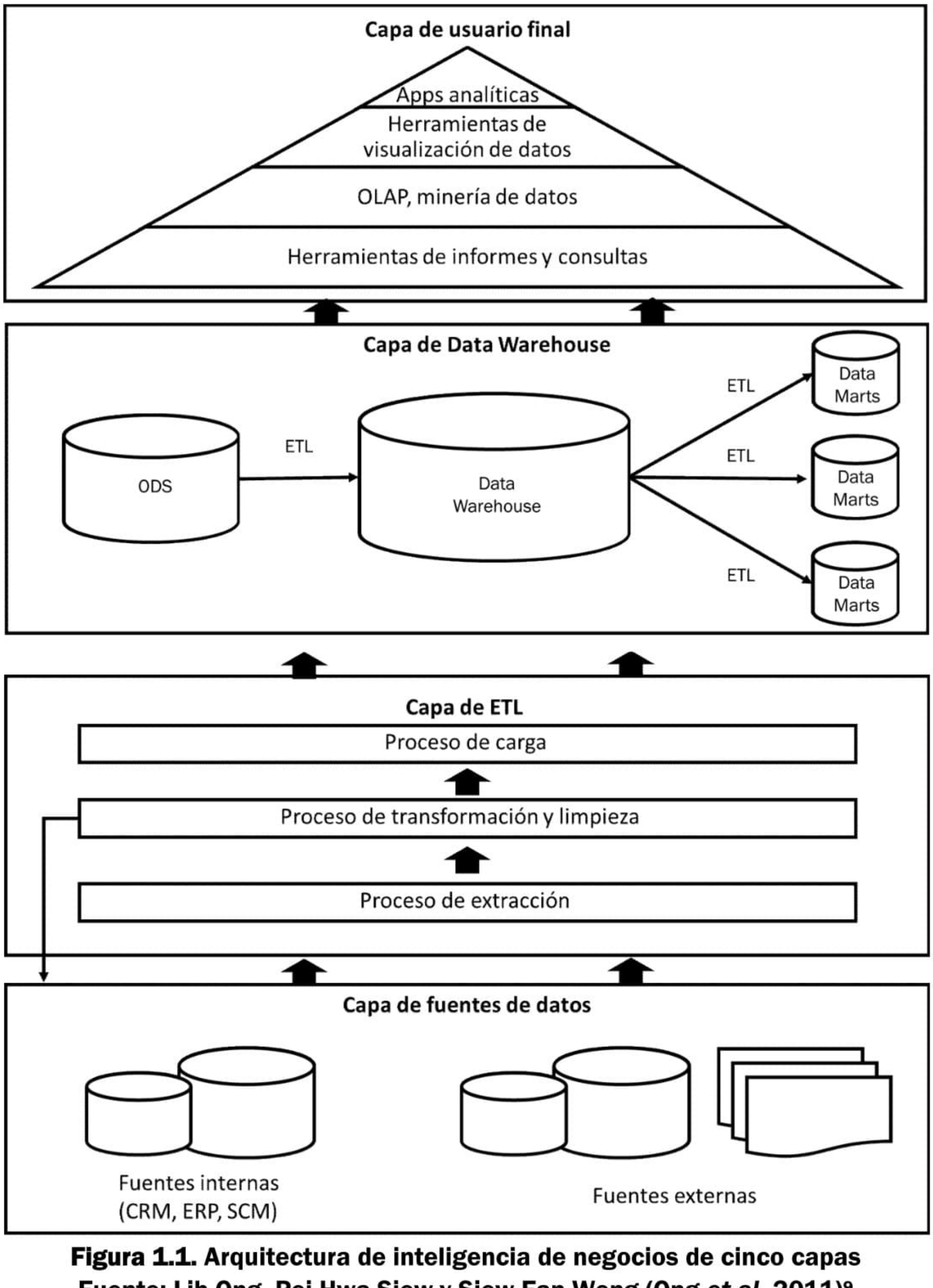


Figura 1.1. Arquitectura de inteligencia de negocios de cinco capas
Fuente: Lih Ong, Pei Hwa Siew y Siew Fan Wong (Ong *et al*, 2011)[9]

1.4.1 FUENTES DE DATOS

Los datos del entorno de negocio son, en la actualidad, de tres tipos diferentes: estructurados, no estructurados y semiestructurados, que deben ser entregados

de modo efectivo y en el momento que se necesiten. Estos datos proceden de diferentes fuentes, incluyendo *Big Data*, y se adquieren de dos tipos de fuentes: internas y externas.

Las **fuentes de datos internas** se refieren a los datos que son capturados y mantenidos por los sistemas operacionales dentro de las organizaciones, tales como sistemas CRM, ERP, SCM o GIS. Las fuentes de datos internas incluyen los datos relacionados con las operaciones de negocio (por ejemplo, datos de clientes, productos y ventas). Estos sistemas operacionales tradicionalmente se conocen como sistemas de transacción en línea (transaccionales), ya que ellos procesan grandes cantidades de transacciones en tiempo real y actualizan los datos siempre que sea necesario. Los sistemas operacionales contienen sólo los datos actuales que se utilizan para soportar las operaciones diarias de negocio de una organización. Normalmente, estos sistemas operacionales están orientados a procesos, de modo que se centran en operaciones específicas de negocio tales como ventas, compras, contabilidad o recursos humanos.

Las **fuentes de datos externas** se refieren a las que se originan en el exterior de una organización. Este tipo de datos se pueden recolectar de fuentes externas tales como socios de negocio (*partners*), proveedores de datos, Internet, gobiernos y corporaciones nacionales y locales, organizaciones de investigación de mercados o científicas, datos demográficos. Es importante para las organizaciones identificar sus fuentes de datos y los métodos de acceso a los mismos. Este conocimiento de las fuentes facilitará posteriormente la replicación, limpieza y extracción de los datos. Es muy importante identificar las fuentes, ya que se pueden encontrar con datos innecesarios, no fiables o irrelevantes para las necesidades actuales o futuras del negocio.

Las fuentes de datos en la actualidad son muy diversas y los datos que proporcionan no siempre serán *estructurados* (formatos fijos de tabla, filas y columnas); al contrario, serán en un gran porcentaje (estadísticas fiables hablan del 80 al 90% de los datos manejados por una organización en la actualidad) *no estructurados* (texto, video, audio, imágenes). En estos casos se contemplarán los grandes volúmenes de datos (*Big Data*) y su manipulación requerirá nuevos sistemas de recolección y almacenamiento. Dada su importancia se introducirán más adelante y se estudiará el modo de integración de *Big Data* en los sistemas de inteligencia de negocio. A continuación, diferentes tipos de datos manejados por las organizaciones en la actualidad y sus fuentes respectivas:

- Sistemas operacionales (bases de datos y archivos)
- ERP.
- CRM.
- SCM.
- GIS.
- Sistemas heredados.

- Sistemas de información departamentales.
- Datos del entorno de negocios.
- Datos de la Web.
- Datos de dispositivos móviles.
- Datos de sensores y de dispositivos de ciudades inteligentes.
- Datos de *Social Media* (medios sociales y redes sociales).
- Proveedores.
- Económicos de empresas y de administraciones públicas y gubernamentales.
- Otros datos externos procedentes de fuentes diversas (Internet de las cosas, datos biométricos).

Los datos en bruto (*raw*) extraídos de las fuentes de datos serán integrados y organizados de modo que puedan ser analizados, para poder ser utilizados posteriormente por las personas encargadas en la toma de decisiones.

1.4.2 PROCESO ETL

La capa ***ETL*** (*Extract, Transform, Load*) se centra en tres procesos principales: extracción, transformación y carga de los datos. Extracción, es el proceso de identificación y recolección de datos relevantes o significativos de diferentes fuentes. Normalmente, los datos extraídos de fuentes de datos internas y externas no están integrados y pueden ser incompletos y estar duplicados. El proceso de extracción se necesita para seleccionar datos que sean significativos para la toma de decisiones en las organizaciones.

Los datos extraídos se envían a un área de almacenamiento temporal que se llama *Data Staging*, que es previa al proceso de transformación y limpieza. La transformación es el proceso de conversión de los datos, utilizando un conjunto de reglas de negocio (tales como funciones de agregación) en formatos consistentes para realizar informes o reportes y análisis. Una vez que los datos se han limpiado y transformado se almacenan en la citada área temporal (*Staging Area*).

La última fase del proceso ETL es la carga de los datos del área de *staging* en el repositorio destino (*Data Warehouse* y *Data Marts*), normalmente a través de un almacén de datos operacional (ODS).

1. **Etapa de extracción**: consiste en capturar datos de fuentes heterogéneas y homogéneas. Las herramientas de extracción que se utilizan en esta etapa soportan múltiples formatos de almacenamiento de datos.

2. **Etapa de transformación**: aplica un conjunto de reglas de unificación de datos básicos para transformar los datos desde el origen al destino. Esto incluye la conversión de los datos medidos a la misma dimensión, usando las mismas unidades, para que más adelante se puedan unificar. Una vez transformados los datos, es necesario realizar una serie de operaciones de depuración. Esta etapa es una de las más importantes, ya que garantiza la calidad de los datos por tratar.
3. **Etapa de carga**: es necesario garantizar que esta operación se realiza correctamente y empleando el menor número de recursos posible.

La **figura 1.2** muestra el proceso de flujo de los datos mediante herramientas ETL, desde las fuentes de datos al almacén destino, *Data Warehouse*, y de allí a los *Data Marts* departamentales (extraer, transformar, cargar).

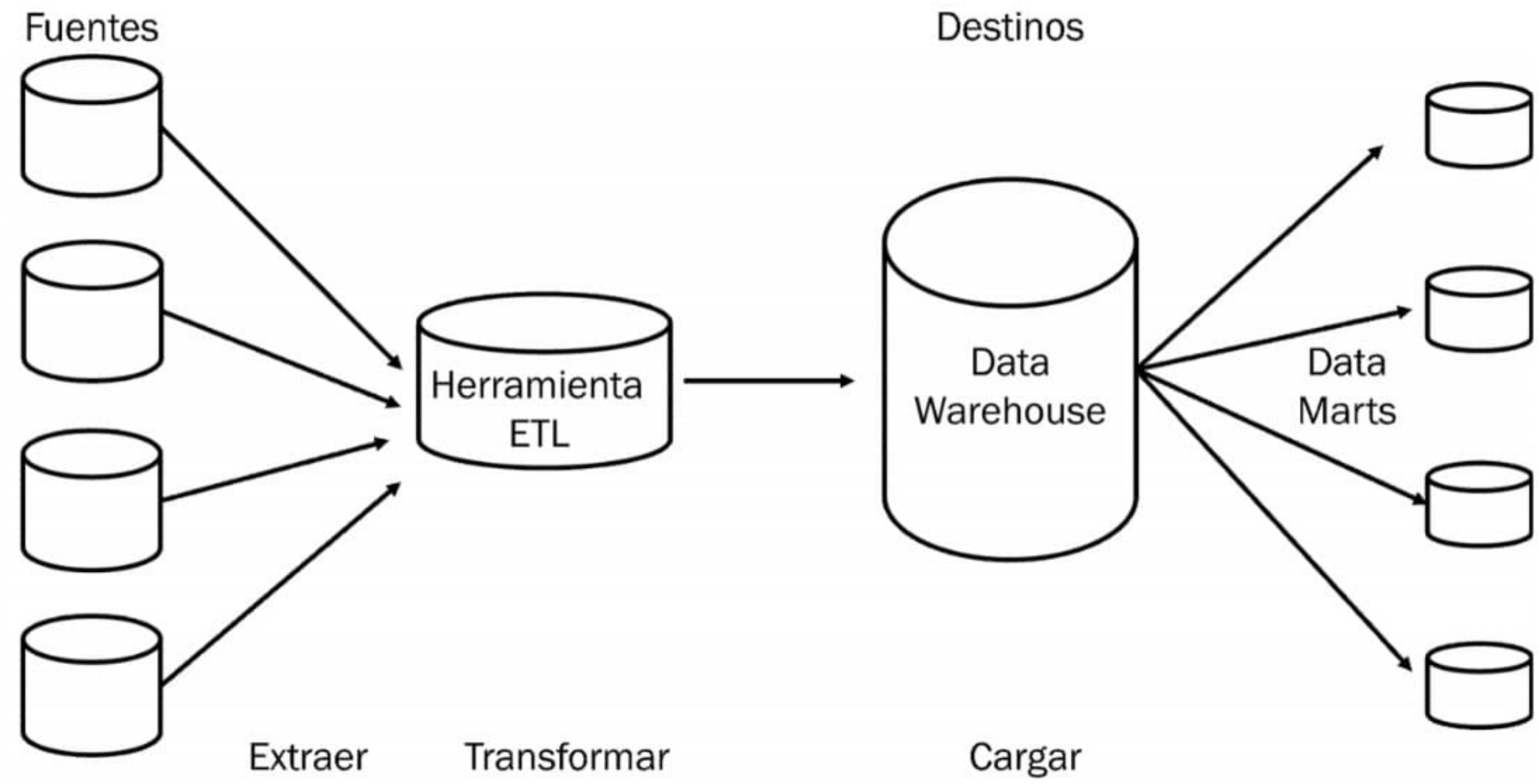


Figura 1.2. Flujo de datos en proceso ETL

El proceso ETL, en algunos casos que comentaremos más adelante, sobre todo en el procesamiento de *Big Data*, puede ser sustituido por procesos ELT. El sistema funciona extrayendo los datos de las fuentes, transformando y cargando directamente en el *Data Warehouse* con una sola operación.

Descargado en: eybooks.com

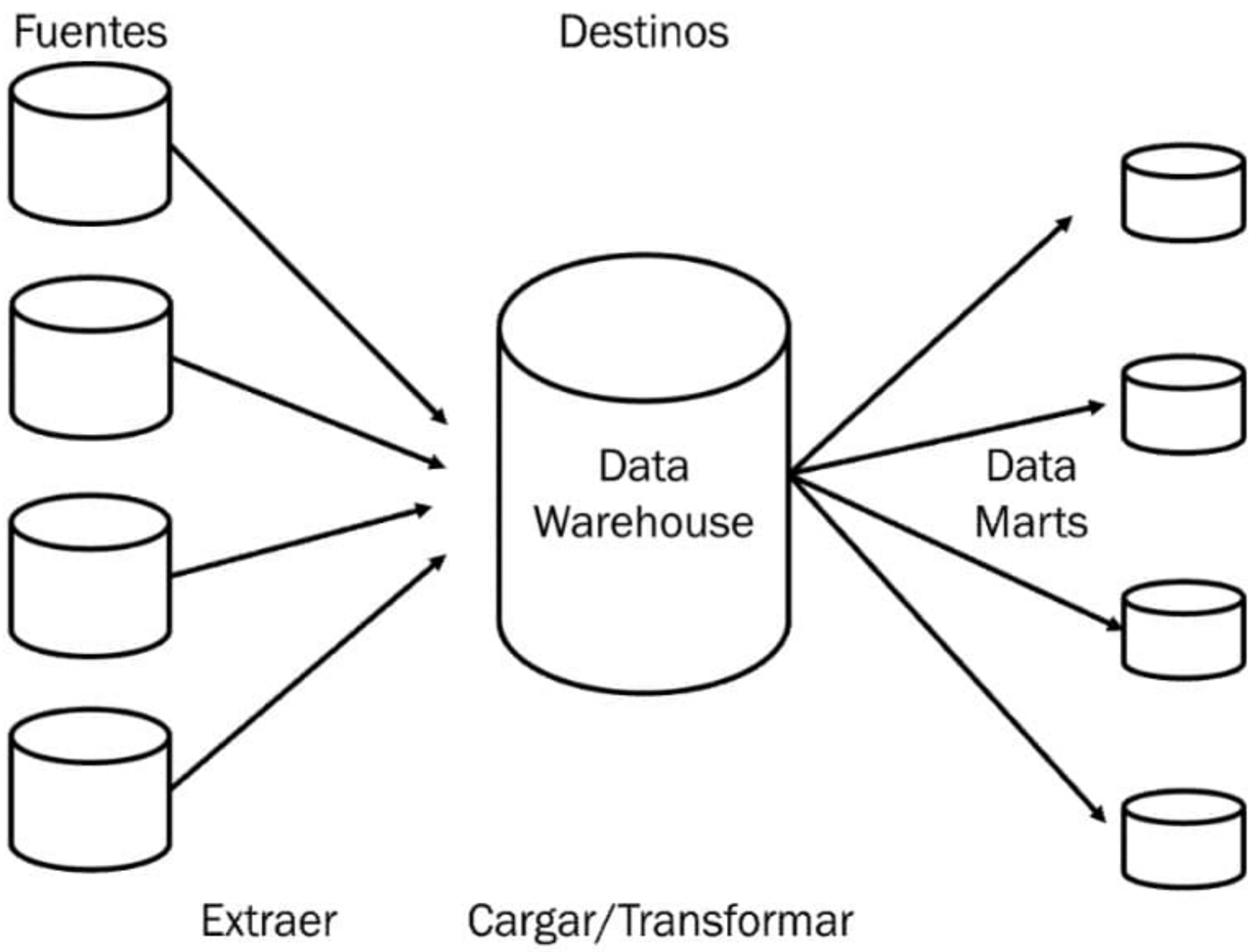


Figura 1.3. Proceso ELT (extraer, cargar, transformar)

1.4.3 ALMACENES DE DATOS (*DATA WAREHOUSES* Y *DATA MARTS*)

La capa de almacenamiento de datos consta de tres componentes: el almacén de datos operacional (ODS del inglés *Operational Data Store*), el *Data Warehouse* (almacén de datos) y los *Data Marts* (almacenes de datos corporativos). Los flujos de datos van del ODS al *Data Warehouse* y, posteriormente, a los *Data Marts*. El *Data Warehouse* es uno de los componentes más importantes de la arquitectura de un sistema de Inteligencia de Negocios (en el capítulo 4 se describirá en detalle la descripción de los almacenes de datos (*Warehouses* y *Data Marts*), así como las nuevas infraestructuras de lagos de datos (*Data Lakes*).

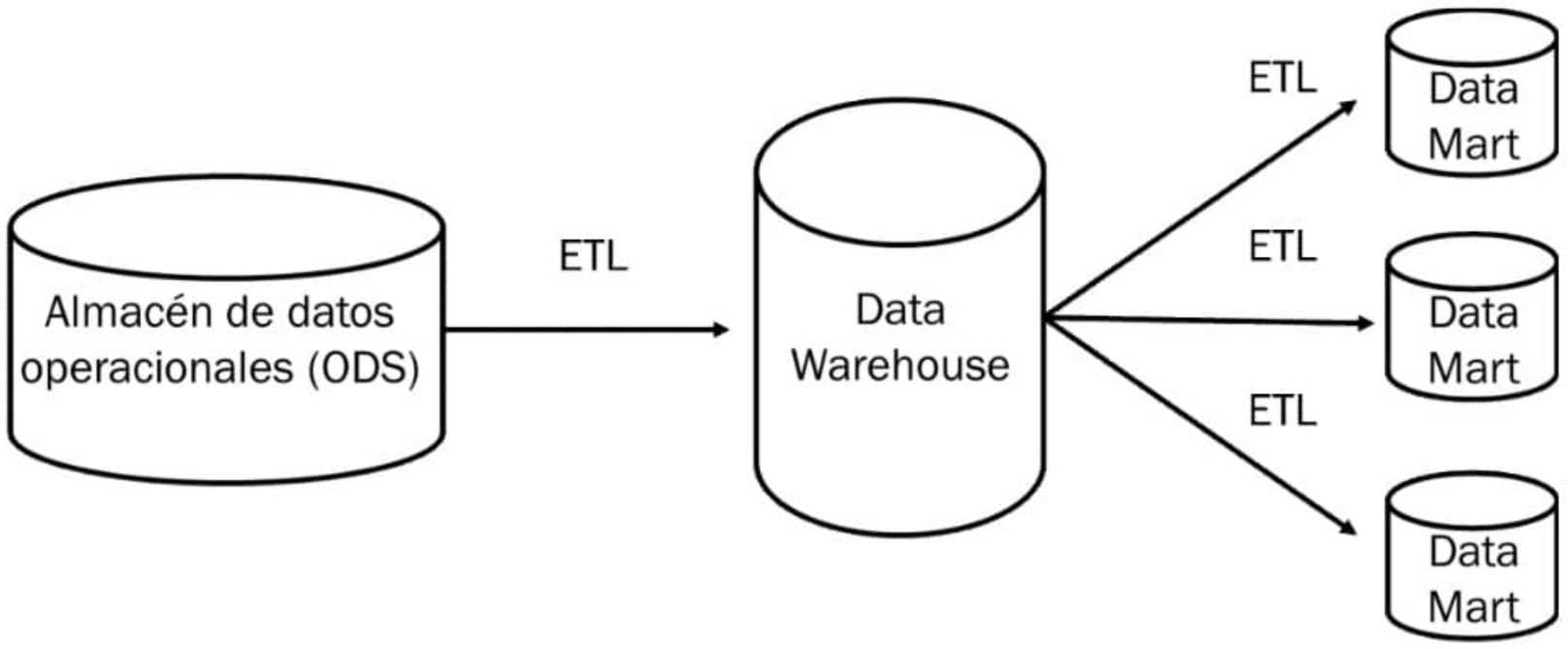


Figura 1.4. Capa de *Data Warehouse* (almacén de datos)

La herramienta tradicional de almacenamiento de datos corporativa ha sido —y sigue siendo— durante años el almacén de datos. Un ***almacén de datos*** (*Data*

Warehouse) es una base de datos que almacena datos históricos y actuales de interés potencial para tomar decisiones en la empresa.

Los datos se originan en numerosas fuentes de datos, tales como sistemas de transacciones operacionales (sistemas de ventas, cuentas de clientes, sistemas de fabricación e, incluso, transacciones de sitios web). Un almacén de datos extrae datos internos actuales o históricos de múltiples sistemas operacionales de la organización. Estos datos internos se combinan con datos procedentes de fuentes externas. Todos estos conjuntos de datos se han de transformar y dejar preparados para la gestión de informes y de consultas, mediante operaciones de limpieza y reestructuración de datos, antes de ser cargados en el almacén de datos, mediante las operaciones ETL (extraer, transformar y cargar). Un *Data Warehouse* pone los datos disponibles a disposición de cualquier persona que pueda necesitarlos, pero no se pueden alterar.

Las empresas construyen normalmente un almacén de datos, donde un almacén central sirve a la totalidad de la empresa. Sin embargo, el almacén de datos central no suele estar equipado para soportar las necesidades específicas y el requerimiento de los departamentos específicos y se requieren nuevos componentes para cumplir estas funcionalidades. Estos componentes se llaman *Data Marts* (almacenes de datos corporativos) y se pueden construir como almacenes de datos más pequeños que son descentralizados y que sirven a un departamento o división de la empresa. Un *Data Mart* es un subconjunto de un almacén de datos (*Data Warehouse*) que se almacena en bases de datos independientes y se pone a disposición de un público específico, perteneciente a un determinado departamento. Así, por ejemplo, una empresa puede desarrollar *Data Marts* de datos de venta y *marketing* (datos de puntos de venta de almacenes minoristas "*retail*"). De este modo, existirá un gran almacén de datos (*Data Warehouse*) y varios almacenes de datos departamentales (*Data Marts*) para los departamentos de ventas, mercadotecnia (*marketing*), recursos humanos, etcétera.

Esta capa es la encargada del almacenamiento de datos, previa organización y preparación de los datos. Sus componentes son:

- Procesos ETL (extracción, transformación y carga).
- Sistemas ODS.
- Almacenes de datos (bases de datos, *Data Warehouses* y *Data Marts*).
- Metadatos.
- Sistemas de *Big Data* (Hadoop: HDFS, MapReduce y Spark).
- Plataformas de Analítica.

El sistema de *Data Warehouse* tiene una capa previa de enlace entre las fuentes de datos y el citado sistema **ETL** (*Extraction, Transformation, Load*) que consta de tres procesos: extracción, transformación y carga. En el proceso de

extracción se realiza la recolección o captura de datos; una vez recolectados los datos, se pasa al proceso de transformación, donde los datos se transforman, integran y limpian; una vez limpiados los datos, el siguiente proceso carga y actualiza los datos en los almacenes de datos.

La infraestructura de Inteligencia de Negocios se soporta en un sistema potente de bases de datos que captura todos los datos relevantes de operación del negocio. Los datos se pueden almacenar en bases de datos operacionales (transaccionales) o combinadas e integradas en un almacén de datos (*Data Warehouse*) o almacenes de datos departamentales (*Data Marts*). Originalmente, los almacenes de datos incluían los datos históricos de las compañías que se organizaban, preparaban y resumían para que los usuarios finales pudieran visualizar o manipular datos e información. En la actualidad, los almacenes de datos pueden manejar datos en tiempo real y, en numerosas ocasiones, requieren la integración de los sistemas de *Big Data*.

Un **almacén de datos** (***Data Warehouse***) es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, con independencia de cómo se vayan a emplear, posteriormente, por los diferentes usuarios. Los ***Data Marts,*** o ***Data Warehouses*** **departamentales**, son un subconjunto de los almacenes de datos enfocados y de valor para un departamento determinado de la empresa, para un conjunto de usuarios o, incluso, para un análisis de datos específico.

En algunos sistemas de almacenamiento de datos, existen almacenes intermedios entre el ETL y los *Data Warehouses*, denominados **ODS** (sistemas de datos operacionales). El otro componente de la infraestructura son los **metadatos**, que son, a su vez, generadores de datos y que alimentan a todos los *Data Warehouses* y *Data Marts*.

1.4.4 CAPA DE METADATOS

Los metadatos se refieren a datos acerca de los datos. La capa describe donde se utilizan y almacenan los datos, las fuentes de datos, cuales cambios se realizan a los datos y cómo una pieza de datos se refiere a otra información. El repositorio de datos (depósito) de los metadatos se utilizan para almacenar información técnica y de negocio acerca de datos, así como reglas de negocio y definiciones de datos. Los metadatos más usuales propuestos por la arquitectura Ong son:

- Fuentes de datos
- ETL
- Informes (Reporting)
- OLAP (Procesamiento analítico en línea)
- Minería de datos

- Repositorios (depósitos de datos)

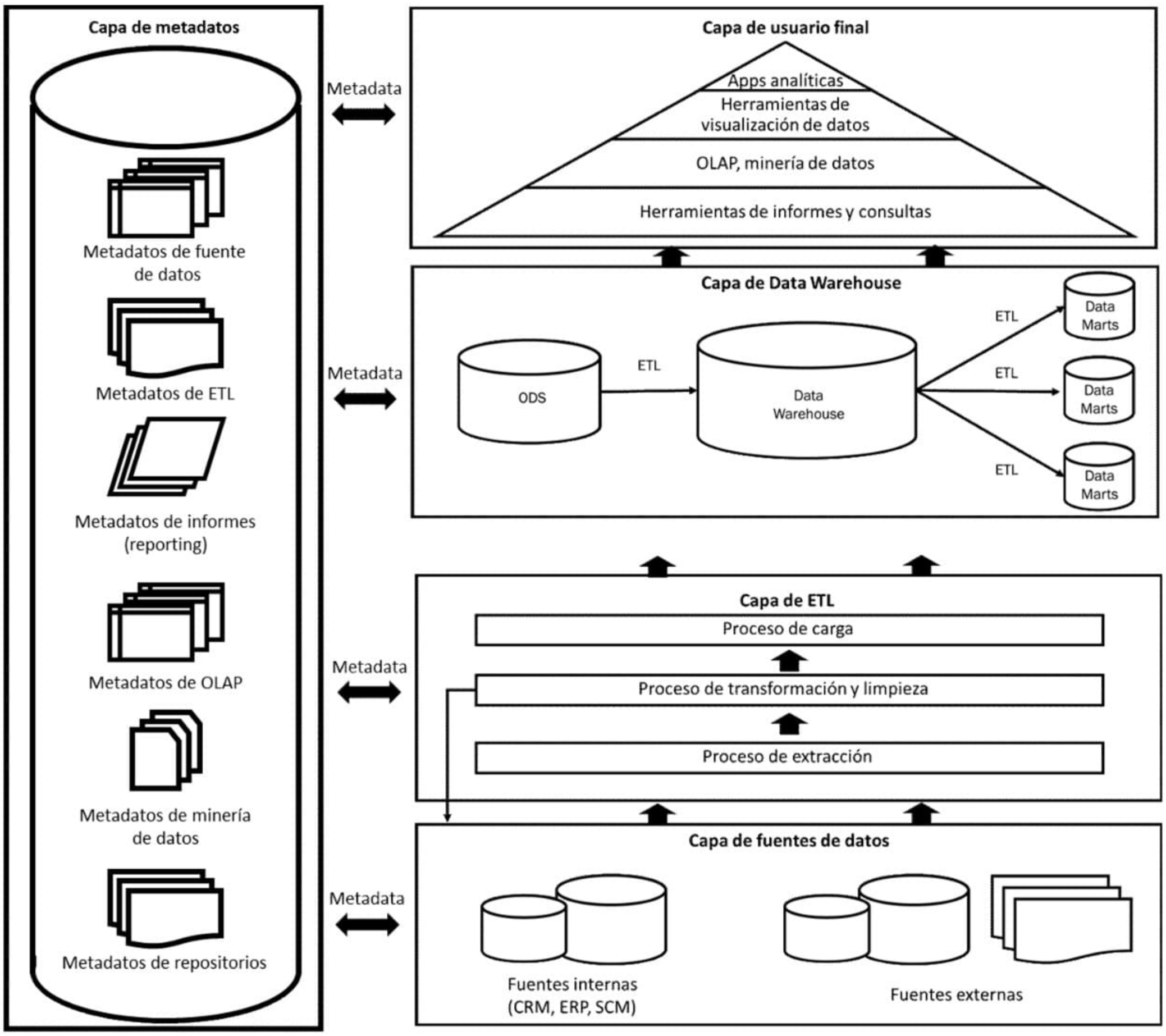


Figura 1.5. Capa de metadatos. Fuente: (Ong *et al*, 2011)[10]

- Metadatos de fuentes de datos
- Metadatos de ETL
- Metadatos de informes (*Reporting*)
- Metadatos de OLAP (Procesamiento analítico en línea)
- Metadatos de minería de datos
- Metadatos de repositorios (depósitos de datos)

1.4.5 CAPA DE USUARIO FINAL

La capa de usuario final se compone de una serie de herramientas que visualizan la información en diferentes formatos y para diferentes usuarios. Estas herramientas se pueden agrupar de modo jerárquico (Ong *et al*, 2011) en sentido ascendente y en forma de pirámide, clasificadas por la facilidad de comprensión creciente de la presentación de los datos a medida que son procesados. Así, por ejemplo el nivel más alto de la pirámide consta de aplicaciones analíticas, normalmente utilizadas por los directivos y gerentes de alto nivel estratégico, mientras que el nivel más bajo, herramientas de consultas y reportes, se utilizan, principalmente, por el nivel de gestión operacional.

Las herramientas de la capa de usuario más usuales son: aplicaciones de analítica de datos, procesamiento analítico en línea (OLAP), herramientas de informes (*reporting*) y consultas (*query*), herramientas de analítica (minería de datos) y herramientas de visualización.

1.5 INTRODUCCIÓN A *BIG DATA* Y SU IMPACTO EN INTELIGENCIA DE NEGOCIOS

En las empresas se generan diariamente una ingente cantidad de datos, por lo que se necesitan herramientas para su procesamiento y análisis, su conversión en conocimiento útil y rentable para las organizaciones y que les ayuden en la toma de decisiones.

El término *Big Data* se refiere al conjunto de datos de gran volumen y complejos que las herramientas tradicionales, como las bases de datos relacionales, son incapaces de procesar en un rango de tiempo aceptables o dentro de un rango de costes razonables. Los problemas se producen en la extracción, búsquedas, flujos o movimientos, almacenamiento, procesamiento y análisis de datos, ya que las herramientas tradicionales, como ya se ha comentado, no pueden resolverlos.

Así, pues, han nacido en estos últimos años las tendencias y el concepto de *Big Data* para referirse a los conjuntos de datos voluminosos que exceden a la capacidad de manipulación de las herramientas tradicionales (normalmente, en el rango de terabytes, petabytes y magnitudes superiores). El volumen de datos, sin embargo, no es la única propiedad importante para su definición, tal como se verá posteriormente.

Las fuentes de datos son muy numerosas, pero, en la actualidad, además de los datos procedentes de las fuentes tradicionales, sistemas de información, bases de datos corporativas (transaccionales), y archivos que manipulan normalmente datos estructurados con formatos definidos, se alimentan de grandes volúmenes de datos que tienen formatos diferentes, no estructurados y semiestructurados. Estos grandes volúmenes de datos se conocen como *Big Data*

(macrodatos o datos masivos) y no se pueden procesar utilizando las herramientas tradicionales de bases de datos relacionales o, en caso de poder realizar esas tareas, los tiempos de proceso serían enormes. En consecuencia, se necesitan unas técnicas y herramientas diferentes de las tradicionales para su procesamiento eficiente y fiable. Los datos en la actualidad proceden de numerosas fuentes, tal como ya se viene comentando:

- Sistemas de información (ERP, CRM, SCM, GIS).
- Datos heredados (de bases de datos antiguas).
- Bases de datos relacionales y archivos.
- Correos electrónicos.
- Mensajes de texto.
- Archivos XML.
- Portales web.
- Medios sociales (*blogs*, redes sociales, wikis).
- Redes privadas.
- Multimedia (imágenes, sonido, video).
- Datos en *streaming* (flujo continuo de datos, texto, video, audio).
- Datos de máquinas (M2M, máquina a máquina).
- Sensores.
- Datos biométricos.
- Datos generados por humanos.

Los datos estructurados, o datos tipo tabla, proceden de las bases de datos y archivos tradicionales. El resto de los datos se conocen como datos no estructurados o datos semiestructurados y son muy difíciles de manejar por las herramientas tradicionales. Por estas razones ha aparecido la nueva tendencia denominada *Big Data*.

1.5.1 DEFINICIÓN DE *BIG DATA*

El término Big Data fue acuñado por Doug Laney[11], analista de la consultora Gartner, en 2001, para referirse a todo el conjunto de datos cuya cantidad o volumen —normalmente terabytes o petabytes—, velocidad y variedad exceden a la capacidad de manipular y procesar la información que tienen las herramientas tradicionales. Laney se refería no sólo al volumen de datos, sino a su velocidad de generación y a la gran variedad de formatos. Este modelo se conoce como el modelo de las 3V de Big Data:

- **Volumen:** Tamaño global del conjunto de datos, terabytes y petabytes, aunque ya muchas empresas generan exabytes de información.
- **Velocidad:** Tiempo utilizado en la generación de los datos así como la rapidez en que necesitan ser procesados: en tiempo real o casi en tiempo real.
- **Variedad**: Amplia gama de datos que pueden contener los conjuntos de datos que proceden de fuentes muy diversas: páginas web, texto, audio, video, fotografías, sensores, datos de máquinas, datos de dispositivos móviles, etcétera. Los datos se clasifican en tres tipos: **estructurados** (los datos de las bases de datos relacionales y heredadas, en formato tabla), **no estructurados** (audio, texto, fotografías), **semiestructurados** (archivos de texto, archivos XML, etcétera).

Posteriormente, en el capítulo 5 dedicado a *Big Data*, se verán otras características o dimensiones que vienen a configurar de un modo más preciso el concepto de *Big Data*, en los modelos conocidos como las 3V, las 4V, las 5V, las 7V e, incluso, las 8V como visión global de todas sus dimensiones.

Bernard Marr, uno de los grandes *gurús* de *Big Data*, considera naturalmente que se rastrean y almacenan datos de todo tipo y se tienen acceso a grandes volúmenes de datos; sin embargo, Marr plantea que "el valor real de *Big Data* no se encuentra en los grandes volúmenes de datos y sus tres propiedades fundamentales, sino a lo que podemos hacer con ellos. No es la cantidad de información lo que marca la diferencia, sino que se trata de nuestra capacidad para analizar series extensas y complejas que van más allá de todo lo que hubiéramos podido hacer anteriormente y su impacto global es análisis de esos datos, la gran capacidad para convertir enormes cantidades de datos complejos en valor".

1.5.2 TIPOS DE DATOS EN *BIG DATA*

- **Datos estructurados**: Datos tradicionales almacenados en filas y columnas (tablas) y que son los más empleados en archivos y bases de datos ordinarios de las organizaciones.
- **Datos semiestructurados:** No se ajustan a un esquema fijo y explícito; no se limitan a campos determinados, mantienen marcadores para separar elementos. Tienen información poco regular, de forma que no puede ser gestionada de un modo estándar; utilizan lenguajes de marcación de hipertexto o de marcas extensibles. Ejemplos de estos datos son los documentos XML, HTML, datos de sensores, etcétera.
- **Datos no estructurados:** Son los datos más complejos; se presentan en formatos que no pueden ser fácilmente manipulados por las bases de datos relacionales: archivos Word, pdf, ppt, hojas de cálculo, documentos multimedia, audio, voz, video, fotografías, correos electrónicos.

1.5.3 EL IMPACTO DE *BIG DATA* EN INTELIGENCIA DE NEGOCIOS

Big Data es la gestión y análisis de grandes volúmenes de datos que, normalmente, no se pueden tratar con los métodos tradicionales, no sólo por su volumen, sino por los formatos de los datos (no estructurados y semiestructurados en su gran porcentaje) y la velocidad a la que se generan dichos datos. La mayoría de los datos recopilados por las organizaciones, hasta que apareció la tendencia de *Big Data*, eran datos transaccionales que podían alojarse fácilmente en filas y columnas (tablas) de los sistemas de gestión de bases de datos relacionales tradicionales. La explosión de datos que se ha producido en los últimos años ha dado origen a una avalancha de datos —mensajes de texto, correos electrónicos, mensajes de redes sociales, archivos de audio, de fotografías, de videos, datos generados por sensores (utilizados en medidores inteligentes de energía, agua, eléctricos), datos biométricos—. Estos datos, en su gran mayoría, pueden ser no estructurados —no siguen el formato de tabla, filas, columnas— o semiestructurados —datos de archivos web, programación— y, por consiguiente, no son adecuados para soluciones de bases de datos relacionales, que, como ya se ha señalado, organizan sus datos en formatos de filas y columnas.

En la actualidad, el término *Big Data* describe a estos grandes conjuntos de datos y su almacenamiento, gestión y análisis es una de las grandes tareas estratégicas en las organizaciones. La inteligencia tradicional se ha preocupado siempre del procesamiento y análisis de los datos estructurados, sin embargo, una nueva tendencia moderna de Inteligencia de Negocios está surgiendo para tratar de conservar los principios fundamentales de la inteligencia empresarial para su apoyo en la toma de decisiones, pero tomando como soporte los *Big Data* y sus herramientas más significativas, no sólo en sus infraestructuras físicas sino, y sobre todo, en el análisis de esos grandes volúmenes de datos, con el objeto de facilitar las tareas de toma de decisiones de los empleados corporativos.

Los *Big Data* (a los que dedicaremos un capítulo exclusivo) se producen en grandes cantidades y a una mayor velocidad que los datos estructurados tradicionales. En este libro, dedicaremos e integraremos la gestión y análisis de *Big Data* junto con los restantes datos históricos y actuales no estructurados, para conformar herramientas de ayuda a la toma de decisiones en la vida diaria de las organizaciones y empresas modernas y embebidas en la economía digital y en la transformación digital de las mismas con las tecnologías, técnicas y métodos que iremos describiendo a lo largo del libro.

Las herramientas actuales de Inteligencia de Negocios han de dar soporte a *Big Data* y deberán cumplir características específicas para su correcto tratamiento. Así, deberán cumplir con las siguientes funcionalidades y características:

- Carga y gestión de grandes volúmenes de datos de forma eficiente (Volumen).
- Facilitar la integración de un mayor número de fuentes y soportar el amplio abanico de formatos existentes (Variedad).
    - Nuevos formatos: XML, JSON, BD's NoSQL, API's de servicios web.

- Facilitar el diseño de procesos de verificación de la calidad (Veracidad).
- Integración de fuentes en tiempo real (Velocidad), tanto recolección como integración en tiempo real.

En resumen, se necesitan herramientas que simplifiquen la aplicación de las técnicas para el desarrollo de aplicaciones de Inteligencia de Negocio —*Big Data*— de forma lo más eficiente y efectiva posible. La integración de las soluciones de *Big Data* en los procesos de Inteligencia de Negocios será una de las grandes tareas por realizar en las empresas y, a lo largo del libro, veremos herramientas y técnicas que integren *Big Data* en inteligencia de negocios de un modo eficiente y rentable.

1.6 ARQUITECTURA DE INTELIGENCIA DE NEGOCIOS CON INTEGRACIÓN DE *BIG DATA*

La arquitectura *Big Data* está compuesta generalmente por cinco capas: recolección de datos, almacenamiento, procesamiento de datos, visualización y administración. Esta arquitectura no es nueva, sino que ya es algo generalizado en las soluciones de *Business Intelligence* que existen hoy en día. Sin embargo, debido a las nuevas necesidades, cada uno de estos pasos ha ido adaptándose y aportando nuevas tecnologías, a la vez que han abierto nuevas oportunidades.

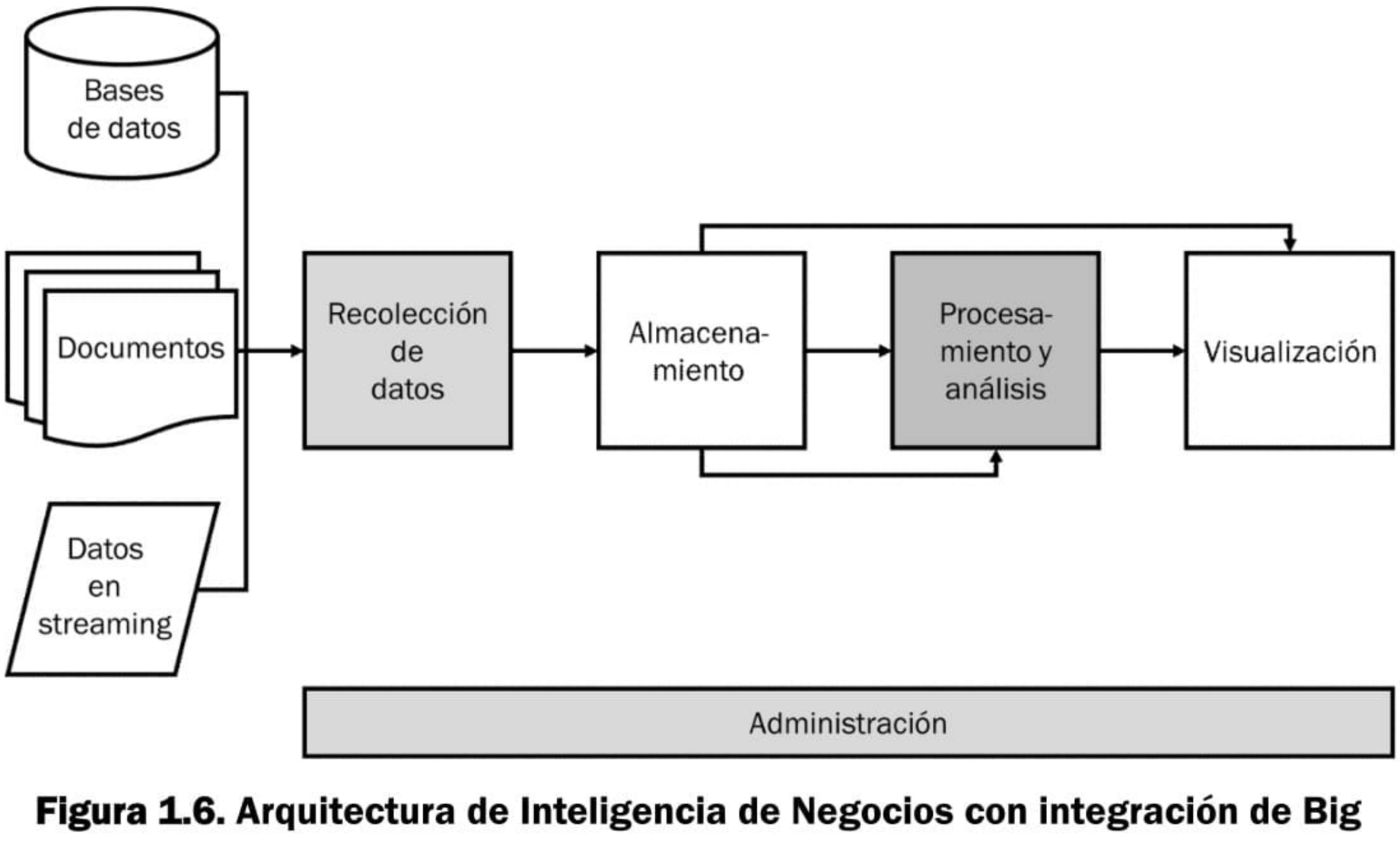


Figura 1.6. Arquitectura de Inteligencia de Negocios con integración de Big Data
Fuente Chunmei Duan, (2014)

En términos generales, una arquitectura de *Big Data* (Duan, 2014)[12] está constituida por cinco componentes: recolección de datos, almacenamiento, procesamiento de datos, visualización y administración (gestión). Además, cada

uno de estos componentes ha ido añadiendo nuevas tecnologías, las cuales dependen de las necesidades que se vayan dando, y también es necesaria su adaptación para dar una solución eficiente a las empresas, actualmente.

En la **figura 1.6** se puede observar el flujo que la información tendría en una arquitectura *Big Data*, con orígenes de datos diversos —bases de datos, documentos o datos recibidos en *streaming*— que se reciben y almacenan a través de la capa de recolección de datos, con herramientas específicamente desarrolladas para tal función. Los datos recibidos pueden procesarse, analizarse y/o visualizarse tantas veces como haga falta y lo requiera el caso de uso específico.

1.6.1 RECOLECCIÓN DE DATOS

En esta etapa, el sistema debe conectarse a sus diferentes fuentes de información y extraerla para obtener los datos que posteriormente serán almacenados, procesados, analizados y visualizados sus resultados. Las herramientas o métodos de recolección de datos pueden dividirse en dos grupos, según cómo se conecten al origen de los datos:

1. ***Batch* o por lotes**: se conectan de manera periódica a la fuente de datos (archivos o bases de datos) buscando nueva información. Generalmente, se usan para conectarse a sistemas de archivos (ficheros) o bases de datos, buscando cambios desde la última vez que se conectaron. Una herramienta para migrar datos periódicamente (una vez al día, por ejemplo) de una base de datos a otra es un ejemplo de este tipo de recolección.
2. ***Streaming* o transmisión en tiempo real**: están conectados de manera continua a la fuente de datos, descargando información cada vez que ésta transmite. Suele utilizarse para monitorización de sistemas (para aumentar la seguridad y la detección de fallos), de conjuntos de sensores o para conectarse a redes sociales y descargar información en tiempo real.

Actualmente, las herramientas han evolucionado de manera que muchas de ellas ya pueden usarse de las dos maneras, tanto como para descargarse información en *streaming* como con procesos *batch*.

En esta etapa, los datos pueden sufrir algún tipo de proceso o cambio si la aplicación así lo requiere, por ejemplo, el filtrado de información no deseada o el formateo con el que se guardará finalmente en el sistema de almacenamiento.

1.6.2 ALMACENAMIENTO

La capa de almacenamiento tiene, de modo general, dos elementos básicos: el sistema de archivos (ficheros) y las bases de datos. Hasta hace poco, los sistemas tradicionales de tratamiento de la información se centraban

principalmente en las bases de datos relacionales, pero, debido a los grandes volúmenes de datos manejados en la actualidad junto con la gran variedad (estructurados y no estructurados) y la velocidad de descarga de los datos (los sistemas *Big Data*), los sistemas de almacenamiento de información se han tenido que adaptar a estas nuevas características. Los sistemas de almacenamiento tradicionales de bases de datos relacionales no pueden manejar estos nuevos tipos de datos y sus grandes volúmenes. Se busca la mayor variedad posible —las bases de datos suelen ser poco flexibles—, los sistemas de archivos han cobrado mayor importancia en la manipulación de *Big Data*, mediante sistemas de archivos HDFS que manipulan de un modo más fácil y sencillo las nuevas características de volumen, velocidad y variedad (esencialmente) para el almacenamiento de la información.

SISTEMAS DE ARCHIVOS Y SISTEMAS DE ARCHIVOS DISTRIBUIDOS

Los sistemas de archivos (ficheros) son una parte fundamental de la arquitectura *Big Data*, ya que es por encima de ellos que el resto de las herramientas están construidas. Además, el hecho de querer trabajar con datos no estructurados los hace aún más importantes, ya que son el medio principal para trabajar con este tipo de información.

Adicionalmente; un objetivo que buscan los sistemas *Big Data* es la **escalabilidad**, es decir, un sistema que pueda variar su tamaño (ya sea aumentándolo o disminuyéndolo) según las necesidades y que esto no afecte al rendimiento general de todo el sistema. Esta necesidad fue la que motivó la aparición de los sistemas de archivos distribuidos, que consisten en una red o *clústeres* de computadores (o nodos) interconectados, que están configurados para tener un sólo sistema de archivos lógico.

En las arquitecturas *Big Data* más recientes se está intentando aprovechar lo mejor de los dos paradigmas. Se crea un sistema de almacenamiento (ya sea un sistema de archivos distribuido o una base de datos NoSQL) para almacenar la información no estructurada en grandes volúmenes de datos y, posteriormente, se almacenan los resultados de los procesos y análisis realizados sobre estos datos en un sistema SQL, obteniendo una mayor velocidad de respuesta al consultar los resultados.

1.6.3 PROCESAMIENTO Y ANÁLISIS

Una vez se tienen los datos almacenados, el siguiente paso en un sistema *Big Data* es explotar la información para llegar a los resultados deseados. Las herramientas de análisis y procesamiento de información han evolucionado considerablemente, especialmente aquellas que trabajan sobre datos no estructurados. Una vez que se tienen almacenados los datos, se ha de obtener

conocimiento o valor mediante el procesamiento y análisis de toda la información almacenada.

La necesidad de crear nuevas aplicaciones, y que éstas ya estén adaptadas a los sistemas de almacenamiento más recientes (como los comentados en el punto anterior, los sistemas distribuidos y las bases de datos NoSQL), ha promovido la aparición de nuevos paradigmas para el análisis de datos y la presentación de resultados que veremos en el siguiente apartado.

En la actualidad, y debido a la existencia de grandes volúmenes de datos (*Big Data*), tanto estructurados como no estructurados o semiestructurados, se ha de realizar el establecimiento de conexiones entre los almacenes de datos tradicionales (bases de datos, *Data Warehouses* y *Data Marts*) con las plataformas de *Big Data*. Se requieren unos conectores entre las bases de datos relacionales y los almacenes de datos, y las bases de datos NoSQL y *"en memoria"* (*in-memory*), así como los sistemas de archivos distribuidos de las plataformas Hadoop y Spark, tales como MapReduce, HDFS, Hive, HBase, etcétera.

1.6.4 VISUALIZACIÓN

El componente de visualización, como tal, es el que menos ha cambiado respecto de las arquitecturas más tradicionales, aunque sí han cambiado radicalmente las herramientas de visualización modernas, como se verá en el capítulo 7 (por ejemplo, las "narraciones de datos" o *data storytelling).* Los datos originales se han convertido en conocimiento y sus resultados se presentan a las organizaciones para su estudio y toma de decisiones correspondiente, mediante herramientas de visualización de datos adecuadas al nivel organizativo que se trate.

Como se ha comentado en el apartado de **almacenamiento**, los resultados por visualizar del procesamiento se acostumbran a consultar sobre bases de datos relacionales o SQL, ya que son las que ofrecen un menor tiempo de respuesta.

1.6.5 INTEGRACIÓN DE *BIG DATA* EN SISTEMAS DE INTELIGENCIA DE NEGOCIOS

Un sistema de Inteligencia de Negocios (IN/BI) representa una amplia categoría de aplicaciones, tecnologías y procesos que tienen como objetivo la recolección, almacenamiento, acceso y análisis de datos para la ayuda a los usuarios en la toma de mejores decisiones.

Una arquitectura de Inteligencia de Negocios tradicional tiene, normalmente, los componentes descritos en las metodologías de Turban y Laudon (apartado 1.7). Sin embargo, la expansión en los últimos años de las tecnologías de *Big Data* obliga a su consideración e integración con las infraestructuras tradicionales

de datos estructurados, tales como las bases de datos relacionales. En consecuencia, los componentes de una arquitectura de Inteligencia de Negocios son:

- Fuentes de datos.
- Infraestructuras de datos.
- Analítica de Datos.
- Gestión empresarial y de usuario.
- Interfaces de usuario.

Por otra parte, cada día se necesita utilizar más los ***Big Data***. Una La infraestructura de Inteligencia de Negocios (propiamente dicha) incluye bases de datos relacionales tradicionales, pero, sobre todo, **almacenes o repositorios de datos** (*Data Warehouses* y *Data Marts*), **almacenes de datos Hadoop y Spark** (*Big Data*) y ***bases de datos en memoria.*** Una infraestructura de Inteligencia de Negocios moderna debe contar con capacidades y herramientas para analizar grandes cantidades de datos y de diferentes formatos, procedentes de múltiples fuentes; estas herramientas han de ser fáciles de utilizar para realizar informes (reportes) y consultas para usuarios ordinarios, gerentes, directivos, administrativos, y herramientas de analítica más sofisticadas para usuarios avanzados como analistas de datos, ingenieros de datos o científicos de datos.

ALMACENES DE DATOS: HADOOP Y SPARK

La profusión de grandes volúmenes de datos en las corporaciones (*Big Data*) requiere de nuevas herramientas para su gestión. La razón es que *Big Data* son datos de diferentes formatos y las bases de datos relacionales de los almacenes de datos sólo están preparadas para lecturas y consultas de datos estructurados en forma de tablas (organizados en filas y columnas). La manipulación de datos no estructurados y semiestructurados requiere de nuevas infraestructuras de almacenamiento: las más usuales Hadoop y Spark (una versión avanzada de Hadoop para procesamientos de datos en tiempo real), bases de datos "en memoria" (*in-memory*) o las clásicas bases de datos analíticas MPP (siempre que puedan procesar datos no estructurados y semiestructurados).

Hadoop es un marco de trabajo creado, diseñado y actualizado por la Fundación Apache, que facilita el procesamiento distribuido y paralelo de grandes volúmenes de datos. El sistema, en el que se profundizará más adelante, se apoya en la distribución de los datos en miles de nodos de procesamiento más pequeños, que ya pueden operar sobre datos en menor tamaño y más fáciles de analizar. Hadoop es el sistema utilizado, mayormente, por las grandes compañías como Google, Facebook, Amazon, LinkedIn, empresas multinacionales como líneas aéreas (Iberia, Aeroméxico, Avianca o Latam) o empresas del sector industrial, turístico, negocios, etcétera.

Hadoop consta de dos componentes clave: HDFS (un sistema de archivos para almacenamiento de datos) y MapReduce, un algoritmo para procesamiento de datos en paralelo y de alto rendimiento. HDFS enlaza juntos los sistemas de archivos en los numerosos nodos de un clúster Hadoop para convertirse en un sistema de archivos procesados con el algoritmo MapReduce. El componente MapReduce es un algoritmo que facilita el procesamiento en paralelo de los miles de nodos en donde se almacenan los datos del *clúster* Hadoop, que puede procesar grandes cantidades de datos de cualquier tipo de formato, tanto estructurados de las bases de datos relacionales como no estructurados, texto, audio, video, fotografías o datos de redes sociales o de la Web.

Los datos de *Big Data* se almacenan en bases de datos no relacionales (conocidas como NoSQL), que proporcionan un acceso rápido a los datos almacenados en los sistemas de archivos HDFS. La base de datos estándar de Hadoop, HBase, será la más utilizada para la ejecución de aplicaciones a gran velocidad, aunque, como veremos más adelante, existe una gran cantidad y variedad de bases de datos NoSQL.

El marco de trabajo Hadoop se ejecuta en un *clúster* (grupo) de servidores de bajo coste, de forma que se pueden añadir o eliminar procesadores a medida que se necesitan. Las empresas utilizan Hadoop para analizar grandes volúmenes de datos y también como área de acondicionamiento de datos no estructurados y semiestructurados antes de que se carguen en el almacén de datos.

Los proveedores de *software*, tanto los de *software* propietario (IBM, Oracle, Microsoft, Hewlett-Packard) como los de fuente abierta (Cloudera, Pentaho, Jaspersoft), tienen sus propias distribuciones de *software*, de forma que ofrecen herramientas para mover datos dentro y fuera del sistema Hadoop o bien para analizar los datos dentro de Hadoop.

Bases de datos en memoria

Otro método para realizar el análisis de *Big Data* es utilizar tecnologías de computación en memoria (*in-memory*), mediante el uso de bases de datos en memoria. En estas bases de datos, los datos se almacenan en memoria (los sistemas de bases de datos relacionales tradicionales utilizan sistemas de almacenamiento en disco), con lo que se producen accesos muy rápidos.

Los usuarios acceden a los datos almacenados en memoria principal del sistema, eliminando así los tiempos de lectura y recuperación de datos en una base de datos tradicional, basada en discos que necesitan periodos de tiempo para transportar los datos entre la unidad de disco y la memoria. De este modo, al almacenar todos los datos en la memoria central, se acortan drásticamente los tiempos de respuesta a consultas y almacenamiento.

Las tecnologías de computación en memoria facilitan que los grandes conjuntos de datos, como los equivalentes al tamaño de un almacén de datos (*Data Warehouse*) o almacén de datos corporativo (*Data Mart*), residan

totalmente en memoria, ahorrando grandes cantidades de tiempo que se necesitan en las bases de datos relacionales, reduciendo los tiempos a velocidades de segundos y casi en tiempo real.

El inconveniente de los sistemas de procesamiento en memoria es la tecnología de *hardware* de computadoras, que requiere procesadores de gran velocidad, procesamiento multinúcleo, así como el alto coste de las infraestructuras *hardware* y la necesidad de un *software* muy especializado. Sin embargo, estas tecnologías ayudan a las empresas a optimizar el uso de la memoria y acelerar el rendimiento del procesamiento, a la vez que se reducen los costes.

Curiosamente, el primer proveedor de bases de datos en memoria fue SAP —reconocido como el primer proveedor mundial de *software* de gestión corporativo—, que, en alianza con grandes fabricantes de *hardware* como IBM, diseñó hace unos años HANA (*High Performance Analytics Appliance*), un *software* para procesamiento en memoria. A SAP e IBM han seguido otros grandes proveedores, como Oracle con sus soluciones Oracle Exalytics. Estas soluciones de almacenamiento en memoria proporcionan un conjunto de componentes de *software* integrado, que incluyen al *software* de la base de datos integrado con *software* de analítica especializada (*analytics*), y que corren sobre las arquitecturas de computación en memoria.

1.7 VISIÓN GERENCIAL DE LA INTELIGENCIA DE NEGOCIOS

Una visión general de un entorno de Inteligencia de Negocios ha de tener presente los componentes de *hardware, software* y de gestión empresarial que ofrecen los diferentes proveedores comerciales y que las empresas instalan y despliegan para conseguir sus objetivos. Existen diferentes arquitecturas de Inteligencia de Negocios de los proveedores de soluciones (Oracle, SAS, IBM, Microstrategy, SAP, Microsoft) y modelos de expertos y consultores de Inteligencia de Negocios, así como diferentes artículos de investigación. En nuestro caso, analizaremos las arquitecturas definidas por Turban (2014) y Laudon (2014) que tienen componentes y funcionalidades similares, aunque difieren en el número de componentes, y realizaremos una propuesta específica que trata de contener la mayoría de las características de ambas metodologías. También, como casos de estudio analizaremos la arquitectura de BI de algunos de los grandes proveedores de soluciones.

1.7.1 METODOLOGÍA TURBAN

Turban *et al* (2014) en su obra sobre Inteligencia de Negocios (referencia obligada en la materia) considera que un sistema de BI tiene cuatro componentes importantes:

- *Data Warehousing* —almacenamiento de datos, con sus fuentes de datos—.

- *Business Analytics* —**analítica de negocios**, colección de herramientas para manipulación, minería y análisis de los datos de los almacenes de datos o *Data Warehouse*—.
- *Business Performance Management* —BPM, **gestión del rendimiento (desempeño) del negocio,** mediante la monitorización y análisis del desempeño—.
- **Interfaces de usuario** por ejemplo, cuadros de mando o *dashboards*—.

En la arquitectura se plantean también los roles o responsables de cada una de las capas o componentes. Así, la capa de almacenamiento de datos es responsabilidad principal de la plantilla técnica (ingenieros de sistemas, ingenieros informáticos, ingenieros de *software*, etcétera) y son los responsables de la construcción de *Data Warehouse* y de *Data Marts* (almacenes de datos); la capa de Analítica de Negocios es responsabilidad de los usuarios de negocios, que accederán a los almacenes de datos y también actuarán en la gestión de los datos; los directivos y ejecutivos utilizarán las herramientas de BPM y las interfaces de usuario; y, por último, cualquier usuario puede utilizar la interfaz de usuario, tales como navegadores, portales, cuadros de mando (*dashboards*), hojas de cálculo, etcétera.

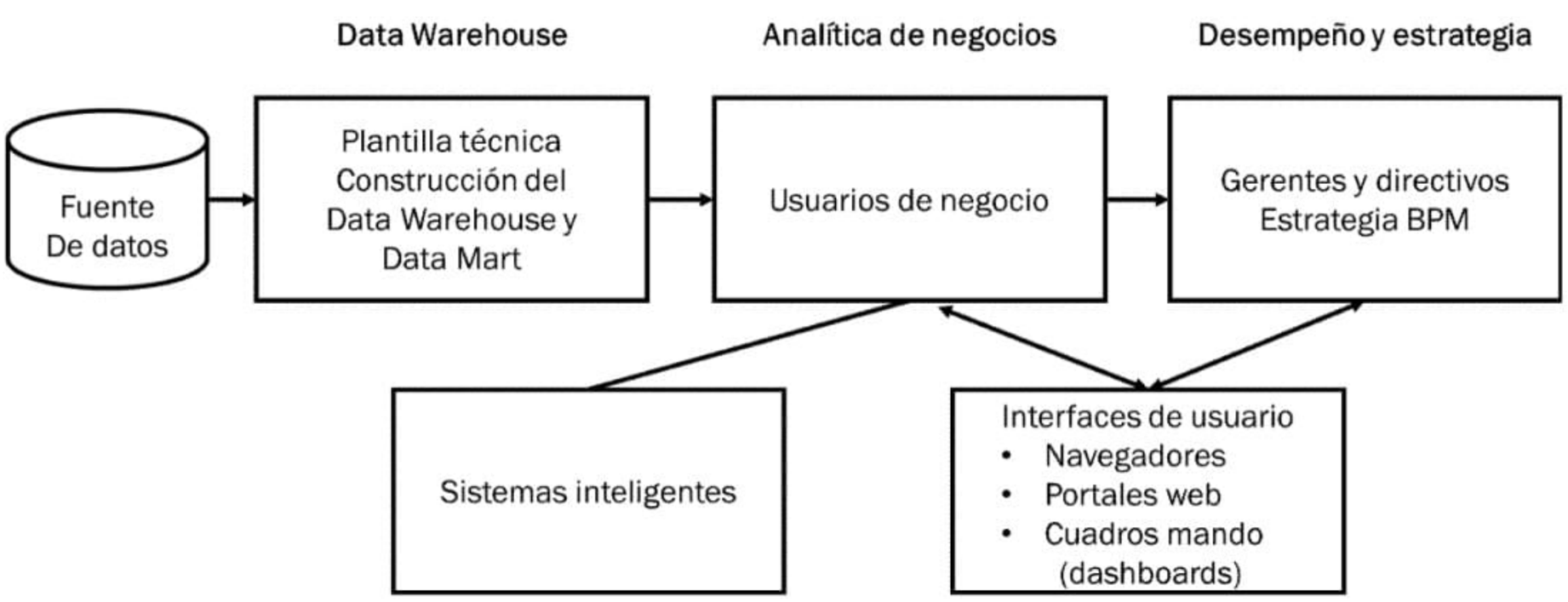


Figura 1.7. Arquitectura de inteligencia de negocios (BI).
Fuente: Turban et al (2011 y 2014)[13],[14] (Adaptada del original)

1.7.2 METODOLOGÍA LAUDON

Laudon (2014), dos de los autores más reconocidos en el área de Sistemas de Información, plantea una arquitectura de Inteligencia de Negocios, apoyada en las sucesivas ediciones de su obra *Sistemas de Información*, que consta de seis componentes en lo que ellos denominan **entorno de inteligencia de negocios**.

- Fuentes de datos (datos del entorno de negocio)
- Infraestructura de Inteligencia de Negocios

- Herramientas de Analítica de Negocios (*Business Analytics*)
- Métodos y usuarios gerenciales
- Plataformas de entrega de información: sistemas de información, MIS, DSS y ESS
- Interfaces de usuario

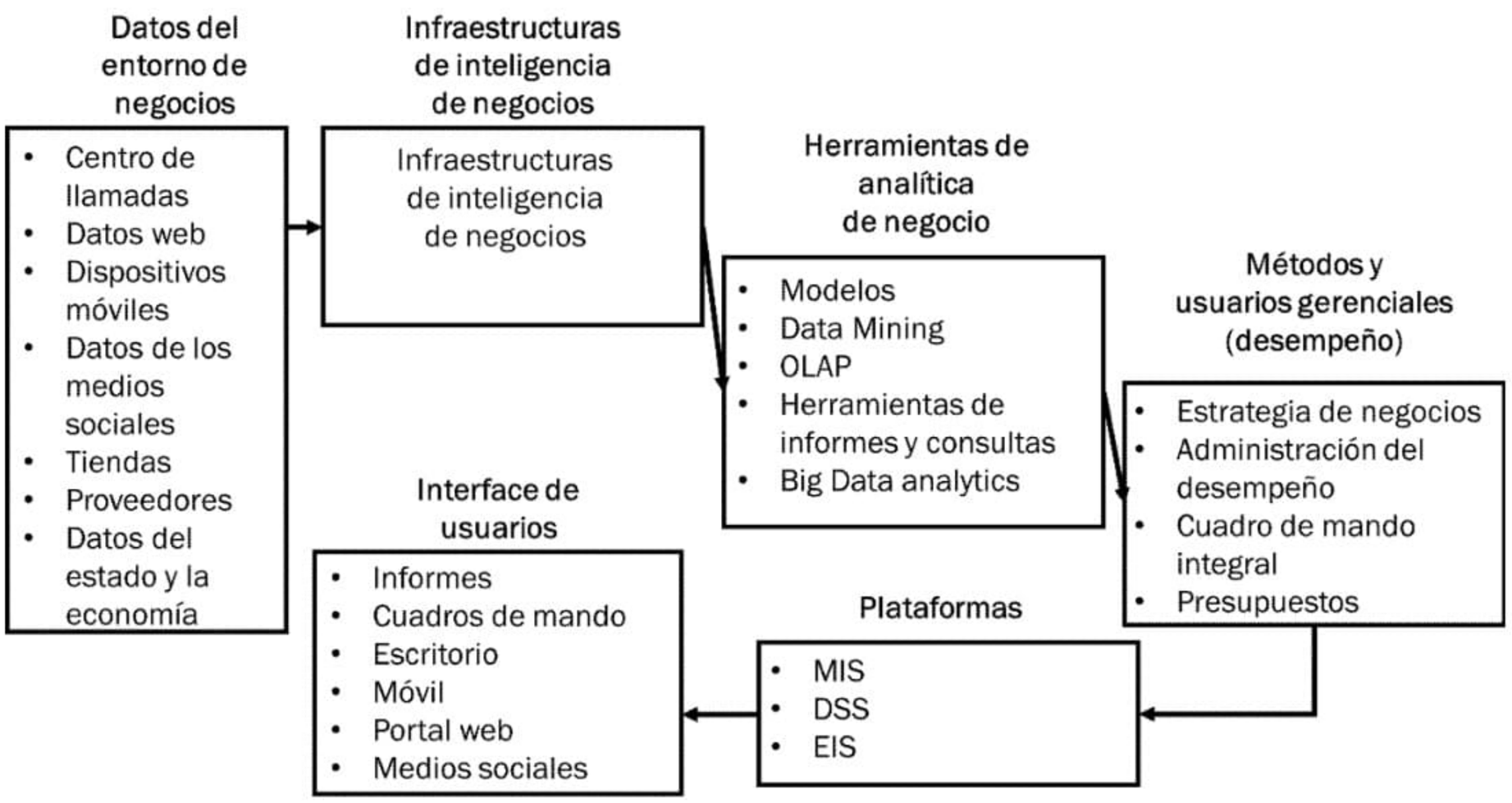


Figura 1.8. Arquitectura de Inteligencia de Negocios (BI)
Fuente: Laudon y Laudon (2014)[15] (Adaptada del original)

Datos del entorno de negocios (*Data from Business Environment*): centros de llamada (*call centers*), datos de la Web, dispositivos móviles, datos de medios sociales (*social media*), almacenes, proveedores, datos económicos y gubernamentales. En los sistemas modernos, los negocios deben tratar datos estructurados, no estructurados y semiestructurados procedentes de numerosas fuentes, incluyendo *Big Data*. Los datos necesitan ser integrados de modo que puedan ser analizados y usados por los empleados tomadores de decisiones.

Infraestructura de Inteligencia de Negocios (*Business Intelligence* Infrastructure): los componentes fundamentales de Inteligencia de Negocios son las bases de datos, *Data Warehouses* (almacenes de datos), *Data Marts*, plataformas de analítica y, cada día con mayor frecuencia, los *Big Data* en infraestructuras como Hadoop o Spark, fundamentalmente, y bases de datos NoSQL y "en memoria" (*in-memory*).

Herramientas de Analítica de Negocios (*Business Analytics Toolset*): conjunto de herramientas de *software* que se utilizan para analizar productos y producir informes (reportes) que respondan a las necesidades de los administradores, directivos y ejecutivos. Estos informes deberán adaptarse a los indicadores clave de desempeño o rendimiento KPI de las organizaciones. Las herramientas más

utilizadas son: modelos, minería de datos, **OLAP**, herramientas de visualización (cuadros de mando), herramientas de *reporting* y *query*, analítica de *Big Data.*

Desempeño y estrategia (*Managerial Users and Methods*): los gerentes y directivos (cada vez con mayor frecuencia, empleados de cualquier nivel) deben utilizar el análisis de datos para definir objetivos estratégicos de negocio y especificar la medición del progreso corporativo. Se han de incluir cuadros de mando y cuadros de mando integral (*balanced scorecards*) centrados en KPI, soluciones de gestión del rendimiento o desempeño del negocio (BPM, *Business Performance Management)* y, también, se debe considerar la gestión de los procesos de negocio (BPM, *Business Process Management*).

Plataformas (*Platform*): los resultados de Inteligencia de Negocios y Analítica de Negocios se entregan a los gerentes y empleados en una variedad de formas, dependiendo de lo que necesitan conocer para realizar sus trabajos. Las plataformas tradicionales son los sistemas de información en sus diferentes modalidades: ERP, CRM, SCM, GIS. Los sistemas de información antiguos trabajan como sistemas independientes. En la actualidad, existen soluciones integradas (*suites*) que permiten la conexión entre diferentes sistemas de información y las herramientas típicas de Inteligencia de Negocios, que iremos analizando en la obra, facilitando la integración de toda la información circulante por las compañías.

Interfaces de usuario (*User Interface)*: los empleados de negocio se conectan hoy en día de modo *omnicanal* con todo tipo de dispositivos, preferentemente dispositivos móviles (teléfonos inteligentes y tabletas) y a través de redes sociales. Por estas razones, las entregas de informes y consultas se deben adaptar a los dispositivos móviles, así como a los medios sociales y, en consecuencia, los informes, consultas, presentaciones en cuadros de mando, infografías, entre otros, deberán poderse presentar en cualquier tipo de dispositivo, móvil, escritorio, Web, etcétera.

1.8 ANALÍTICA DE NEGOCIOS (*BUSINESS ANALYTICS*)

Analítica de Negocios *(Business Analytics*) es un conjunto de técnicas y herramientas de *software* utilizadas para analizar los datos debidamente preparados y producir informes para responder a las cuestiones y consultas de los gerentes y directivos, y seguir el proceso del negocio mediante indicadores clave de rendimiento o desempeño, KPI. En resumen, la Analítica de Negocios o Analítica Empresarial es un conjunto de herramientas de *software* diseñadas para crear informes, realizar consultas y analizar datos, como minería de datos, de texto y Web, y herramientas de estadística avanzada.

Las aplicaciones y tecnologías utilizadas en la Analítica de Negocios son (Turban 2011: 20-21):

- "**Informes (reportes) y consultas (*reports* y *queries*).** La Analítica de Negocios incluye informes estáticos y dinámicos, análisis multidimensional (OLAP) o cubos OLAP.
- **Minería de Datos, minería web y minería de texto, así como otras herramientas matemáticas y estadísticas específicas.** La Minería de Datos es un proceso de búsqueda de relaciones desconocidas (patrones) o información en grandes bases de datos, almacenes de datos (y en la actualidad, sistemas de *Big Data*), que utilizan herramientas inteligentes, tales como computación neuronal, métodos estadísticos avanzados, técnicas de Analítica predictiva (y, cada vez con mayor frecuencia, técnicas de aprendizaje automático y de Inteligencia Artificial)".

La Analítica de Negocios es un conjunto de herramientas de *software* que se utilizan para analizar y comprender los datos y producir informes que respondan a cuestiones planteadas por los directivos y ejecutivos para seguir el progreso de los negocios, utilizando los indicadores KPI. Las herramientas de analítica entregan información correcta y fiable y cuando se necesita en tiempo real, que le ayudan a los usuarios de la empresa a comprender con rapidez la información solicitada u ofertada, con el objetivo de tomar las acciones oportunas en su toma de decisiones.

1.8.1 GESTIÓN DEL DESEMPEÑO (RENDIMIENTO) DEL NEGOCIO

Una vez analizados los datos empresariales, los gerentes y administradores necesitan utilizar una variedad de métodos gerenciales que definan los objetivos estratégicos de negocio y especifiquen cómo se medirán sus progresos (Laudon 2014). Técnicamente, se conocen como administración del rendimiento (desempeño) gerencial (**BPM,** *Business Performance Management)* y, también, administración del desempeño corporativo (**CPM**, *Corporate Performance Management*). Turban (2011) considera que BPM es un porfolio de aplicaciones —emergente en esos primeros años de la década— y metodologías que contienen la evolución de la arquitectura de BI y sus herramientas nucleares.

BPM comprende la monitorización, medición y compartición de ventas, beneficios, costes, rentabilidad y otros indicadores de rendimiento, mediante la gestión y realimentación. La gestión del desempeño empresarial se realiza con la utilización de la popular metodología **BSM** (*Balanced Scorecard Methodology*) cuya traducción a nuestro idioma es **"Cuadro de Mando Integral"**. Tanto BPM como BSM se centran en los indicadores clave de desempeño (KPI, *Key Performance Indicator*) en el entorno general de negocios y prestan atención especial a la competencia.

Laudon (2014) considera que, sin una supervisión de la alta dirección, el proceso de análisis, los informes y consultas pueden no ayudar en la manera debida a la toma de decisiones. A medida que los procesos de análisis se hagan más inteligentes, como comentaremos más adelante, con el uso de aplicaciones

de Inteligencia Artificial, la toma de decisiones será más fácil para los directivos, ya que, además de las características propias de los usuarios, se añadirán las recomendaciones y decisiones de los agentes inteligentes (los robots o *bots* virtuales).

1.8.2 INTERFACES DE USUARIO

Los directivos, gestores o administradores y los usuarios en general hoy se encuentran conectados globalmente con diferentes canales de comunicación (escritorio de PC y portátiles *laptops*, teléfonos inteligentes, tabletas, portales web). Sin embargo, la toma de decisiones es más fácil cuando se utilizan verdaderas herramientas de visualización de datos integradas en los diferentes medios de acceso.

Las herramientas de visualización más empleadas son los cuadros de mando o tableros de control (*dashboards* y *scorecards*), que permiten la presentación de resultados e informes apoyados en los indicadores clave de rendimiento, KPI. Además de estas herramientas, se ofrece información en las presentaciones, en redes sociales como Facebook, Twitter, LinkedIn, o en sistemas de mensajería instantánea como WhatsApp, Telegram, Viber, Skype, Facebook Messenger, Line, etcétera.

Los cuadros de mando o tableros de control, como señala su nombre, son similares a los tableros de control de los automóviles y proporcionan una visión global o específica de las métricas de desempeño corporativo, indicadores de diferentes tipos, tendencias, excepciones. Los cuadros de mando presentan gráficos que, a primera vista, muestran el rendimiento real comparado con las métricas deseadas y permiten presentar una visión del estado real de la organización.

Una de las herramientas que comienza a utilizarse ya con relativa frecuencia, y que en 2018 y en años siguientes se popularizará en las organizaciones y empresas, son los *bots* (robots virtuales) y en especial los *chatbots* (*bots* conversacionales). Los *bots* que se podrán utilizar son los generalistas, tales como Siri de Apple, Cortana de Microsoft, Google Assistant de Google, Alexa de Amazon, etcétera, o bien bots personalizados de la propia empresa, que le permitirán su conexión directa a cualquier tipo de aplicación o red social, mediante el desarrollo del *bot* empresarial que cada día es más fácil diseñar, desarrollar e implantar.

Además de las técnicas inteligentes de Inteligencia Artificial antes citadas, y otras de ayuda a la toma de decisión, comienzan ya a integrarse en las herramientas de visualización técnicas de realidad aumentada y realidad virtual (o mixta, integración de ambas).

Los sistemas de información gerencial proporcionan también medios de presentación; además de ser herramientas de recolección y entrega de datos, hay

algunos sistemas de información especiales que ofrecen muy buenas herramientas de visualización, como es el caso de los sistemas de información geográfica (GIS).

1.8.3 PLATAFORMAS DE ANALÍTICA

Los proveedores de bases de datos comerciales (relacionales, Hadoop, NoSQL, "en memoria") han desarrollado plataformas analíticas especializadas de alta velocidad utilizando tecnologías relacionales, no relacionales y "en memoria", que se optimizan para analizar grandes volúmenes de datos (*Big Data*). Las plataformas analíticas incluyen herramientas para gestión de las diferentes bases de datos y almacenes de datos.

Una infraestructura moderna de Inteligencia de Negocios debe utilizar todas las tecnologías actuales y tener capacidad para gestionar y analizar grandes volúmenes de datos *(Big Data*) de tipos diferentes (estructurados, no estructurados y semiestructurados) procedentes de múltiples fuentes, externas e internas. Las herramientas de consulta y de reportes (*query* y *reporting*) deben ser fáciles de utilizar por todo tipo de usuarios, tanto empleados ordinarios, ejecutivos y directivos, como usuarios avanzados, tales como analistas de datos, ingenieros de datos o científicos de datos.

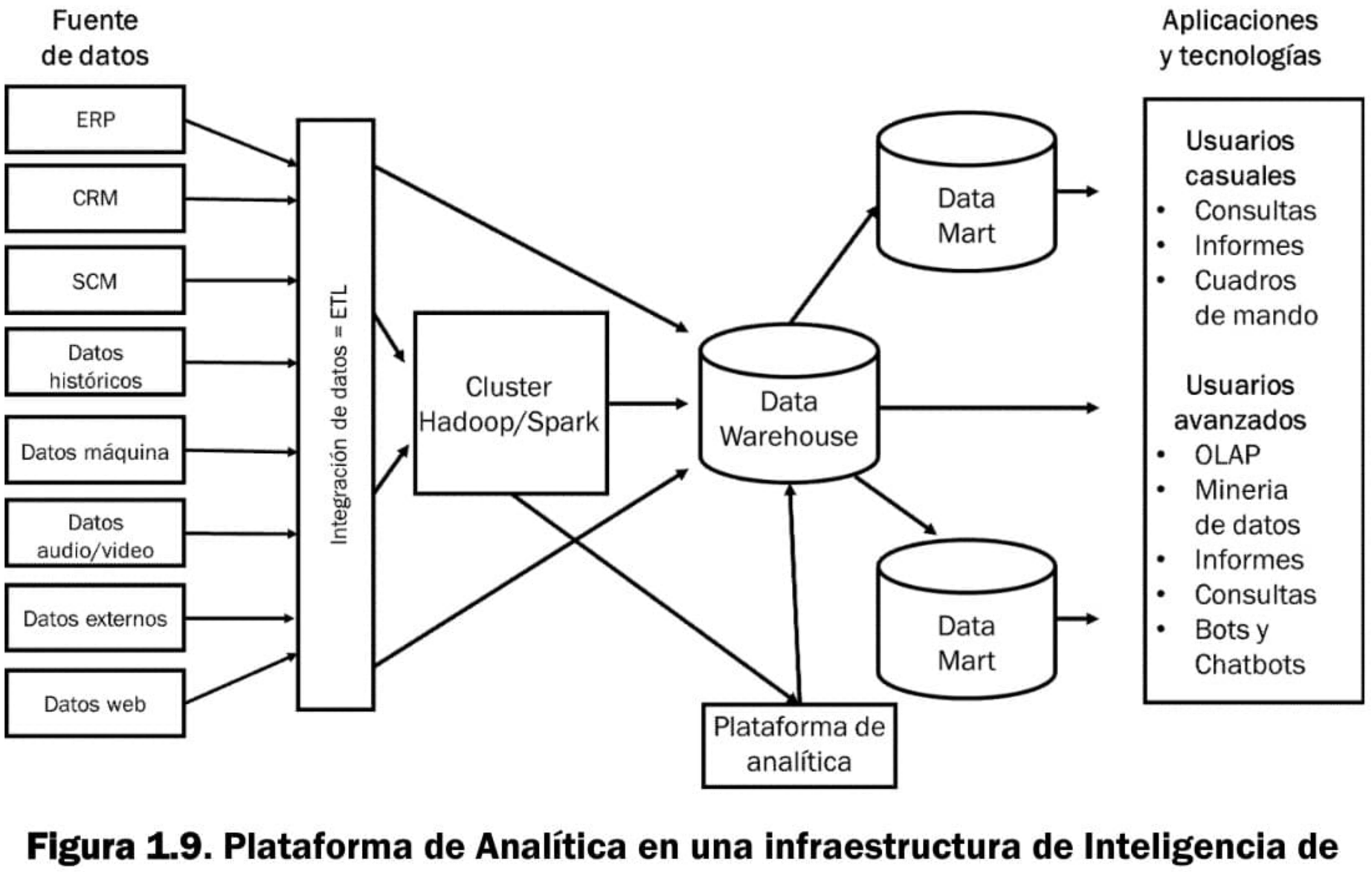


Figura 1.9. Plataforma de Analítica en una infraestructura de Inteligencia de Negocios
Fuente: Laudon (2014) (Adaptada del original)[16]

Los datos utilizados por las empresas pueden ser datos internos (datos históricos, datos transaccionales, datos generados por máquinas, datos de audio,

video) y datos externos (fuentes externas). Los datos se extraen, transforman y cargan en el *clúster* de Hadoop y en los *Data Warehouses* (y *Data Marts*). Los *clústeres* de Hadoop preprocesan *Big Data* para su uso en los almacenes de datos, *Data Marts* o en una plataforma analítica, para consultas directas de usuarios potenciales. Una vez que los datos han sido capturados y organizados utilizando las tecnologías correspondientes, se ponen disponibles para un análisis profundo utilizando *software* de consultas e informes, análisis de datos con técnicas OLAP y de Minería de Datos, principalmente. Los resultados (salidas del proceso) incluyen reportes, consultas, cuadros de mando (*dashboards*), sistemas de procesamiento analítico en línea OLAP, Minería de Datos y herramientas de Inteligencia Artificial para aplicaciones avanzadas, tal como veremos en los capítulos 7, 8, 9 y 13.

INFRAESTRUCTURA DE INTELIGENCIA DE NEGOCIOS

Las empresas necesitan información concisa y fiable sobre operaciones actuales, tendencias y cambios en toda la compañía. Si trabaja en una empresa grande (como una multinacional), los datos que se necesitan pueden estar agregados o desagregados en departamentos, tales como ventas, fabricación, contabilidad, recursos humanos, internacional e, incluso, fuentes externas con datos estadísticos, datos demográficos, datos de la competencia o datos abiertos procedentes de portales gubernamentales.

Una infraestructura moderna de Inteligencia de Negocios consta de un conjunto de herramientas para la obtención de información útil de todos los diferentes tipos de datos utilizados por los negocios (estructurados, no estructurados o semiestructurados). Un sistema de Inteligencia de Negocios constará de componentes organizados en dos grandes bloques: Inteligencia de Negocios propiamente dicha y Analítica de Negocios (*Business Analytics*) organizadas como **plataformas de Analítica**.

1.9 INTELIGENCIA DE NEGOCIOS EN *BIG DATA*

Las técnicas y herramientas de *Big Data* se integran en las soluciones de *Business Intelligence*; incluso hay expertos en análisis de datos y consultoras que consideran que *Big Data* es la evolución de Inteligencia de Negocios. En realidad, *Big Data* adquiere todo su potencial cuando se realiza el análisis de dichos grandes volúmenes de datos. Las técnicas y herramientas tradicionales utilizadas en Inteligencia de Negocios se podrán utilizar para la gestión y procesamiento de *Big Data*, aunque, como ya se ha comentado, se requerirán herramientas específicas para su tratamiento.

Big Data e Inteligencia de Negocios (BI) son dos tecnologías utilizadas para análisis de datos y ayuda en la toma de decisiones empresariales. Son dos

tecnologías muy similares y, de hecho, existen especialistas de Inteligencia de Negocios que plantean dos vistas diferentes de ambas soluciones: *Big Data* es una evolución de Inteligencia de Negocios, o bien la *Inteligencia de Negocios moderna* ha de integrar las soluciones de *Big Data* en las empresas. En nuestro caso, consideramos que *Big Data* se ha de integrar dentro de la arquitectura de Inteligencia de Negocios, aunque analizaremos las diferencias entre ambas soluciones de empresa.

Inteligencia de Negocios y *Big Data* tienen en común su objetivo principal: el análisis de datos para el soporte en la toma de decisiones empresariales. Difieren, esencialmente, en su infraestructura, en el tipo de datos que analizan y en el modo que realizan el análisis.

Las soluciones tradicionales de Inteligencia de Negocios se basan en el principio de agrupamiento de todos los datos empresariales en un servidor central instalado en la empresa (o, como veremos más adelante, con la opción de servidores virtuales instalados en la nube). Normalmente, los datos se analizan en modo tradicional (*off-line*) desde los componentes de la infraestructura de la IN, después de que se han depositado en el entorno típico del almacén de datos de la empresa (*Data Warehouse*). Los datos se estructuran en una base de datos relacional convencional con un conjunto adicional de índices y formas de acceso en las tablas (con herramientas OLAP de cubos multidimensionales). Pese a las diferencias entre *Big Data* e Inteligencia de Negocios, la situación actual camina a la integración de ambos tipos de herramientas: tradicionales de Inteligencia de Negocios, que manejan datos estructurados, con específicas de *Big Data*, que manipulan datos no estructurados y semiestructurados.

DIFERENCIAS ENTRE UNA SOLUCIÓN DE INTELIGENCIA DE NEGOCIOS Y UNA SOLUCIÓN DE *BIG DATA*

Se pueden utilizar las soluciones de Inteligencia de Negocios para procesar *Big Data*, teniendo en cuenta las diferencias entre ambas tecnologías.

La metodología tradicional analizada en el libro es la arquitectura fundamental para el análisis de datos en la empresa y la tendencia de *Big Data* que ha emergido en estos últimos años ha de integrarse en ella. La diferencia principal reside en el hecho de que los datos de los sistemas de Inteligencia de Negocios tradicionales se suelen alojar en un servidor central de la empresa (o, como ya se ha comentado, en servidores virtuales en la nube, normalmente en la modalidad de nube privada, con dichos servidores al servicio exclusivo de la empresa) y los datos se analizan, normalmente, en modo *off-line*. El procesamiento de datos transaccionales (operacionales) en línea conocido como OLTP (*Online Transaction Proccesing*) transfiere los datos al entorno ya conocido de almacén de datos (*Data Warehouse*). Los datos, normalmente estructurados, residen en un entorno de bases de datos relacionales (RDBMS, en inglés, o SGBDR) con muy pocos datos desestructurados o no estructurados.

Una solución de *Big Data* es diferente en muchos aspectos de un sistema de Inteligencia de Negocios tradicional, especialmente en los siguientes (Sawant, Shah 2013: 2)[17]:

- En un entorno de *Big Data*, los datos se almacenan en un sistema de archivos (ficheros) distribuidos en lugar de en un servidor central.
- En Inteligencia de Negocios, las funciones de procesamiento dirigen los datos en lugar de que los datos dirijan las funciones. Dicho de otra manera, las soluciones de Big Data *se centran en el análisis en torno a los datos, y eso facilita manipular grandes volúmenes de datos de un modo mucho más ágil.*
- *Big Data* puede analizar datos en diferentes formatos, tales como datos estructurados, no estructurados y semiestructurados. En la actualidad, el volumen de datos no estructurados es muy alto y crece a una gran velocidad (datos fiables estiman en un 80 al 90 por ciento, el total de datos no estructurados almacenados en las organizaciones y empresas). Las soluciones de *Big Data* permiten un análisis global de estos grandes volúmenes de datos, así como de las diferentes fuentes de información.
- Los datos procesados por *Big Data* pueden ser datos en tiempo real o datos tradicionales fuera de línea (*off-line*). Con estas técnicas, las empresas pueden tomar decisiones que afectan a sus negocios de un modo más ágil y eficiente.
- Las tecnologías de *Big Data* recurren a sistemas de procesamiento paralelo masivo, MPP *(Massive Paralell Processing).* Estas tecnologías mejoran, sensiblemente, la velocidad de análisis de datos. Las instrucciones de las soluciones se ejecutan de modo simultáneo en gran número de componentes que funcionan en paralelo y que, posteriormente, se reunifican y presentan los resultados globales. Las tecnologías MPP soporte de las bases de datos NoSQL y "en memoria" permiten aumentar las capacidades de procesamiento, lo que reduce los costes de almacenamiento y aumenta el rendimiento de las soluciones de los almacenes de datos.

Numerosos proveedores de Inteligencia de Negocio y proveedores de almacenes de datos ofrecen soluciones de integración de las tecnologías de *Big Data* (Hadoop, Spark) con las suyas específicas, con el objeto de realizar un análisis continuo de datos estructurados y no estructurados, almacenados en entornos diferentes, todos ellos bajo una interfaz de usuarios única para producir resultados de análisis de negocio integrados.

1.10 INTELIGENCIA DE NEGOCIOS MÓVIL

Las organizaciones y empresas están conectadas globalmente mediante todo tipo de dispositivos móviles (celulares). Los empleados, en un gran porcentaje, están conectados mediante herramientas móviles que les permiten tomar decisiones de negocio en modo ubicuo, en cualquier lugar, en cualquier momento y con cualquier dispositivo. La proliferación de teléfonos inteligentes (con pantallas de 5 y 6 pulgadas), tabletas (con tamaños de pantalla de 7 a 12 pulgadas) y computadores portátiles (*laptops*) ya en muchos casos con posibilidad de uso doble (computador y tableta) aumenta la necesidad de que las soluciones de Inteligencia de Negocios estén adaptadas a los dispositivos móviles.

Las aplicaciones empresariales móviles son, hoy en día, un componente fundamental en las empresas. Los gerentes y administradores, ejecutivos y usuarios especializados requieren un acceso rápido y seguro a los datos corporativos, con el objeto de poder tomar decisiones en cualquier momento que se necesite y en tiempo real. Existe un amplio catálogo de *apps* que se pueden utilizar en los diferentes componentes (o capas) de la arquitectura de Inteligencia de Negocios.

Los proveedores de soluciones de Inteligencia de Negocio ofrecen soluciones integradas o específicas, tanto integradas (*suites*) como independientes. Qlik, Microsoft, Oracle, IBM, SAS, SAP ofrecen soluciones de Inteligencia de Negocios específicas para dispositivos móviles.

La Inteligencia de Negocios Móvil (MBI, *Mobile Business Intelligence*) es hoy una tendencia imparable y la mayoría de los proveedores de soluciones, tanto propietarias como de código abierto, ofrecen sus herramientas y aplicaciones de un modo ***omnicanal.*** Algunas soluciones de Inteligencia de Negocios Móvil son:

- Microstrategy Mobile.
- Power View.
- Jaspersoft.
- Pentaho.
- IBM Cognos Mobile.
- SAP Business Object Mobile.

1.11 INTELIGENCIA DE NEGOCIOS EN LA NUBE

La Inteligencia de Negocios tradicional depende normalmente del departamento de TI de la empresa y, en consecuencia, de sus infraestructuras *hardware* y *software*, de modo que cuando se necesita ampliar o reducir los servicios, por

cualquier causa, pueden producirse grandes trastornos económicos y sociales en la propia empresa. La Analítica de Negocios y, en particular, la Analítica de Datos son las herramientas clave para ayudar en la toma de decisiones. El análisis de datos en la nube está creciendo a una gran velocidad ya que la nube ofrece, como ahora detallaremos, grandes ventajas.

El uso de servidores en la propia organización, así como las aplicaciones de *software* instaladas en ellos, implica en la actualidad, para su mantenimiento, unas inversiones iniciales y unos recursos que no sólo serán económicos, sino que en muchos casos requerirán de personal especializado. De igual forma, los gastos de disponer de un almacén de datos (*Data Warehouse*) o un servidor de almacenamiento con los datos propios se pueden ahorrar si se tienen en la nube y con la máxima seguridad si es un proveedor fiable y de calidad.

Inteligencia de Negocios en la nube es un conjunto de aplicaciones que se encuentran alojadas en ella (redes virtuales disponibles en los centros de datos de los proveedores de la nube). De este modo, aplicaciones como las fuentes de datos en los sistemas de información, las herramientas de Analítica de Datos, tableros de control o cuadros de mando (*dashboards*) para medir los indicadores de rendimiento estarán alojadas en la nube y no en las instalaciones de la empresa.

La nube ofrece un método ligero y ágil para acceder a las aplicaciones de BI, ya que una de sus grandes ventajas reside en que las aplicaciones de Inteligencia de Negocios en la nube tienen carácter ubicuo, y se puede acceder a ellas desde múltiples dispositivos (especialmente móviles), navegadores web, en cualquier lugar y en cualquier momento que se pueda necesitar.

VENTAJAS DEL USO DE INTELIGENCIA DE NEGOCIOS EN LA NUBE

Las ventajas son innumerables y destacamos algunas de las más sobresalientes:

- La disponibilidad de datos y del *software* en la nube permite grandes reducciones de costes en inversiones iniciales y de gestión o administración.
- Los gastos son más equilibrados, ya que con las plataformas de Inteligencia de Negocios en la nube sólo se paga lo que se necesita y se consume, por lo que pueden personalizarse las aplicaciones y comenzar a utilizar las versiones básicas e ir contratando versiones avanzadas a medida que se necesite.
- El disponer del *software* en la nube, en lugar de tenerlo instalado en la organización, facilita el acceso en tiempo real a los datos críticos del negocio.
- Permitir mejorar notablemente la velocidad en el despliegue de los componentes de Inteligencia de Negocios, ya que no se requieren

instalaciones de *hardware* y *software* adicionales; normalmente, serán suficientes las infraestructuras TIC existentes de la empresa.

- Altas propiedades de escalabilidad y elasticidad. Las herramientas en la nube permiten un escalamiento rápido para su adaptación al incremento en la cantidad de usuarios de la empresa o en sus capacidades de procesamiento o almacenamiento de datos.
- Posibilidad de almacenar grandes volúmenes de datos y a precios ajustados.
- Mayor velocidad en el momento de realizar informes, consultas, etcétera, dado que los accesos a las aplicaciones de la nube se pueden realizar en cualquier momento y lugar, y con cualquier dispositivo, siempre que se tenga acceso a Internet y a la plataforma de Inteligencia de Negocios.
- Las nuevas herramientas de aplicaciones de BI facilitan la rápida creación de informes de un modo más sencillo y ágil, junto con las consultas en tiempo real y facilidad para realizar los análisis de datos que se puedan requerir y en tiempo real.

USO DEL SERVICIO EN LA NUBE

Una de las ventajas más significativas de Inteligencia de Negocios en la nube es su facilidad de uso de sus aplicaciones y servicios. Para hacer uso del servicio sólo hay que suscribirse, pagar una cuota mensual —como cualquier otro servicio de la empresa, como la luz, el agua o la calefacción—, que dependerá del número de usuarios y el volumen de información que requiera el cliente y las aplicaciones ofertadas por la plataforma de Inteligencia de Negocios contratada.

La mayoría de los proveedores de Inteligencia de Negocios, tanto tradicionales con *software* propietario (IBM, Oracle, SAS, SAP, Microstrategy, Microsoft) o *software* abierto (Pentaho, Jaspersoft) o nuevos proveedores de soluciones como Birst o Good Data, ofrecen porfolios con soluciones de Inteligencia de Negocios en la nube.

INTELIGENCIA DE NEGOCIOS EN LA NUBE CON LA INTEGRACIÓN DE *BIG DATA*

Los sistemas de bases de datos y almacenamiento de datos tradicionales no pueden gestionar bien grandes volúmenes de datos (*Big Data*) además del coste y consumo de tiempo que exigen y las otras grandes características como variedad y velocidad de los *Big Data*; por otro lado, no son flexibles y tardan mucho tiempo en responder a las consultas y peticiones por parte de los usuarios cuando requieren decisiones en cuestión de horas o a lo sumo días.

Además, los datos no estructurados y semiestructurados son difíciles de manipular por las herramientas tradicionales, por lo que se requiere la

integración de las técnicas de *Big Data* en las soluciones de *software* clásicas, tales como Office 365, Google Apps, Salesforce, Google Analytics, etc. Las soluciones de *software* en la nube están facilitando la resolución e integración de herramientas de *Big Data* tales como Hadoop, Spark, MapReduce, Hive, etcétera. Google BigQuery es una herramienta que facilita la integración de herramientas de BI con *Big Data*. La misma situación se está dando con la integración de herramientas y aplicaciones de Inteligencia de Negocios con las arquitecturas de bases de datos **NoSQL** y "en memoria", que se conectan con todo tipo de información en la propia empresa (*on-premise*) o con la información existente en otras nubes de socios, clientes, patrocinadores, externas.

SEGURIDAD DE LOS DATOS: RIESGOS

La migración a la nube en las empresas ha alcanzado grandes cotas de penetración en todo tipo de aplicaciones y servicios. La alternativa de Inteligencia de Negocios en la nube se está convirtiendo en prioritaria para todo tipo de empresas.

Sin embargo, la nube también ofrece riesgos a la hora de su aceptación por la empresa. Existen todavía empresas que consideran altos los riesgos de llevar sus datos y aplicaciones a la nube. Efectivamente, esos riesgos existen aunque los servicios de la nube suelen tener unos índices de seguridad que, en la mayoría de los casos, son mucho más altos que en las instalaciones corporativas.

La mejor alternativa que ha de seguir una organización o empresa para migrar sus servicios a la nube es la elección de un proveedor de *cloud computing* fiable y de calidad, que ofrezca todos los servicios de Inteligencia de Negocios que requiera la empresa y que cumpla rigurosamente todas las normativas nacionales e internacionales de protección de datos y de privacidad.

LA INTELIGENCIA DE NEGOCIOS EN LA NUBE

Las predicciones de tendencias de Inteligencia de Negocios para 2017 y 2018 publicadas por las grandes consultoras multinacionales, empresas proveedoras de soluciones de Inteligencia de Negocios, etcétera consideran de modo muy sobresaliente las grandes ventajas de las soluciones en la nube.

Tableau, uno de los grandes proveedores de *software* para Inteligencia de Negocios y líder en herramientas de visualización de datos, publica todos los años una serie de informes sobre la previsión de tendencias tecnológicas para el año siguiente. En el caso de Inteligencia de Negocios, el informe publicado en diciembre de 2016, *Las 10 tendencias principales de Inteligencia de Negocios para 2017*, destaca como tendencia emergente y clave que "la transición a la nube se acelera en 2017". Tableau considera que "los modelos híbridos de la nube están llegando a todas las organizaciones y la Analítica de Datos se está

trasladando allí donde están los datos. Las ventajas de la rapidez, escalabilidad y pago por uso, así como los ahorros en infraestructuras locales, señalan la tendencia de implantación de la Inteligencia de Negocios en la nube".

La transición a la nube por parte de las empresas es una tendencia creciente en 2018 y, en particular, en el caso de Inteligencia de Negocios en la nube. El éxito de la Inteligencia de Negocios en las organizaciones y empresas depende, en gran medida, de la incorporación de la empresa a la nube

1.12 PROVEEDORES DE INTELIGENCIA DE NEGOCIOS: CUADRANTE MÁGICO DE GARTNER DE BI & ANALYTICS

Los productos de Inteligencia de Negocios y Analítica de Negocios se proporcionan como soluciones de *software*, tanto en código propietario como en *software* libre (*open source*, fuentes abiertas). Los proveedores de *software* propietario más conocidos son Oracle, SAP, IBM, SAS, Microsoft, MicroStrategy, Tableau, que durante años llevan proporcionando soluciones a grandes empresas, y en los últimos años también a pequeñas y medianas empresas. Últimamente, se han implantado en el mercado, proveedores de soluciones de *software* abierto (licencias gratuitas), que ofrecen, a su vez, plataformas de desarrollo (además del *software* libre con licencias gratuitas) y que constituyen su baza fundamental en la competencia con los proveedores de *software* propietario (los proveedores más conocidos son Pentaho, Jaspersoft y Birt).

Los proveedores de soluciones de *software* de Inteligencia de Negocios son muy numerosos. Una de las mejores herramientas que tienen las empresas de cualquier índole para evaluar soluciones de *software* adecuado a sus líneas de negocio son los cuadrantes mágicos de Gartner y, en nuestro caso particular, el cuadrante mágico de Inteligencia de Negocios.

La consultora multinacional Gartner, desde finales de la primera década del siglo XXI, publica todos los años numerosos informes de impacto en soluciones de tecnologías de la información y, en particular, en el caso de Inteligencia de Negocios, es muy apreciado por las empresas el popular informe conocido como *Cuadrante Mágico de Gartner* y que, en el caso de negocios, se denomina *Cuadrante Mágico de Gartner de Inteligencia de Negocios y Plataformas de Analítica* (*Business Intelligence & Analytics*), donde estudia a un gran número de proveedores y los evalúa en función de una serie de indicadores específicos. El informe clasifica a los proveedores de soluciones de *software* propietario y *software* abierto y examina sus fortalezas y debilidades. El informe se titulaba *Business Intelligence* y, desde el año de 2012, cambió su nombre por el ya citado de *Business Intelligence & Analytics Platforms*, coincidiendo con el despegue empresarial de las necesidades de Analítica de Negocios. A continuación,

describirnos más en detalle, los cuadrantes mágicos de Gartner de Inteligencia de Negocios y Plataformas de Analítica de Gartner de los años 2017 y 2018.

1.12.1 CUADRANTE MÁGICO DE GARTNER 2017

En el mes de febrero de 2017 —tradicionalmente, es el mes elegido para su presentación del estudio que realiza a lo largo del año— presentó su edición de 2017. Gartner evalúa un gran número de proveedores y selecciona un buen número de ellos, a los que aplica un conjunto de indicadores relacionados con la Inteligencia de Negocios y plataformas de Analítica. La evaluación final se publica y se señalan los indicadores (capacidades exigidas a cada fabricante) del estudio que se han analizado en las empresas finalistas. Algunas de las capacidades críticas de la edición de 2017 han sido:

- Cloud BI.
- Herramientas de almacenamiento de datos y ETL.
- Analítica avanzada.
- Cuadros de mando (*dashboards*) para Analítica.
- Exploración visual interactiva de datos.
- Descubrimiento inteligente de datos.
- Mobile BI.

El cuadrante mágico se divide, a su vez, en cuatro cuadrantes donde se sitúan y clasifican las empresas en función de la evaluación alcanzada. Los cuatro cuadrantes son: empresas líderes, empresas visionarias, empresas retadoras (*challengers*) y empresas jugadoras de nicho de mercado. Las empresas se sitúan en los respectivos cuadrantes en función de los índices alcanzados en dos enfoques importantes: completitud de visión y capacidad/habilidad de ejecución del *software*. El cuadrante más deseado por las empresas es el de líderes y es el que Gartner recomienda seguir estrechamente justificando a todas las empresas, sea cual sea su situación en el cuadrante. El informe (77 páginas en 2017) destaca los indicadores, los resultados finales y un examen personalizado de todas las empresas que han llegado a la fase final, en los que proporciona un informe personalizado de fortalezas y de precauciones/advertencias para tener presente por los clientes y, lógicamente, por las propias empresas.

En 2017, la empresa líder indiscutible fue Microsoft, seguida muy de cerca por Tableau y, un poco más distanciada, por Qlik. Empresas visionarias (que se han de seguir también con rigor) fueron SAP, SAS, IBM, Microstrategy, Tibco y Salesforce, entre otras; empresas de nicho (especializadas en un sector determinado) han sido Alteryx, Birst, Oracle, Pentaho (líder en *software* de código abierto) y Datameer; empresas retadoras o aspirantes a líderes, en la edición de 2017, no han sido destacadas.



Figura 1.10. Cuadrante mágico de Gartner Business Intelligence & Analytics Platforms 2017

Las conclusiones más destacadas de interés general para soluciones de Inteligencia de Negocios y plataformas de Analítica son las siguientes: El estudio apunta a la progresiva evolución de las herramientas de Inteligencia de Negocios, que cada vez sean más fáciles de utilizar (requieran una menor dependencia de los departamentos de TI) y que incorporen mayor número de funcionalidades. Estas características permiten analizar los procesos de análisis y toma de decisiones, además de proporcionar una mayor autonomía y economía al usuario (esta tendencia comienza a denominarse **Inteligencia de Negocios de autoservicio o democrática**). Gartner considera que las organizaciones deberán acometer nuevos proyectos de Inteligencia de Negocios y de Analítica, a través de plataformas modernas, con el objeto de aprovechar la innovación reinante en el mercado.

1.12.2 CUADRANTE MÁGICO BI & ANALYTICS 2018

El cuadrante mágico BI & Analytics 2018 publicado también en el mes de febrero arroja resultados muy similares en cuanto a proveedores, aunque varían sus posiciones de los cuadrantes.

El cuadrante mágico de Gartner BI & Analytics Platforms 2018 publicado en febrero de 2018 arroja resultados muy similares en cuanto a proveedores aunque varían sus posiciones en los cuadrantes. El cuadro de LÍDERES continúa encabezado por Microsoft seguido de Qlik y Tableau. Destaca MicroStrategy que sale del cuadro de VISIONARIO y pasa a CHALLENGER y también sale de todo el cuadrante, el proveedor de código abierto (open source) Pentaho y solo queda en esta categoría, el proveedor Birst que sigue en el cuadrante de "Jugador de nicho".



Figura 1.11. Cuadrante mágico de Gartner Business Intelligence & Analytics Platforms 2018

1.13 INTELIGENCIA DE NEGOCIOS FUTURA: INTEGRACIÓN DE *BIG DATA*, INTERNET DE LAS COSAS E INTELIGENCIA ARTIFICIAL

La Inteligencia de Negocios tradicional, como ya hemos visto, ha estado soportada por los sistemas de almacenamiento (*Data Warehouse* y *Data Marts*, además de bases de datos relacionales y archivos) junto con herramientas de informes (reportes) y consultas, cuadros de mando o tableros de control (*dashboards*) y herramientas de Analítica basadas en Minería de Datos y técnicas estadísticas. Herramientas de *software* propietario muy reconocidas eran IBM Cognos, Oracle Business Intelligence Suite, Tableau, Pentaho, SAP o SAS.

A medida que han ido creciendo los datos a disposición de las organizaciones y las tendencias de *Big Data* basadas en el gran volumen de datos, variedad y velocidad (esencialmente), se han ido consolidando, han ido apareciendo nuevas herramientas para almacenamiento, procesamiento, análisis y presentación de resultados mediante técnicas de visualización avanzada de datos. El almacenamiento y procesamiento de los grandes volúmenes de datos ha evolucionado en los dos métodos ya citados: procesamiento por lotes (*batch processing*) y procesamiento continuo o en *streaming*, sobre todo para aplicaciones de tiempo real. Han aparecido nuevos sistemas de almacenamiento de datos, que han sido la evolución natural de los *Data Warehouses* y *Data Marts*, para un procesamiento y análisis más rápidos. Las nuevas bases de datos NoSQL y "en memoria" principalmente, han permitido almacenar y procesar datos de todo tipo (estructurados, no estructurados y semiestructurados), al contrario que los almacenes de datos tradicionales que sólo soportaban datos con esquema fijo (tablas, fila y columnas). Las nuevas tecnologías se han ido integrando en plataformas o ecosistemas de datos en torno a *Big Data*. La plataforma dominante en el mercado ha sido Hadoop gestionada por la Fundación Apache Software, aunque en estos dos últimos años ha aparecido la plataforma Spark también soportada por la Fundación Apache Software, pero diseñada y construida para tratamiento de datos en tiempo real y para reducir algunas de las debilidades que tiene Hadoop por sus pequeñas latencias que en ocasiones pueden producir retardos no deseados. Tanto Hadoop como Spark son marcos de trabajo de *software* de código abierto, desarrollados y mantenidos por profesionales de reconocido prestigio en torno a la mencionada fundación Apache. En el futuro, todas estas plataformas soportarán Inteligencia Artificial y los algoritmos de aprendizaje automático y aprendizaje profundo, que comentaremos con más detalle en el capítulo 13, y comienzan a desarrollarse plataformas como Apache Flink y Apache Beam, soportes de la Inteligencia de Negocios basada en Inteligencia Artificial y en Ciencia de Datos.

Al principio del desarrollo de las infraestructuras de *Big Data*, las aplicaciones funcionaban con independencia de las aplicaciones de Inteligencia de Negocios

tradicional. Desde hace poco tiempo, (y sobre todo lo será en el futuro) la palabra mágica en Inteligencia de Negocios ha sido la integración de las tendencias tradicionales y las tendencias de *Big Data*, con herramientas que facilitan la comunicación entre ambos tipos de sistemas, y tanto los proveedores de soluciones de Inteligencia de Negocios, de *software* propietario como los de código abierto, contemplan la integración como la solución más eficiente y recomendada para organizaciones y empresas. Por estas razones, han ido apareciendo tendencias como ***Data Lake*** (**Lago de Datos**), que veremos en el capítulo 4, y que, según las consultoras o empresas que recomiendan este nuevo sistema de almacenamiento, lo considerarán como una evolución del *Data Warehouse* tradicional (DW 4.0) o bien la integración en *Big Data*, para cubrir determinadas carencias que han ido surgiendo en las tecnologías y técnicas de proceso de los grandes volúmenes de datos.

1.14 LA EVOLUCIÓN HACIA LA INTELIGENCIA DE NEGOCIOS EN LA NUBE *(CLOUD BI)*

Ya hemos visto en un apartado anterior la tendencia creciente a utilizar soluciones de Inteligencia de Negocios en la nube (en sus tres modelos de servicio: *software* como servicio, plataformas como servicios e infraestructuras como servicio) y la mayoría de los proveedores de *software* ofrecen ya una amplia gama de sus aplicaciones basadas en la nube (pública, privada o híbrida). El prestigioso columnista de *software* de negocios, Louis Columbus[18], en la revista *Forbes* (donde publica sus artículos y posee un *blog* tecnológico muy reconocido) publicó en abril de 2017 un artículo muy ilustrativo sobre las tendencias de Inteligencia de Negocios y, en especial, el modelo de *Cloud BI* que consideraba la solución de BI de mayor impacto y crecimiento en ese año y en los siguientes. Para ello, se apoyó en el sexto estudio anual de la consultora especializada Dresner Advisory Service *Cloud Computing and Business Intelligence Market Study* (informe de 101 páginas, en pdf, y que está accesible a clientes previo registro, pero cuyos resultados más sobresalientes publicó Columbus en su artículo). El estudio realizado por la consultora sobre la base de informes relevantes y entrevistas a un gran número de profesionales y directivos especializados en TI, en BI y en la nube, mostraba resultados agrupados en categorías:

- Importancia del modelo *Cloud BI* por funcionalidades.
- Presencia de la Inteligencia de Negocios en la nube, en la industria vertical seleccionada (educación, finanzas, negocios, fabricación, salud, telecomunicaciones).
- Informes similares de organismos públicos.
- Modelos de despliegue geográfico a lo largo del mundo.

Los requerimientos más demandados por los usuarios de Inteligencia de Negocio, de la mayoría de los niveles corporativos de las empresas encuestadas e investigadas, fueron:

- Visualización avanzada.
- Consulta *ad-hoc*.
- Integración de datos.
- Herramientas de calidad de datos.
- Herramientas de ETL.
- Herramientas de autoservicio (*self-service*) en la Inteligencia de Negocios.

Columbus analiza las tecnologías estratégicas más requeridas por las organizaciones y empresas, de un total de 31 categorías seleccionadas. Las cinco tecnologías más demandadas son:

- *Reporting* (creación de informes y reportes).
- *Dashboards* (cuadros de mando) personalizados.
- Funcionalidad de *self-service* o autoservicio (facilidad de uso por cualquier tipo de usuario).
- Visualización avanzada.
- Almacenamiento de datos (*Data Warehousing*).

De las restantes tecnologías propuestas en el estudio, hemos seleccionado las que consideramos más innovadoras, y que iremos viendo a lo largo de nuestra obra:

- Minería de Datos y algoritmos avanzados.
- *Big Data* (Hadoop, preferentemente).
- Descubrimiento y visualización de datos.
- Soporte de dispositivos móviles: Inteligencia de Negocios Móvil.
- *Data Storytelling* (narración o historias de datos).
- Inteligencia de Negocios con técnicas y soluciones empresariales embebidas.
- Analítica “en memoria”.
- Inteligencia de Negocios cognitiva (basada en Inteligencia Artificial).
- *Software* de código abierto (*open source*).
- *Social Media Analytics* (analítica social).
- Internet de las cosas.

- Analítica de textos.
- Visualización de datos avanzada.
- Integración de datos / Herramientas de calidad de *software* / Herramientas ETL.

1.15 RESUMEN

El término inteligencia de negocios -también admitido en singular, inteligencia de negocio- (business intelligence) es una combinación de arquitecturas, herramientas, bases de datos, herramientas analíticas, aplicaciones y metodologías (Turban 2008). Analítica de negocios (business intelligence) es un término cuya implantación ha crecido a medida que inteligencia de negocios se ha ido asentando en organizaciones y empresas como una estrategia vital en las corporaciones; en los últimos años, el término convive con Analytics, acuñado por proveedores y consultores de soluciones de TI y que se refiere a las herramientas y técnicas para el análisis y comprensión de datos; ambos términos conviven y se suelen utilizar como sinónimos. Con la aparición del término Big Data, hoy día BI, BA y Big Data junto con la integración de todas ellas en Ciencia de Datos (Data Science) constituyen los pilares de la organización empresarial.

La arquitectura de un sistema de inteligencia de negocios tradicional consta de cuatro capas o etapas: fuentes de datos, proceso ETL (extraer, transformar y cargar), almacenes de datos y visualización de resultados por el usuario final.

- Big Data es el conjunto de grandes volúmenes de datos de gran volumen, variedad y velocidad de procesamiento que mayor impacto ha tenido en los últimos años en la inteligencia de negocios.
- La arquitectura de un sistema de inteligencia de negocios con integración de Big Data es una adaptación de su modelo tradicional: fuentes de datos, recolección de datos, almacenamiento, procesamiento y análisis, visualización y una capa transversal de administración o gestión.
- Las aplicaciones y tecnologías utilizadas en BI y BA son: informes (reportes), consultas, almacenamiento de datos, minería de datos, minería web y de texto, estadística, visualización de datos.
- La inteligencia de negocios tradicional sobre herramientas de escritorio, hoy día se extiende a BI en la nube (cloud computing) y BI móvil, integradas con Big Data.
- Los proveedores comerciales y de código abierto (gratuito) de soluciones de software son numerosos. Las recomendaciones para la elección del proveedor o proveedores más idóneos para una empresa son las dadas por las grandes consultoras de TI, especialmente Gartner, Forrester, McKinsey... a los que se unen las consultoras y auditoras como Accenture, PwC,

Deloitte, KPM, Indra, etc. Se describen las recomendaciones de proveedores comerciales dadas por Gartner en sus Cuadrantes Mágicos de Gartner de Business Intelligence & Analytics Platforms de 2017 y 2018.

- La inteligencia de negocios se clasifica en la actualidad en dos grandes categorías: tradicional o clásica que se centra en los datos históricos y moderna que integra las nuevas tendencias tecnologías de Big Data, Internet de las Cosas e Inteligencia Artificial que incorpora Analytics con soluciones integradas o soluciones particulares.

NOTAS

[1] *Efrain Turban, Ramesh Sharda y Dursun Denle. Decision Support and Business Intelligence Systems 9th edition, Pearson.*

[2] *Gartner. "Business Intelligence". IT Glossary.* www.gartner.com/it-glossary/business-intelligence-bi

[3] *Gartner. "Business Analytics". IT Glossary.* www.gartner.com/it-glossary/business-analytics

[4] *Searchdatamanagement.* techtarget.com/definition/business-intelligence

[5] *Josep Curto Introducción al Business Intelligence. Editorial UOC, p 18, (2016).*

[6] Ibidem 5, p 21.

[7] *Lih Ong1, Pei Hwa Siew1 and Siew Fan Wong. A Five-Layered Business Intelligence Architecture. Communications of the IBIMA. IBIMA Publishing. Vol. 2011 (2011), Article ID 695619, 11 pages. DOI: 10.5171/2011.695619.*
http://www.ibimapublishing.com/journals/CIBIMA/cibima.html

[8] Ibidem 7

[9] Ibidem 7

[10] Ibidem 7

[11] *Doug Laney. "3D Management: Controlling Data, Volume, Velocity and Variety", Gartner, 2001.*

[12] *C. M. Duan. "Design of Big Data Processing System Architecture Based on Hadoop under the Cloud Computing", Applied Mechanics and Materials, Vols. 556-562, pp. 6302-6306, 2014.*

[13] *TURBAN E, SHARDA, R. y DELEN, D. Decision Support and Business Intelligence Systems. Ninth edition. New Jersey: Pearson, 2011.*

[14] *TURBAN E, SHARDA, R. y DELEN, D. Business Intelligence & Analytics. Decision Support Systemas. Tenth edition. New Jersey: Pearson, 2014.*

[15] *Kennet Laudon y Jane Laudon. Management Information Systems. Managing the Digital Firm. 13a ed. Pearson. 2014.*

[16] *Ibídem 14. Kennet Laudon y Jane Laudon*

[17] *Sawant, Nitin, Shah, Himanshu. Big Data Application Architecture Q&A. A Problem – Solution Approach. Apress, 2013.*

[18] *Louis Columbus. "2017 State Of Cloud Business Intelligence", Forbes, 9 de abril.* https://www.forbes.com/sites/louiscolumbus/2017/04/09/2017-state-of-cloud-business-intelligence/#5d2e23b145c2

CAPÍTULO 2

ANALÍTICA DE NEGOCIOS *(BUSINESS ANALYTICS):* UNA VISIÓN GLOBAL

CONTENIDO

2.1. Introducción
2.2. Conceptos básicos de analítica de negocios (business analytics)
2.3. Business Analytics versus Data Analytics
2.4. Analítica avanzada (AA)
2.5. Caso de estudio: Cuadrante mágico de Gartner de BI & Analytics
2.6. Organización, tipos y fuentes de datos
2.7. Ciclo de vida de los datos
2.8. Analítica de datos: conceptos y tipos
2.9. Big Data Analytics
2.10 Ciencia de datos: Evolución de la analítica de negocios y el análisis de datos
2.11 Tendencias de Analytics
2.12 RESUMEN

OBJETIVOS

- Definir los conceptos fundamentales de la analítica de negocios
- Conocer las semejanzas y diferencias entre la inteligencia de negocios y analítica de negocios.
- Aprender los conceptos fundamentales de la analítica de datos soporte de la analítica de negocios.
- Conocer los diferentes tipos de analítica de datos y el ciclo de datos empresariales.
- Conocer las técnicas más utilizadas en la analítica de datos.
- Conocer las herramientas básicas y proveedores de analítica de negocios
- Aprender los conceptos básicos de la analítica de big data (big data analytics)
- Aprender los conceptos fundamentales de ciencia de datos como integración de la analítica e ingeniería de datos.
- Introducción a la analítica de big data.
- Conocer las competencias básicas del analista de datos y del ingeniero de datos
- Introducción a la disciplina de ciencia de datos (data science) y el rol de científico de datos

2.1 INTRODUCCIÓN

Analítica de Negocios (*Business Analytics)* es un componente fundamental en la arquitectura de los sistemas de Inteligencia de Negocios tradicionales y vigentes en la mayoría de las organizaciones y empresas, como responsable del análisis de los datos que han sido depositados en la infraestructura de datos de una organización empresarial (bases de datos relacionales y heredadas, *Data Warehouse* y *Data Marts*).

Sin embargo, la aparición y llegada de los grandes volúmenes de datos a las empresas (*Big Data*) han hecho evolucionar a los sistemas de Inteligencia de Negocios, dándoles mayor preponderancia a todo el análisis de datos por la dificultad de manejar no sólo los grandes volúmenes, sino también sus diferentes formatos (estructurados, no estructurados) así como la velocidad de generación de dichos datos corporativos y otras características que ampliaremos en el capítulo 5, específico de *Big Data*.

Esta situación ha hecho que, aunque Inteligencia de Negocios sigue siendo el término más utilizado todavía en la empresa, proveedores de soluciones *hardware* y *software*, y nosotros, consideramos que la Analítica de Negocios sigue siendo un componente fundamental de un sistema de BI. La evolución de los negocios ha hecho que ambos términos puedan vivir de modo independiente y así lo muestran las numerosas soluciones dadas por muchos proveedores de reconocido prestigio.

Así, Inteligencia de Negocios se puede considerar como la infraestructura de recolección y gestión de los datos corporativos (internos y externos) y sigue utilizando las técnicas y herramientas que vimos en el capítulo 1 (y que seguiremos ampliando en el resto del libro); y Analítica de Negocios (*Business Analytics*) es la etapa siguiente, conectada a las infraestructuras de datos, y las últimas etapas de presentación y visualización de resultados.

En el libro consideraremos la Inteligencia de Negocios como el sistema predominante en la empresa para tratar todo el tema del ciclo de datos, pero la Analítica de Negocios se considerará también como una disciplina integrada en todo el sistema de Inteligencia de Negocios, pero con entidad propia, dado que así comienza a ser entendida por numerosos proveedores de soluciones informáticas, de los que al final depende la empresa para su toma de decisiones gerenciales.

2.2 CONCEPTOS BÁSICOS DE ANALÍTICA DE NEGOCIOS (*BUSINESS ANALYTICS*)

Analítica de Negocios (*Business Analytics*) es un componente fundamental de la Inteligencia de Negocios y es el conjunto de técnicas y herramientas que ayudan a los usuarios a aprovechar los datos empresariales, con el objetivo de tomar decisiones de empresa mejores, acertadas y en consonancia con sus líneas de negocio. Una buena definición de los objetivos de la Analítica de Negocios es la dada por *Information Builder*[1] (uno de los grandes proveedores mundiales de soluciones de Inteligencia de Negocios): "Las soluciones de BA permiten a las organizaciones identificar al instante y en tiempo real los factores que afectan a su rendimiento (desempeño), crear estrategias más precisas orientadas al futuro, mejorar la eficiencia, aumentar la rentabilidad y lealtad del cliente".

Muchas empresas comienzan a utilizar indistintamente los términos *Business Analytics* y *Analytics*. Ambos términos existen en la vida diaria de la empresa, dado que los proveedores de soluciones utilizan uno u otro términos según sus estrategias corporativas y ofrecen herramientas globales (*suites*) de Inteligencia de Negocios, mientras que otros ofrecen sólo las herramientas específicas de analítica de datos, para conectar e integrar al resto de las herramientas y técnicas de los diferentes componentes, como las fuentes de datos, infraestructura de los datos, visualización de datos y presentación de resultados.

BI y BA (también denominada *analítica empresarial*) giran en el entorno empresarial con el objetivo de aprovechar los datos disponibles para mejorar su toma de decisiones. La característica que diferencia a cada una de estas tendencias se refiere al tiempo y al uso de los datos. Ambos sistemas utilizan datos, pero Inteligencia de Negocios utiliza los datos históricos para aprender de

la toma de decisiones pasadas, mientras que Analítica de Negocios mira hacia adelante e intenta predecir lo que sucederá en el futuro.

DIFERENCIAS PRÁCTICAS DE ESTRATEGIA EMPRESARIAL

La Inteligencia de Negocios mira hacia atrás en el tiempo, descubriendo que es lo que se tiene y lo que ha ocurrido, adquiriendo conocimiento (*insights*) para conseguir mejoras empresariales. Desde un punto de vista práctico, la IN se refiere a la captura y acceso a los datos mediante *software* y a la infraestructura necesaria para conducir los datos debidamente preparados y organizados al componente de Analítica de Negocios para su análisis de datos. Una vez que se analizan los datos, es preciso optimizarlos para mejorar el rendimiento empresarial. La Analítica de Negocios se refiere más a la optimización de tendencias y necesidades del futuro.

Inteligencia de Negocios es un conjunto de técnicas y herramientas para la adquisición y transformación de datos en bruto en información significativa y útil para el propósito de análisis de la empresa. BI trata, sobre todo, de la utilización de datos históricos para ver lo que ha sucedido y ayudará a tomar mejores decisiones informadas basadas en los resultados pasados. La Inteligencia de Negocios se apoya en los informes que se generan por desarrolladores profesionales y que se ponen a disposición de toda la organización.

La **Analítica de Negocios** es un conjunto de tecnologías y habilidades que, basada en el rendimiento de la empresa, adquiere conocimiento basado en lo que ha sucedido y trata de anticiparse en las futuras tendencias por seguir y las necesidades del futuro.

Inteligencia de Negocios *versus* Analítica de Negocios *y Big Data*

La Inteligencia de Negocios moderna ha adquirido una gran notoriedad, ya que se apoya en las herramientas tradicionales de Inteligencia de Negocios y en la tendencia de *Big Data*. Aunque fabricantes y proveedores de *software* están adoptando el término de BA, y sobre todo *Analytics*, como sistemas independientes, nuestra opinión es más global y consideramos la Inteligencia de Negocios como estrategia empresarial a nivel global de organización, incluyendo en esta estrategia a la Analítica de Negocios como un subconjunto de BI basado en estadística, minería de datos, modelos de optimización y, esencialmente, analítica predictiva. Analítica de Negocios implica el uso de análisis estadístico y modelado predictivo con el objeto de establecer tendencias.

No obstante, dedicaremos gran atención a la Analítica de Negocios (o *Analytics*) y a la Analítica de Datos, con sus cuatro tipos de análisis de datos pero centrados en los métodos cualitativos para la predicción y prescripción del futuro.

2.3 BUSINESS ANALYTICS *VERSUS* DATA ANALYTICS

En los espacios de trabajo corporativos, la Analítica y los datos están jugando un rol mucho mayor que antes del advenimiento de *Big Data* y de herramientas analíticas de fácil uso, como es el caso de la popular Google Analytics, con la cual los empleados pueden medir sus capacidades analíticas de un modo real y práctico dentro de sus organizaciones. Los términos *Business Analytics, Data Analytics*, y una síntesis de ambas *Analytics*, simplemente se enfocan en las corporaciones y en los ámbitos científicos y de investigación. ¿Cuáles son las diferencias?

BA es un conjunto de habilidades, tecnologías y prácticas que permiten identificar tendencias y comportamientos a partir de datos empíricos, producir escenarios adecuados al trabajo y tomar decisiones óptimas para alcanzar objetivos del negocio.

BA en los sistemas de Inteligencia de Negocios tradicional estaba centrada en los métodos tradicionales de analítica, como era el caso de OLAP — *On-line Analytical Processing* — y herramientas básicas de estadísticas —como hojas de cálculo y similares—. Implicaba el uso de modelos directamente a los datos del negocio, con el objeto de mejorar el rendimiento de la organización. Así se diferenciaba entre la infraestructura de Inteligencia de Negocios tradicional (*Data Warehousing* y *Data Marts*), responsabilidad de la plantilla técnica de TI, y el entorno analítico, que era la responsabilidad de los usuarios de negocio.

Hoy la situación ha cambiado radicalmente desde la aparición de las tendencias *Big Data* y sus herramientas de análisis de grandes datos, facilitando la consolidación de ambas disciplinas y, a su vez, la integración de las mismas.

BA es una colección de herramientas para manipulación, minado y análisis de datos de los almacenes de datos (repositorios de datos). Los usuarios finales del negocio pueden trabajar con los datos y la información en un *Data Warehouse*, utilizando una variedad de herramientas y técnicas:

- **Reportes y consultas.** La Analítica de Negocios incluye reportes estáticos y dinámicos, todo tipo de consultas, descubrimientos de datos e información, vistas multidimensionales, componentes *drill-down* (desglose) para detalles, etcétera.
- **Minería de Datos, minería web y minería de textos**, junto con otras herramientas más sofisticadas de matemáticas y estadística.
- **Cuadros de mando (*dashboards*) y otras herramientas de visualización.**

Minería de Datos es un proceso de búsqueda de relaciones desconocidas o información en grandes bases de datos o almacenes de datos, utilizando herramientas inteligentes como computación neuronal, técnicas predictivas,

métodos estadísticos avanzados, etcétera. Su importancia exige un amplio conocimiento de los términos y herramientas de Minería de Datos (véase capítulo 8).

Analítica de Negocios o simplemente Analítica (*Analytics*) es un término definido por los proveedores de soluciones de *software*, que se centra más en herramientas y técnicas para análisis y comprensión de datos, tales como OLAP, modelos estadísticos y Minería de Datos.

2.3.1 HERRAMIENTAS DE ANALÍTICA DE NEGOCIOS

Son un conjunto de herramientas de *software* que se utilizan para analizar datos y producir informes, responder a cuestiones y consultas planteadas por empleados corporativos para seguir el progreso del negocio, utilizando indicadores clave de rendimiento o desempeño (KPI). Entre estas herramientas se contemplan:

- Modelos de datos.
- Minería de Datos.
- OLAP.
- Herramientas de *reporting* y *quering* (reportes o informes y consultas).
- Analítica de *Big Data*.

A todas las anteriores comienzan a introducirse recientemente técnicas de Inteligencia Artificial, redes neuronales artificiales, aprendizaje automático y aprendizaje profundo.

En función del rol profesional del empleado, varían las herramientas utilizadas. Así, los usuarios casuales están interesados, normalmente, en informes de producción. Los ejecutivos senior (directivos) tienden a utilizar las herramientas de BA para monitorizar las actividades de la empresa, utilizando interfaces visuales como *dashboards* y *scorecard*. Los empleados de nivel medio y analistas utilizan, preferentemente, los datos y el *software* específico. Los empleados operacionales, junto con clientes y proveedores, suelen utilizar informes preempaquetados.

Comienzan a utilizarse con gran profusión herramientas de visualización de datos y una rama de la Analítica conocida como **analítica visual**.

En las pequeñas organizaciones, los roles del analista de negocios y analista de datos suelen superponerse y el analista de negocios realiza tareas que implica manejo de datos y análisis de sistemas. Por el contrario, en organizaciones grandes, normalmente, los dos analistas realizan tareas que si bien están centradas en los datos se diferencian en su modo de proceder. Los analistas de negocios se centran en la gestión del negocio, en una visión global, mientras que

los analistas de datos se centran y circunscriben en el análisis riguroso de los datos.

2.3.2 CONCEPTOS BÁSICOS

Analista de datos y analista de negocios

Las principales tareas del analista de datos es la recolección, manipulación y análisis de los datos. Preparan informes que pueden adoptar diferentes formatos de visualización tales como gráficos, diagramas, infografías, *dashboards*[2], detallando los resultados significativos que se deducen de ellos.

Data Analytics

Son todas las tareas orientadas a la exploración de los datos con la intención de encontrar patrones o conocimiento útil que permita optimizar o rentabilizar un proceso de negocio.

Descubrimiento de datos (*Data Discovery*)

Si se desea responder a la pregunta ¿por qué está pasando?, la BI y la Analítica de Datos tradicional nos da respuesta a lo que está sucediendo en el negocio.

Analytics

Es la investigación, el descubrimiento y la comunicación de patrones relevantes a los datos.

Business Analytics

Se refiere a las habilidades, destrezas, conocimientos, metodologías, tecnologías y prácticas para la exploración continua e investigación del desempeño, pasado de las empresas con el propósito de revelar y comprender eventos pasados, predecir eventos futuros y evaluar posibles resultados de acciones en un contexto de negocio.

Visual Analytics

Ciencia del razonamiento analítico facilitado por interfaces visuales interactivas. El gran beneficio para las organizaciones de las herramientas de visualización analíticas es poder compartir el conocimiento con la alta administración. Las herramientas de Analítica visual hacen las decisiones inteligentes más rápidas.

***Visual Analytics for Making Smarter Decision Faster* – TDWI**

Los usuarios de negocio de hoy en día desean superar los límites de las hojas de cálculo y los informes de inteligencia de negocios buscan obtener una experiencia más rica y personalizada de los datos.

Con la explosión de los grandes volúmenes de datos y la gran variedad de datos disponibles para ellos, los usuarios de negocios desean explorar más datos y descubrir nuevas percepciones (*insights)* que puedan aplicar fácilmente para mejorar las estrategias de negocio, procesos, operaciones y compromisos con el cliente. Junto con la exploración de datos más fácil, los usuarios están tratando de aumentar la profundidad y la frecuencia de sus análisis de datos.

En lugar de consumir datos, hacen preguntas, prueban diferentes puntos de vista y enfoques, construyen modelos predictivos, etcétera.

Analítica de Datos

Analytics es una ciencia de los datos. Si BI se refiere a la toma de decisiones, DA (*Data Analytics*) es el proceso de realizar preguntas correctas. Las herramientas de *Analytics* se utilizan cuando una compañía quiere hacer previsiones o pronósticos (*forecast*) para predecir el futuro y realizar análisis de escenarios.

Las herramientas de BI transformarán esas previsiones y modelos predictivos en decisiones que impactarán directamente en el negocio, mostrándolos de manera que faciliten la comprensión rápida de las decisiones.

BI analiza el presente y el pasado. Se basa en el análisis de datos históricos. Su fortaleza es la tendencia de los datos históricos en cuanto a los resultados. Se especializa en la presentación de informes, cuadros de mando y visualización interactivas. Es un concepto pensado en el usuario final; desarrolla un sinfín de opciones para acceder a los datos.

Con BI se realiza generación de informes, gestión del desempeño, lectura de informes y todos los mecanismos válidos para evaluar el desempeño empresarial se desarrollan a partir de las tendencias con mayor penetración en el mercado.

Es la última tendencia en la Gestión Inteligente de Datos, esencialmente, no estructurados. Requiere desarrollo avanzado de matemáticas y estadística conectadas con la innovación de los medios sociales y de la Internet de las cosas.

Analytics mira hacia el futuro y la capacidad de vislumbrar el futuro es el centro. Los principales atributos son la predicción y la capacidad para realizar proyecciones, segmentaciones, modelamiento de datos.

El usuario final de BI es el negocio, en tanto que el de *Analytics* es tanto el negocio como los sistemas. BI es reactivo y *Analytics*, proactivo.

En la actualidad, se necesita tanto BI como BA, por ello se considera a BI como un superconjunto. Una empresa debe ser capaz de medir sus resultados actuales

y pasados, y tomar decisiones basadas en los resultados, y, a la vez, debe ser capaz de predecir el futuro para poder adaptarse a distintos escenarios de mejor manera y realizar planteamientos a corto y largo plazo.

2.3.3 GLOSARIO DE TÉRMINOS DE INTELIGENCIA DE NEGOCIOS (GARTNER)

La consultora Gartner en su prestigioso *IT Glossary* define los siguientes términos de *Business Intelligence* y *Analytics*.

- **Inteligencia de Negocios (*Business Intelligence*)**[3].Es un término paraguas que incluye las aplicaciones, infraestructuras, herramientas y mejores prácticas que facilitan el acceso y el análisis de información para mejorar y optimizar las decisiones y el rendimiento o desempeño.

- **Analítica de Negocios (*Business Analytics*)**[4]. Comprende las soluciones utilizadas para construir modelos de análisis y simulaciones para crear escenarios, comprender realidades y predecir estados futuros. La Analítica de Negocios incluye Minería de Datos, Analítica predictiva, análisis y estadística aplicada, y se entrega como una aplicación adecuada al usuario de negocios. Estas soluciones de Analítica, con frecuencia, vienen con contenidos de la industria preconstruida, que se dirige a un proceso de negocio de la industria (por ejemplo, reclamos, suscripción o un requisito regulatorio específico).

- **Analítica avanzada**[5]. Es el examen del contenido o datos autónomos o semiautónomos utilizando técnicas y herramientas sofisticadas, normalmente más especializadas que las tradicionales de Inteligencia de Negocios, para descubrir conocimientos (ideas, *insights*) más profundas, hacer predicciones o generar recomendaciones. Las técnicas de Analítica avanzada incluyen Minería de Datos y de texto, aprendizaje automático, reconocimiento de patrones, previsiones (*forecasting*), visualización, análisis semántico, análisis de sentimiento, análisis de redes y clústeres, estadística multivariante, análisis de gráficos, simulación, procesamiento de eventos complejos, redes neuronales.

- ***Analytics* (Analítica)**[6]. Ha emergido como un término general para una variedad de iniciativas relacionadas con aplicaciones e Inteligencia de Negocios. Para algunos, es el proceso de analizar la información de un dominio específico, como la Analítica de sitios web o Analítica web. Para otros, es la aplicación de las grandes capacidades de BI a un área de contenido específico (por ejemplo, ventas, servicios, cadena de suministro, etcétera). En particular, los vendedores de BI utilizan el término "*Analytics*"

para diferenciar sus productos de la competencia. Cada vez con mayor frecuencia, la Analítica se utiliza para describir el análisis de datos matemáticos y estadísticos que agrupa, segmenta resultados y predice cuáles son los escenarios más probables de suceder. Cualquiera sea el caso de uso, la *Analytics* se ha desplazado con gran fuerza a los núcleos vitales de los negocios. Analytics ha ganado un creciente interés en los negocios y en los profesionales de TI que buscan explotar enormes montículos de datos generados internamente y disponibles externamente.

Nota. En el día de la última consulta (mayo 2018), curiosamente, los términos *Data Science* y *Data Analytics* no figuran como términos independientes, sino asociados a otros términos, como es el caso de *Data Scientist*.

2.4 ANALÍTICA AVANZADA (AA)

Otro término utilizado en los últimos años por las empresas proveedoras de *software* de Inteligencia de Negocios y que ha venido a "enredar" más la jerga de la Analítica y la Inteligencia de Negocios, pero que deseamos hacer constar por la frecuencia de uso de los grandes fabricantes, ha sido el de **Analítica Avanzada** (Advanced Analytics). Gartner, Microsoft, IBM y Oracle han utilizado el término indistintamente. En el caso de Gartner como ya hemos comentado, su último cuadrante mágico de Analítica Avanzada fue el año 2016 y ya en la edición de 2017 y 2018 ha optado por cambiar directamente el término por Data Science (Ciencia de Datos). Nuestra opinión coincide con Gartner y nos quedaremos con los dos términos soporte del contenido del libro, Inteligencia de Negocios y Analítica de Negocios (en su caso Analítica o Analytics) con las diferencias que estamos señalando.

Aunque Gartner ya ha cambiado el término, muchos otros proveedores, consultores y asesores siguen utilizando Analítica avanzada, por lo que hemos optado por seguir una infografía —creemos que muy elocuente— realizada por Microsoft sobre esta disciplina.

Microsoft plantea que la Analítica avanzada es utilizar los datos de manera correcta para obtener información de valor que permita tomar mejores decisiones de negocio y en la infografía diseñada para ver las diferencias entre Inteligencia de Negocios y Analítica avanzada (figura 2.1) se pueden apreciar las diferencias entre ambas disciplinas.



Figura 2.1. Inteligencia de Negocios *versus* Analítica avanzada
Fuente: Microsoft: https://news.microsoft.com/es-xl/diferencias-entre-inteligencia-de-negocios-y-analisis-avanzados-infografia/#sm.0000klxjikbblf91ucz2acedwdrwb#lz6GZb581fw0WQPH.97

La Inteligencia de Negocios (BI) realiza el análisis del pasado (histórico) para describir y diagnosticar. En esta capa incluye: análisis descriptivo (¿Qué pasó?: análisis de la información histórica) y análisis de diagnóstico (¿Por qué sucedió?: análisis de retrospección).

Los análisis avanzados son métodos cuantitativos sofisticados para la predicción y prescripción del futuro. Hay de dos tipos: *análisis predictivo* (¿Qué pasará?) y *análisis prescriptivo* (¿Cómo hacemos que suceda?). Buscan la optimización mediante la comprensión y la prospección.

Microsoft resume su concepto de análisis avanzados: "Los análisis avanzados comprenden estadísticas, análisis descriptivo y predictivo, simulación, optimización y métodos cognitivos, puestos en marcha para describir la información de valor que la Inteligencia de Negocios no obtiene. Todo esto mediante capacidades de *Big Data* para obtener conocimientos (*insights*) de

todas las formas de datos (estructurados, no estructurados y semiestructurados) y administración de información para convertir esta información en valor en procesos de negocio".

Beneficios de la AA según Microsoft

Búsqueda proactiva de patrones e información de valor a partir de los datos y la habilidad de actuar en consecuencia. Se ha de pasar del análisis de lo que sucedió a predecir lo que pasará. Se trata de cambiar modelos manuales de toma de decisiones a procesos automatizados que aceleran a los negocios y crean ventajas competitivas.

Herramienta de BI y Analítica de Microsoft

La herramienta de Microsoft de BI y Analítica es Microsoft Cortana Analytics. Es un sistema dotado de tecnologías e infraestructuras para Analítica avanzada; incluye capacidades de aprendizaje automático, almacenamiento de *Big Data*, procesamiento en la nube con inteligencia perceptual, por ejemplo, reconocimiento y análisis de visión, voz y rostro, con la meta de ayudar a los negocios a acelerar sus procesos. La inteligencia perceptual permite la interacción con los clientes de nuevas formas.

Inteligencia de Negocios moderna

La Inteligencia de Negocios moderna se sustenta en la Inteligencia de Negocios tradicional de manipulación de datos estructurados y evolucionará al BI moderno, donde prevalecerá la Analítica de Negocios con especial incidencia en el análisis de datos y, en especial, *Big Data*, que será la fuente de alimentación de la Inteligencia de Negocios.

2.5 CASO DE ESTUDIO: PLATAFORMAS DE BI *& ANALYTICS* DE GARTNER

La consultora tecnológica Gartner, con presencia a nivel mundial, publica anualmente numerosos informes sobre todo tipo de tecnologías, herramientas y aplicaciones de gestión empresarial. Uno de ellos, y de los más populares, son los cuadrantes mágicos, informes que evalúan a las empresas más influyentes en función de unos indicadores determinados y agrupa a las empresas evaluadas en cuatro grandes categorías: empresas líderes, visionarios, de nicho y aspirantes a líderes. Uno de los cuadrantes mágicos más populares es el de Inteligencia de Negocios, cuyo objetivo principal es la evaluación de Plataformas de Inteligencia de Negocios (*Business Intelligence Platforms*) y la presentación de información y las funcionalidades analíticas de usuarios y fabricantes especializados en *software* de BI.

En el capítulo 1 se realizó una introducción al cuadrante mágico de Gartner de Inteligencia de Negocios y Analítica, una de las herramientas de consultoría más acreditadas en el sector de negocios. En particular se analizó el cuadrante mágico de Gartner de *BI* y *Analytics Platforms* de 2017 y 2018.

En este apartado y para ver la evolución del cuadrante y, en consecuencia, de *BI* y *Analytics*, analizaremos el cuadrante de 2012 (último año en el que Gartner denominaba al informe con el título *Business Intelligence Platforms*) y 2016 (donde ya destacaba el impacto del BI móvil y herramientas de visualización: *reporting* y *dashboards* -el capítulo 7 se dedica a visualización-*).* Gartner quiso resaltar con el actual nombre (*Business Intelligence & Analytics Platforms*) vigente desde 2013 la importancia de la nueva disciplina *Analytics* integrada totalmente con *Business Intelligence* y también su carácter autónomo en la nueva era de los "datos" como herramienta indispensable para organizaciones y empresas.

Gartner define una Plataforma de BI como una plataforma de *software* que ha de cumplir con 14 capacidades y que agrupa en tres categorías de funcionalidades: integración, entrega de información y análisis. En concreto, las categorías y capacidades que Gartner define son la mayoría de las herramientas que hemos analizado —o analizaremos posteriormente— en el apartado de la arquitectura de Inteligencia de Negocios:

1. Integración
   - Infraestructura BI.
   - Manejo de metadatos.
   - Herramientas de desarrollo.
   - Colaboración.
2. Entrega de información
   - *Reporting* (informes o reportes).
   - Cuadros de mando (*dashboards*).
   - Consultas *ad hoc*.
   - Integración con Microsoft Office.
   - Búsqueda basada en BI.
   - BI móvil (*Mobile BI*).
3. Análisis
   - Procesamiento analítico *online* (OLAP).
   - Visualización interactiva.

- Modelos predictivos y Minería de Datos (*Data Mining*).
- *Scorecards.*

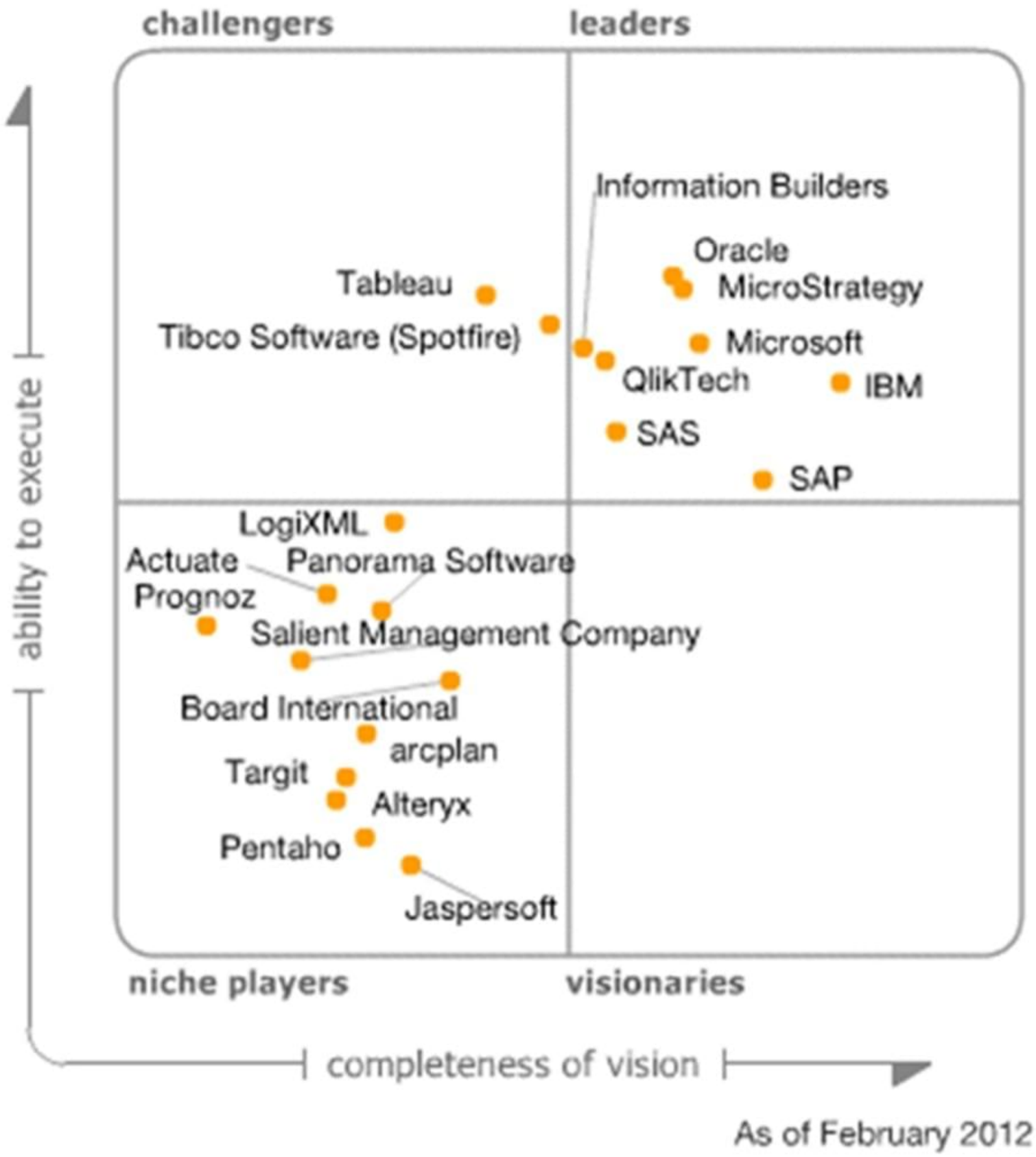


Figura. 2.2. Cuadrante mágico de Gartner para Plataformas de *Business Intelligence*. Fuente: Gartner; Febrero, 2012

Los proveedores de soluciones de Inteligencia de Negocios más destacados en 2012 eran: IBM, Jaspersoft (código abierto), MicroStrategy, Pentaho (código abierto), QlikView, SAP, SAS, Tableau, Oracle, Information Builder, entre otros.

2.5.1 CUADRANTE MÁGICO DE GARTNER DE *BUSINESS INTELLIGENCE& ANALYTICS PLATFORMS* 2013 Y SIGUIENTES

Gartner, reconociendo la importancia de la evolución de las herramientas de Inteligencia de Negocios hacia su capacidad de análisis y predicción, decidió

cambiar su nombre tradicional del cuadrante mágico de Plataformas de Inteligencia de Negocios por el nuevo de Plataformas de Inteligencia de Negocios y Analítica (*Business Intelligence &* Analytics *Platforms*) a partir de la edición de 2013. Dada la evolución que está sufriendo el binomio Inteligencia de Negocios y Analítica de Negocios (Analytics), también Gartner en 2016 mantiene el cuadrante mágico de BI y Analytics *Platform* y en su estudio define el nuevo modelo de Inteligencia de Negocios al que llama **BI moderno***("modern BI")*, pero añade un nuevo informe complementario más centrado en los proveedores tradicionales de *reporting* corporativo, análisis de consultas *adhoc*, procesamiento analítico multidimensional OLAP, denominado *Market Guide for Enterprise-Reporting-Based Platforms 2016*.

2.5.2 CUADRANTE MÁGICO DE GARTNER DE PLATAFORMAS DE INTELIGENCIA DE NEGOCIOS Y ANALÍTICA 2016[7]

El cuadrante mágico de Gartner ha cambiado su estructura de una visión centrada en IT (tecnologías de la información) a una visión centrada en el negocio, en un nuevo modelo conocido como Inteligencia de Negocios moderno. El cuadrante mágico de Gartner mantiene el número de 14 capacidades pero las redefine con nuevos nombres y define, a su vez, cinco nuevos grupos funcionales:

Infraestructuras

1. Administración de la plataforma de BI.
2. BI en la nube.
3. Administración de la seguridad y del usuario.
4. Conectividad frente a las fuentes de datos (contempla la posibilidad de que los usuarios se pueden conectar a datos estructurados y no estructurados, contenidos en diferentes tipos de plataformas, tanto en la empresa (*onpremise*) o en la nube.

Gestión de datos

1. Gobierno y gestión de datos.
2. Herramientas ETL más interactivas y almacenamiento de datos.
3. Preparación de datos en autoservicio.

Análisis y creación de contenidos

1. Analítica avanzada embebida.

2. *Dashboards* de Analítica fáciles de funcionamiento.
3. Exploración visual interactiva.
4. Exploración y creaciones de autor móviles.

Compartición de recursos

1. Contenido de Analítica embebida.
2. Contenido de Analítica de edición.
3. Inteligencia de Negocios social y colaborativa.

Las empresas proveedoras de Inteligencia de Negocios líderes en 2016 son Microsoft, Qlik y Tableau. En segundo nivel "visionarios" se destacan SAP, SAS, IBM, MicroStrategy, Oracle, y Pentaho. Otros proveedores destacados son Birst, Information Builders y Salesforce.



Figura 2.3 Cuadrante mágico de Gartner de *Business Intelligence & Analytics Plataforms* 2016

2.6 ORGANIZACIÓN, TIPOS Y FUENTES DE DATOS

Los sistemas de Inteligencia de Negocios tienen un subsistema de gestión de datos que se encarga de las diferentes fases por las que han de fluir los datos desde las fuentes externas e internas del sistema de información de la empresa hasta los almacenes de datos donde serán analizados y, posteriormente, puestos a disposición de los usuarios con herramientas de visualización de datos.

El análisis de los datos, tanto los tradicionales de los archivos y bases de datos, integrados en los almacenes de datos, como los grandes volúmenes de datos, almacenados en las infraestructuras de procesamiento y almacenamiento de *Big Data*, exige el conocimiento previo de los diferentes tipos de datos y los formatos en que se almacena la información, así como las propiedades fundamentales que deben cumplir los datos en su proceso y ciclo de vida.

2.6.1 TIPOS DE DATOS

Los tipos de datos manejados hoy en día por los sistemas de Inteligencia de Negocios se dividen en dos grandes grupos, en función de su estructura: **estructurados** y **no estructurados**. A su vez, los datos **no estructurados** se subdividen en otras dos categorías: **no estructurados** —propiamente dichos— y **semiestructurados**. De esta forma, las clasificaciones más frecuentes de tipos de datos son: **estructurados**, **no estructurados** y **semiestructurados**. La mayoría de las bases de datos relacionales existentes en la actualidad —aunque todavía quedan bases de datos heredadas (*legacy*) que siguen otros modelos— contienen datos estructurados en formato tabla y las bases de datos modernas que implementan *Big Data* contienen datos no estructurados y semiestructurados, aunque pueden contener también datos estructurados.

Datos estructurados

Los datos estructurados son aquellos que tienen una estructura predefinida y fija. Se localizan en un campo fijo de un registro o archivo determinado. Los datos estructurados nombran a cada campo de la base de datos y determinan las relaciones entre los campos. Son fáciles de introducir, almacenar, procesar y analizar (se almacenan en filas y columnas, en tablas); se almacenan en bases de datos y hojas de cálculo. Las bases de datos relacionales sólo tienen capacidad para almacenar y analizar datos estructurados. Los datos estructurados se organizan en modelos de datos con diferentes campos y relaciones entre ellos, que se han de crear para la introducción, almacenamiento, procesamiento, análisis y visualización de los datos. Son datos estructurados:

- Datos de estudiantes, empleados, clientes.
- Datos financieros.
- Datos logísticos.

Un modelo de datos es aquel que contiene los tipos de datos de las organizaciones y empresas que se registran, se almacenan, se procesan, se analizan y se pueden acceder a ellos. Así, por ejemplo, en una base de datos de clientes de una empresa tradicional o de comercio electrónico, se deben establecer los tipos de datos (alfabéticos, numéricos) y los campos determinados (nombres, dirección, número de teléfono de contacto, dirección postal, dirección de correo electrónico, etcétera.), así como las convenciones o restricciones que se pueden almacenar (por ejemplo, el número de teléfono sólo puede aceptar información numérica, rango de valores numéricos, etcétera). También se podrán incorporar menús desplegables que limiten las opciones de los datos que se pueden introducir en un campo, tal como el campo "tratamiento", que puede tener las siguientes opciones: "Sr.", "Sra.", "Doctor/a", "Ingeniero/a", "licenciado/a", etcétera.

Los datos estructurados nombran los campos de las bases de datos relacionales y determinan las relaciones entre los campos. La ventaja más notable de los datos estructurados es su facilidad para introducir, almacenar, analizar y visualizar en una base de datos. Las tecnologías de bases de datos relacionales y hojas de cálculo sólo tienen capacidad para almacenar, procesar y analizar datos estructurados; el resto de los datos no se pueden procesar y deberán guardarse en otros soportes físicos (carpetas, archivadores) u otros dispositivos magnéticos.

Los datos estructurados se gestionan, normalmente, con el lenguaje de programación SQL (*Structured Query Language*), creado en la década de los setenta por IBM, y sigue siendo el lenguaje estándar de gestión de las bases de datos.

Datos no estructurados

Los datos no estructurados o sin estructurar son aquellos que no contienen una estructura definida y no se pueden incluir fácilmente en columnas, filas y campos; es decir, su formato no se puede indexar con facilidad en tablas relacionadas para su análisis. Ejemplos de datos no estructurados son:

- Imágenes, fotografías, dibujos.
- Audio.
- Video.
- Publicaciones en redes sociales.
- Archivos de texto, documentos, presentaciones visuales.
- Correos electrónicos.
- Archivos PDF.

- Entradas de redes sociales.
- Hojas de cálculo.
- Publicaciones en redes sociales.

El proceso de datos no estructurados por un computador es muy complejo y el almacenamiento de este tipo de datos requiere unas bases de datos más complejas que las tradicionales. Estas bases de datos son las denominadas bases de datos NoSQL y analíticas “en memoria”, y no tienen formato fijo.

Datos semiestructurados

Los datos semiestructurados contienen propiedades de los datos estructurados y los datos sin estructurar. Pueden tener alguna estructura, pero no la estructura estricta de un modelo de datos. En los datos semiestructurados se utilizan etiquetas y otros marcadores para identificar algunos de sus elementos, pero no tienen una estructura rígida. Por ejemplo, datos etiqueta/valor:

<familia>= Lucas, <madre>= Antonia, y <hija>= Victoria

Algunos ejemplos de datos semiestructurados son:

1. Una publicación de Facebook puede clasificarse por autor, información, longitud, opiniones personales, etcétera, pero el contenido no está estructurado.
2. Un *software* de tratamiento de textos que incluya metadatos como en las bases de datos documentales; puede incluir fichas de autor, de socio, con nombre, fecha de creación, fecha de modificación, etcétera, pero estos datos no son estructurados.
3. Los datos de correo electrónico son datos sin estructurar, pero se pueden considerar híbridos, ya que tienen una parte estructurada (destinatarios, receptores, asunto) y una no estructurada (texto del mensaje).

Los datos semiestructurados están, especialmente, incluidos en archivos de texto para la Web creados con lenguajes HTML, XML y JSON, y por esta razón contienen ciertos elementos estructurados (etiquetas o marcadores).

Datos estructurados	Datos semiestructurados	Datos no estructurados
Fichas de clientes Fechas de nacimiento Nombre Dirección Transacciones en un mes Puntos de venta Tarjeta de crédito Tarjeta de fidelidad	Correos electrónicos *Parte estructurada*: Destinatario, receptores, asunto. *Parte no estructurada:* cuerpo del mensaje	Persona a persona Comunicaciones en las redes sociales. Persona a máquina Dispositivos médicos Comercio electrónico Computadores Teléfonos inteligentes Maquina a máquina (M2M) Sensores, dispositivos GPS Camas de seguridad

Tabla 2.1 Taxonomía de datos y propiedades

2.7 CICLO DE VIDA DE LOS DATOS

El ciclo de vida de los datos en un sistema de Inteligencia de Negocios es similar al ciclo de vida de los sistemas de información o sistemas de gestión de bases de datos relacionales. En el caso de Inteligencia de Negocios, se ha de considerar el ciclo de vida de datos en todos los formatos, así como los datos tradicionales de sistemas tradicionales o heredados y los nuevos sistemas de *Big Data*.

El flujo de información de los datos tiene su origen en las fuentes de datos (internas y externas) y ha de pasar por los diferentes componentes hasta llegar a su etapa final de visualización de datos y presentación de resultados. Las etapas de un sistema de gestión de datos son:

- Recolección de datos.
- Almacenamiento.
- Procesamiento y análisis de datos.
- Visualización de datos.

2.7.1 RECOLECCIÓN DE DATOS

Las organizaciones se enfrentan a una cantidad enorme de datos procedentes de fuentes internas y fuentes externas, como parte del plan de Inteligencia de Negocios. Se debe desarrollar en primer lugar el citado plan de IN, de modo que se puedan definir un conjunto de objetivos por conseguir y, en consecuencia, decidir cuáles son los datos que se necesitan para cumplir esos objetivos.

Herramientas de recolección de datos

La etapa de recolección de datos se refiere a la obtención de los datos. En esta etapa, el sistema se conecta a las diferentes fuentes de información para extraer los datos que luego se han de almacenar, procesar, analizar y visualizar. Las herramientas de recolección de datos se agrupan en dos categorías, en función del modo en que se conectan al origen de datos:

- ***Batch* (por lotes).** Este tipo de recolección de datos se conecta cada cierto tiempo a las fuentes de información (archivos y bases de datos), en las que se buscan cambios en la última conexión realizada.
- ***Streaming* (en tiempo real).** Este tipo de recolección de datos trabaja en modo directo, "continuamente", con las fuentes de información de un modo continuo, de forma que la información se obtiene cada vez que se solicita (o tramita), es decir, en tiempo real.

Fuentes internas

Los datos internos de las organizaciones son todas aquellas que las organizaciones y las empresas tienen a su disposición o podrían tener acceso en la actualidad. Las organizaciones normalmente disponen de registros de contabilidad, activos financieros, recursos humanos, etcétera, almacenados en sus sistemas operacionales (también denominados **OLTP**, *Online Transaction Processing*), que operan sobre bases de datos relacionales que se optimizan para soportar las relaciones entre los datos (por ejemplo, una clave común, como puede ser el DNI o número de pasaporte, en algunos países, como Estados Unidos, se utiliza el número de seguridad social que permite registrar e identificar a las personas relacionadas). Ejemplos de datos típicos son:

- Datos de registros de clientes.
- Datos de empleados.
- Datos de ventas.
- Datos de control de existencias.
- Comentarios de clientes.
- Datos de operaciones.

Fuentes de datos externas

Las fuentes de datos externas pueden ser públicas o privadas y su procedencia es muy variada:

- Datos meteorológicos.
- Datos de organismos oficiales (censos, Instituto Nacional de Estadística, administraciones públicas, etcétera).
- Datos de Twitter, de Facebook o de LinkedIn.
- Datos de perfiles de redes sociales.
- Datos de Google Maps.
- Datos de Google Trends.

Datos de la nube

Uno de los sistemas más utilizados en la actualidad para el almacenamiento de datos personales y corporativos es la nube (*cloud*). El acceso a la nube se realiza mediante los oportunos sistemas de identificación personal o corporativa.

Sin embargo, es preciso tener presente que los proveedores de la nube permiten el acceso a datos de otros clientes —siempre que éstos legalmente permitan dicho acceso—, para los cuales estos proveedores ponen a disposición de todos sus clientes las interfaces de programación de aplicaciones (API) correspondientes, para poder acceder y disponer de esos datos externos. Además de los proveedores comerciales (LinkedIn, Facebook, Twitter, Instagram, Google, Amazon, Salesforce), existen otros proveedores que personalizan sus propios datos y los ponen a disposición de los usuarios en general, como suelen ser los casos de la NASA o el INE. Estas situaciones se suelen manifestar en las tendencias conocidas como Datos Abiertos (*Open Data*) que, fundamentalmente, son datos de las administraciones y gobiernos e instituciones públicas, aunque también pueden ser puestos a disposición por entidades privadas.

2.7.2 ALMACENAMIENTO

Estos datos normalmente se almacenan en bases de datos y representan un aspecto fundamental de la organización de una solución de Inteligencia de Negocios. **Estos datos son estructurados y pueden ser analizados eficientemente.** Los componentes y conceptos implicados en el almacenamiento de datos estructurados internos son:

- Modelos de datos.
- Bases de datos (repositorios de datos).

- Sistemas de gestión de datos (*software* del sistema de gestión de bases de datos, que permite a los usuarios las consultas a las bases de datos, tales como SQL y MySQL).
- Almacenes de datos (*Data Warehouse*). Integración de varias bases de datos de una organización que se utiliza para informes (*reporting*) y análisis (*analysis*). Por ejemplo, una base de datos de reservas de una línea aérea (AeroMéxico, LAN, Avianca, Iberia) permite reservar y emitir billetes (tiquetes) y hacer un análisis de tendencias, encontrando patrones individuales de los viajeros.
- Sistemas de bases de datos no relacionales (NoSQL) y "en memoria" y de procesamiento masivo paralelo (MPP). Aquí se incluyen también todas las bases de datos integradas en los sistemas Hadoop y Spark de *Big Data.*

La situación actual de las soluciones de Inteligencia de Negocios es que los datos, en su mayoría, son **datos no estructurados** y, en consecuencia, no siguen modelos de datos definidos, como suelen ser los casos de correos electrónicos, páginas de redes sociales, tuits, informes de páginas web, etcétera. Estos datos, como ya hemos comentado, no cumplen los requisitos de formato de un modelo de datos relacionales y son difíciles de almacenar y analizar.

La solución a este problema es almacenar estos datos en bases de datos no relacionales (NoSQL), bases de datos "en memoria" y bases de datos distribuidas de almacenamiento paralelo masivo (MPP) de infraestructuras de *Big Data*. Las bases de datos NoSQL son diseñadas para almacenar grandes cantidades de datos no estructurados, pero también pueden almacenar datos estructurados o conectarse a bases de datos relacionales de datos estructurados.

2.7.3 PROCESAMIENTO Y ANÁLISIS DE LOS DATOS

Una vez que los objetivos de Inteligencia de Negocios se han definido y se han seleccionados las fuentes de datos, se ha de decidir el proceso de análisis de datos. Los datos se pueden almacenar, en su totalidad, en la propia empresa (*inhouse*, *onpremise*) o en un proveedor de la nube. En consecuencia, el análisis se puede realizar en la propia empresa o en la nube. La elección del lugar donde residirán las aplicaciones y los datos requiere necesidades técnicas y de negocios:

- Los datos deberán cumplir con unas normas específicas legales y operacionales. Por ejemplo, en el caso de la Unión Europea, los datos almacenados en servidores de empresas u organizaciones y de la nube deben estar situados en países de la Unión Europea o en países que cumplan con sus normativas de privacidad y protección de datos.
- La tecnología disponible puede exigir que el análisis se realice fuera de la empresa por razones de potencia de computación.

- Se pueden comprar datos externos de un proveedor de la nube, y acceder a ellos vía una API, y requieren un cierto tipo de análisis. En este caso, la compra de datos debe ser legal y cumplir con todas las normativas nacionales e internacionales.
- La empresa deberá tener presente los riesgos financieros en el despliegue de datos y aplicaciones en un proveedor de la nube.

Todas estas premisas conducen a la utilización de herramientas de Analítica que ayudan al análisis de los datos para su soporte en la toma de decisiones:

- **Informes y reportes descriptivos.** Consultas y cuestiones básicas que implican informes tabulares clásicos, que utilizan el análisis de los datos históricos. Un caso típico suele ser el de los informes de ventas mensuales por provincias o regiones.
- **Procesamiento analítico en línea (OLAP)**
  - Cubos de datos.
  - Operaciones de *slicing* y *dicing*.
- **Minería de Datos**
  - Asociaciones.
  - Secuencias.
  - Agrupamiento.
  - Clasificación.
- **Análisis de datos no estructurados**
- **Minería de texto**
  - Lenguaje de Procesamiento Natural (NLP).
- **Minería web de uso**
- **Analítica predictiva**

2.7.4 VISUALIZACIÓN DE DATOS

El proceso del ciclo de vida de los datos termina con la presentación de resultados al usuario mediante una interfaz de usuario soportada en herramientas de visualización de datos que faciliten la interpretación de resultados y la toma de decisiones posterior. Tradicionalmente los directivos y profesionales de cualquier nivel empresarial utilizaban los escritorios de las computadoras y dispositivos portátiles y móviles con el software ordinario de ofimática. Todos los paquetes integrados de software (*suites*) de análisis de

negocio cuentan con herramientas de visualización de datos ordinarias (gráficos, diagramas, tableros de control, mapas, etc.)

Las nuevas herramientas de visualización de datos aportan grandes beneficios frente a los típicos informes de filas, columnas y tablas de información. La visualización de datos es la presentación de datos y de los análisis realizados con el objeto de que los responsables de la toma de decisiones puedan entender todos los conceptos y conocimientos necesarios para identificar nuevos patrones de comportamiento.

La visualización de datos ayuda a las personas a descubrir patrones, tendencias, relaciones y eventos que se puedan producir analizando datos ordinarios y complejos, mediante la transformación de datos en un contexto visual. La visualización puede tomar diferentes formas incluidas tablas dinámicas, gráficos circulares, gráficos de líneas, gráficos de columnas, mapas, tableros y cuadros de control, o las modernas herramientas de narración de historias (data storytelling). Así mismo se trata de identificar los indicadores clave de rendimiento (KPI) más relevantes para cada sector, organización o departamento.

La visualización de datos puede representar la información de la manera más intuitiva para conseguir una comunicación simple, clara y efectiva, de modo que el conocimiento se pueda extraer de manera efectiva y eficiente. Las modernas herramientas de cuadros de mando o tableros de control (dashboards) permiten monitorizar el rendimiento de un área o departamento, o monitorizar operaciones y procesos con el objetivo de realizar una explotación de los datos lo más eficiente posible. Dada la importancia en la inteligencia de negocios moderna de la visualización de datos ampliaremos estos conceptos, técnicas y herramientas en el capítulo 7, dedicado exclusivamente a esta disciplina.

Existen numerosas herramientas de visualización que iremos comentando a lo largo del contenido del libro tales como: Qlik View, QlikSesion, Tableau, Power BI, Domo, Pentaho, MicroStrategy, Business Object, etc.

2.8 ANALÍTICA DE DATOS: CONCEPTOS Y TIPOS

La **Analítica de Datos** (*Data Analytics*) es una técnica consistente en capturar, procesar y analizar los datos, con el objetivo de predecir y anticiparse al futuro en la toma de decisiones. Pretende obtener conclusiones sobre la información, con el propósito de encontrar patrones o conocimiento útil que permita optimizar o rentabilizar un proceso de negocio. Es una disciplina antigua que trata de examinar los datos en bruto y convertirlos en conocimiento, con el propósito de que las organizaciones y empresas puedan tomar mejores decisiones empresariales. En el caso de la industria y otros sectores no empresariales, la Analítica de Datos se utiliza para verificar o reprobar modelos o teorías existentes.

Desde el punto de vista semántico, el Diccionario de la Real Academia Española (DRAE), en la última edición publicada, no contempla los términos *analítica* ni *análisis* como acepciones informáticas o de computación. Sólo aparece la acepción de procesamiento de datos, relativa al término *dato* (una de las etapas de la Analítica) y es definida como la "aplicación sistemática de una serie de operaciones sobre un conjunto de datos, generalmente por medio de máquinas, para explotar la información que estos datos representan".

El término Analítica de Datos se utiliza por los proveedores de *software* de Inteligencia de Negocios como sinónimo de Analítica de Negocios y como componente fundamental de la arquitectura de Inteligencia de Negocios. Desde la aparición de *Big Data*, se comenzó también a utilizar el término de Analítica de *Big Data (Big Data Analytics)* como el ciclo de vida completo de los grandes volúmenes de datos, desde su recolección en las diferentes fuentes de datos existentes en la actualidad hasta su presentación de resultados.

El análisis de datos se utiliza en todo tipo de ámbitos y sectores, desde el procesamiento analítico en línea (OLAP) hasta el análisis de los datos de los sistemas de información o centros de almacenamiento de datos internos en las empresas, o bien en los centros de la nube cuando se guardan allí los datos y aplicaciones empresariales. El análisis de datos moderno utiliza herramientas de visualización de datos, especialmente tableros de control (*dashboards*) o cuadros de mando (*scorecards*), para medir los indicadores de rendimiento KPI que se basan en los flujos de datos en tiempo real. El análisis en tiempo real implica análisis e informes dinámicos basados en los datos que se introducen en el sistema, instantes antes de que se efectúe una consulta o se desee un informe actualizado.

Existen tres técnicas utilizadas en Analítica de Datos[8]:

- Analítica descriptiva
- Analítica predictiva
- Analítica prescriptiva

Algunos tratadistas de Inteligencia de Negocio, consultoras y proveedoras de soluciones de *software* incluyen una cuarta técnica, denominada **analítica de diagnóstico**, que reúne algunas funcionalidades de la analítica descriptiva y algunas de la analítica predictiva, y que también analizaremos para completar la visión global del lector.

- **Analítica descriptiva**. Utiliza los datos para saber qué ha pasado o está pasando ahora en un negocio: ventas, ingresos, clientes e incluso reclamaciones.
- **Analítica predictiva**. Bebe de la anterior para poder hacer una estimación a futuro sobre qué va a pasar con un negocio.

- **Analítica prescriptiva.** Como si se tratara de peldaños de una escalera, necesitaremos de las dos analíticas anteriores, basándonos en el presente y en las predicciones de la marcha futura, para saber qué se debe hacer ahora para mejorar un negocio.

2.8.1 ANALÍTICA DESCRIPTIVA

La Analítica descriptiva consiste en almacenar y realizar agregaciones de datos históricos, visualizándolos de forma que puedan ayudar a la comprensión del estado actual y pasado de la empresa. Cuenta cómo ha funcionado la empresa hasta la fecha. En esencia, es un análisis de la información histórica.

Se consultan y visualizan datos de manera agregada provenientes de los diferentes indicadores de negocio, con el objeto de obtener una visión de lo que ha pasado y lo que está pasando.

La Analítica descriptiva es el análisis tradicional y permite a una organización responder ¿qué está pasando? ¿qué ha pasado?, y tomar decisiones basadas en los resultados.

¿Cómo funciona? La analítica descriptiva actúa sobre un sistema de almacenamiento donde se encuentran concentrados todos los datos relevantes del negocio. El sistema de datos puede ser centralizado en un servidor, como el caso de las bases de datos relacionales tradicionales (SQL), o sobre un sistema de archivos (ficheros) distribuidos, como pueden ser Hadoop, bases de datos **NoSQL** o bases de datos "en memoria". Sobre esta infraestructura se realiza el procesado de los datos —bien por *batch* (por lotes) o bien *online*—, de forma que se puedan efectuar las agregaciones y consultas necesarias para el análisis.

Una vez realizado el análisis, se aplican estrategias de visualización de datos para resumir el modelo de negocio. Se puede definir un conjunto de métricas clave (KPI) para ser visualizadas o bien definir una serie de reglas, de forma que se generen avisos o alertas automáticas cuando se desvíen de los valores esperados.

La analítica descriptiva ayuda en la reducción de costes y una gestión inteligente de la empresa y permite conocer la evolución histórica de las demandas de un determinado producto o servicio.

2.8.2 ANALÍTICA PREDICTIVA

La analítica predictiva es la aplicación de técnicas matemáticas y estadísticas avanzadas, así como el aprendizaje automático (*machine learning*), para predecir datos necesarios pero que no se visualizan directamente y están ausentes. Proporciona una información muy valiosa para dar soporte a la toma de

decisiones, aunque no las automatiza en sí mismas; las adiciones de técnicas de aprendizaje automático aumentan las posibilidades de automatización de toma de decisiones si así se desea.

El análisis predictivo se suele realizar tras el análisis descriptivo y permite predecir los posibles escenarios futuros. Estas predicciones se realizan basándose en informaciones pasadas (datos históricos) y presentes, y permite predecir acontecimientos con un alto grado de probabilidad.

Busca extraer conocimiento de los datos en forma de patrones, tendencias o modelos, que nos proporcionan una cierta certeza sobre situaciones potenciales futuras o el resultado de diferentes alternativas sobre las que tomar una decisión. Proporciona herramientas para estimar los datos del negocio que son desconocidos, o inciertos, o que requieren de un proceso manual o automático para su obtención.

La Analítica predictiva se refiere a las prevenciones que se han de tener presentes en la organización. Algunas situaciones analizadas pueden ser determinar la probabilidad de que ocurra un suceso, estimar montos cuantificables, predecir el momento en que ocurrirá un determinado cambio, etcétera.

La Analítica predictiva realiza las predicciones de datos que fortalecen las decisiones del negocio. Responde a las preguntas ¿qué va a pasar? ¿qué es lo que podría pasar?

2.8.3 ANALÍTICA PRESCRIPTIVA

El conocimiento adquirido mediante los modelos de analítica descriptiva y predictiva se capitaliza con la analítica prescriptiva, mediante la optimización del proceso de toma de decisiones. La analítica prescriptiva recopila información del negocio y realiza una integración completa en el mismo, teniendo en cuenta no sólo los datos del negocio sino cómo han impactado las decisiones tomadas en las cuentas de resultados y en los beneficios, y aconseja cuáles son los aspectos que se han de contemplar en las acciones futuras por realizar.

La analítica prescriptiva es la parte de la analítica que se dedica a la automatización de la toma de decisiones, utilizando dos técnicas principalmente: **los sistemas de gestión de reglas del negocio**, cuando el proceso de decisión no es sistemático, y **la optimización matemática** mediante algoritmos —especialmente de aprendizaje automático— cuando el proceso es sistemático, y que permitirá mejorar la eficiencia operativa de las empresas.

El modelo de analítica prescriptiva lleva a efecto una integración completa con el negocio, mediante la propuesta de recomendaciones sobre las acciones que se

han de seguir para reducir costes o mejorar los beneficios tratando de conseguir, en todos los casos, un incremento del retorno de la inversión (ROI).

En la práctica, la Analítica prescriptiva está encargada de la automatización de la toma de decisiones; proporciona una gran ayuda al proceso de decisión, mejorando la eficiencia operativa de las empresas que la utilizan para diferenciarlas de sus competidoras en un mercado cambiante y competitivo.

Realización del análisis prescriptivo

Como ya se ha comentado anteriormente, la analítica prescriptiva toma como base el conocimiento y las técnicas de los análisis descriptivos y predictivos, mediante la optimización numérica y los sistemas de gestión de reglas del negocio. El sistema prescriptivo recopila información del negocio, predice, sobre la base de dicha información, el impacto que tendrán las diferentes políticas o acciones que se vayan a tomar y selecciona la política por seguir con un mayor retorno de inversión mediante el proceso de optimización (IIC-UAM).

El análisis prescriptivo se encarga de recomendar las mejores opciones, basadas en los resultados del análisis descriptivo y predictivo, y presentar el impacto potencial de las acciones recomendadas. Se basa en las aplicaciones de reglas de negocio, algoritmos, aprendizaje automático y procedimientos de modelos numéricos y computacionales. Se utilizan datos históricos, transaccionales (operacionales), en tiempo real y *Big Data*.

La Analítica prescriptiva responde a preguntas tales como: ¿Qué hacer para que pase? ¿Qué necesito hacer? ¿Cómo podemos hacer que algo suceda?

2.8.4 ANALÍTICA DE DATOS (MODELO DE GARTNER)

La consultora Gartner considera un cuarto tipo de analítica que denomina analítica de diagnóstico y que ha tomado bastante popularidad en la jerga y técnicas analítica. Es un tipo de analítica que se encuentra entre la analítica descriptiva y la analítica predictiva y que como define Gartner tiene características de ambas técnicas.

La consultora Gartner considera cuatro etapas en la analítica de datos: descriptiva, diagnóstico, predictiva y prescriptiva

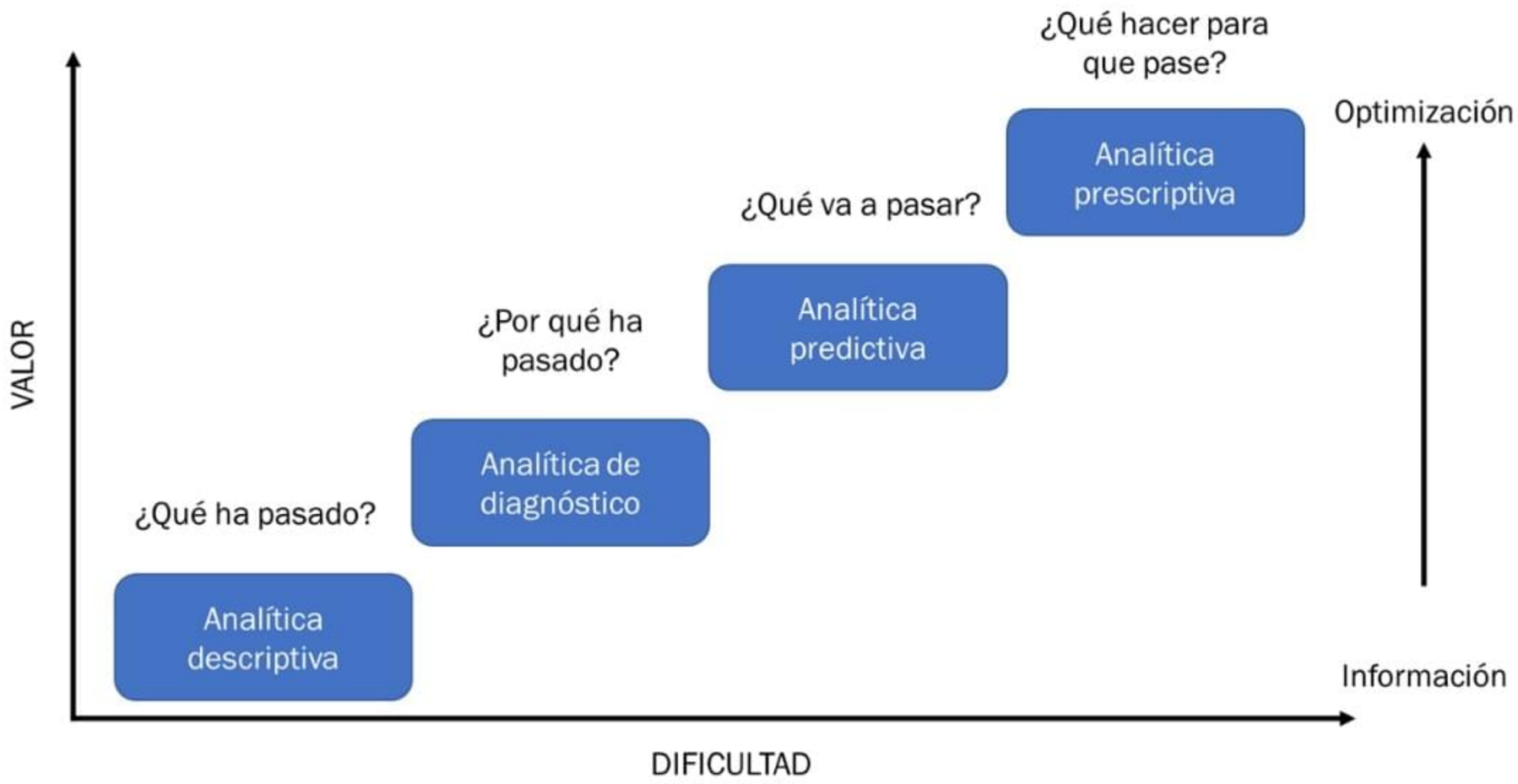


Figura 2.4. Tipos de analítica de datos
Fuente: Gartner (2013)

1. **Descriptiva:** ¿Qué ha pasado? O, lo que es lo mismo, recopilar y verificar la calidad de los datos que describen qué ha pasado en el proceso actual, es decir, las tradicionales tareas de gestión de la información.
2. **Diagnóstica:** ¿Por qué ha pasado? Para interpretar los datos y situarlos en su ámbito y contextos correctos, es necesario conocer y comprender el proceso educativo. Los datos nos ayudan a evaluar si hemos alcanzado nuestros objetivos, por ejemplo, proporcionando una explicación sobre la tasa de abandono escolar entre un grupo de estudiantes. ¿Es necesario realizar cambios en el proceso? En caso afirmativo, ¿qué cambios?
3. **Predictiva:** ¿Qué va a pasar? Nadie sabe cuáles son los indicadores que predicen que los estudiantes obtendrán malos resultados en sus estudios o abandonarán su formación. En esta fase es necesario recurrir a nuevas habilidades que combinen conocimientos del proceso educativo y el uso de enormes cantidades de datos. Con estas nuevas posibilidades podemos predecir los efectos que tendrán las diferentes medidas y usar los resultados para decidir qué vamos a hacer.
4. **Prescriptiva:** ¿Qué hacer para que pase? En este paso, llegamos a la mejor decisión, respaldada por los datos relevantes. Esto podría ser una

recomendación que apoye una decisión particular o un proceso automático que implemente directamente la decisión "calculada".

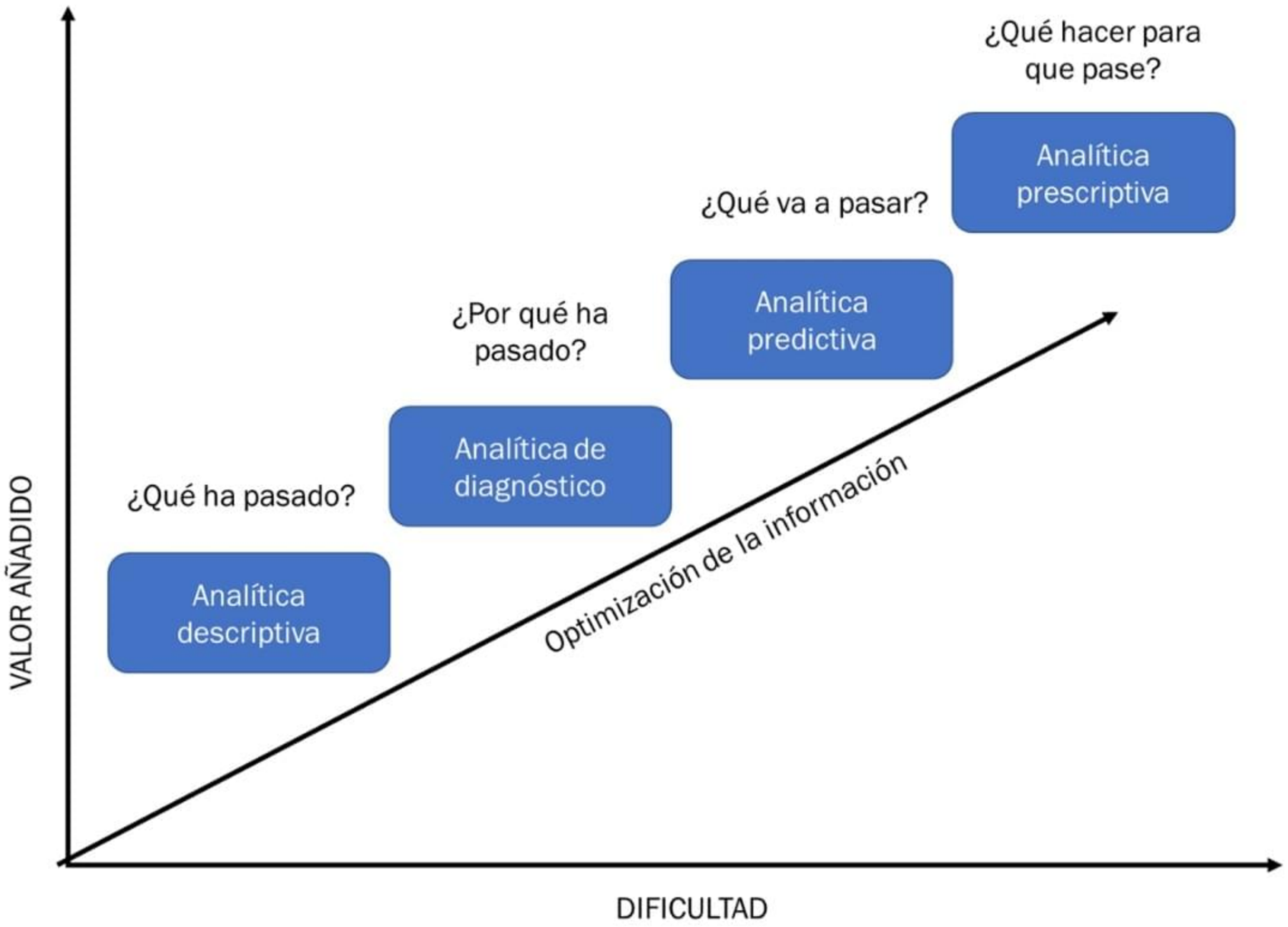


Figura 2.5. Optimización de la información en los deferentes tipos de analítica.
Fuente: Gartner (2013)

CASO DE ESTUDIO: IBM

Las soluciones de **Analítica Prescriptiva de IBM**[9] proporcionan a las empresas de comercio, servicios financieros, asistencia médica, gobierno, y otros sectores, un uso muy intensivo de los datos: "Una forma de analizar datos y transformarlos en acciones recomendadas prácticamente al momento. Estas soluciones combinan modelos predictivos, opciones de desarrollo, reglas localizadas, técnicas de puntuación y optimización para crear una base potente para la gestión de decisiones".

IBM en su página web dedicada a Analítica Prescriptiva pone algunos casos prácticos, donde su utilización resulta muy rentable en la toma de decisiones:

- Automatizar los compromisos y las decisiones complejas para gestionar mejor los recursos limitados.
- Aprovechar las oportunidades futuras o mitigar cualquier riesgo futuro.

- Actualizar de forma proactiva las recomendaciones basadas en sucesos cambiantes.
- Cumplir los objetivos operativos, incrementar la fidelidad del cliente, prevenir las amenazas y los fraudes y optimizar los procesos de negocio".

2.9 *BIG DATA ANALYITCS*

La confluencia de las tecnologías de Analítica de Datos con *Big Data* ha dado lugar al nacimiento de un nuevo tipo de analítica conocido como **Analítica de *Big Data*** (*Big Data Analytics*), que incluye los diferentes tipos de análisis de datos antes examinados y los diferentes tipos de Analítica que veremos posteriormente (cap 10 y 11), como es el caso de la analítica web, analítica de textos, analítica social, etcétera.

El volumen de grandes datos, la variedad de tipos de datos y la velocidad de datos que llegan a una organización están alcanzando niveles sin precedentes. Estos crecimientos enormes en todas sus propiedades fundamentales crean la necesidad no sólo de su conocimiento, sino, y sobre todo, de cómo obtener valor de esos grandes volúmenes de datos mediante el análisis de los mismos, con el objeto de convertirlos en conocimiento para ayuda en la toma de decisiones de las organizaciones.

SAS, uno de los grandes fabricantes de *software*, líder mundial en soluciones de programas de Minería de Datos y Analítica de Datos, define Analítica de Big Data como "el proceso de examinar grandes volúmenes de información para descubrir patrones ocultos, correlaciones desconocidas y otra información útil —como tendencias de mercado o preferencias de clientes— que se puede utilizar para tomar mejores decisiones en el ámbito de la empresa"[10]. Con la Analítica de *Big Data* se pueden analizar enormes volúmenes de datos que las soluciones analíticas y de Inteligencia de Negocios tradicionales no pueden realizar o su análisis sería tan lento que necesitaría de esperas de horas o días. Hasta hace pocos años, no sólo era difícil recoger y almacenar grandes cantidades de datos, sino que el tratamiento de estos grandes volúmenes de datos hacía que incluso su análisis fuese tan lento que era necesario esperar horas o días para obtener resultados. En consecuencia, el análisis de *Big Data* ofrece una gran cantidad de beneficios y oportunidades, así como una reducción considerable en el proceso de análisis de los mismos.

¿POR QUÉ ES TAN IMPORTANTE EL ANÁLISIS DE *BIG DATA*?

La Analítica de *Big Data* ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. SAS considera que esta característica conduce a negocios más inteligentes, operaciones más eficientes,

beneficios más altos y clientes más felices. Tom Davenport[11], uno de los grandes expertos mundiales en Analítica de Datos y en Inteligencia de Negocios, resume los grandes beneficios de utilizar Analítica de *Big Data* (figura 2.6):

- **Reducción de costos.** Tecnologías de *Big Data* tales como Hadoop y Analítica basada en la nube reducen considerablemente los costos cuando se almacenan grandes cantidades de datos en la nube y pueden identificar medios más eficientes para hacer negocios.
- **Toma de decisiones mejores y más rápidas.** Con la velocidad de Hadoop y la Analítica "en memoria", combinadas con la capacidad para analizar nuevas fuentes de datos, las organizaciones pueden identificar medios más eficientes de hacer negocios.
- **Nuevos productos y servicios.** La Analítica de *Big Data* proporciona a las compañías una gran capacidad para crear nuevos productos y cumplir con las necesidades de los clientes.

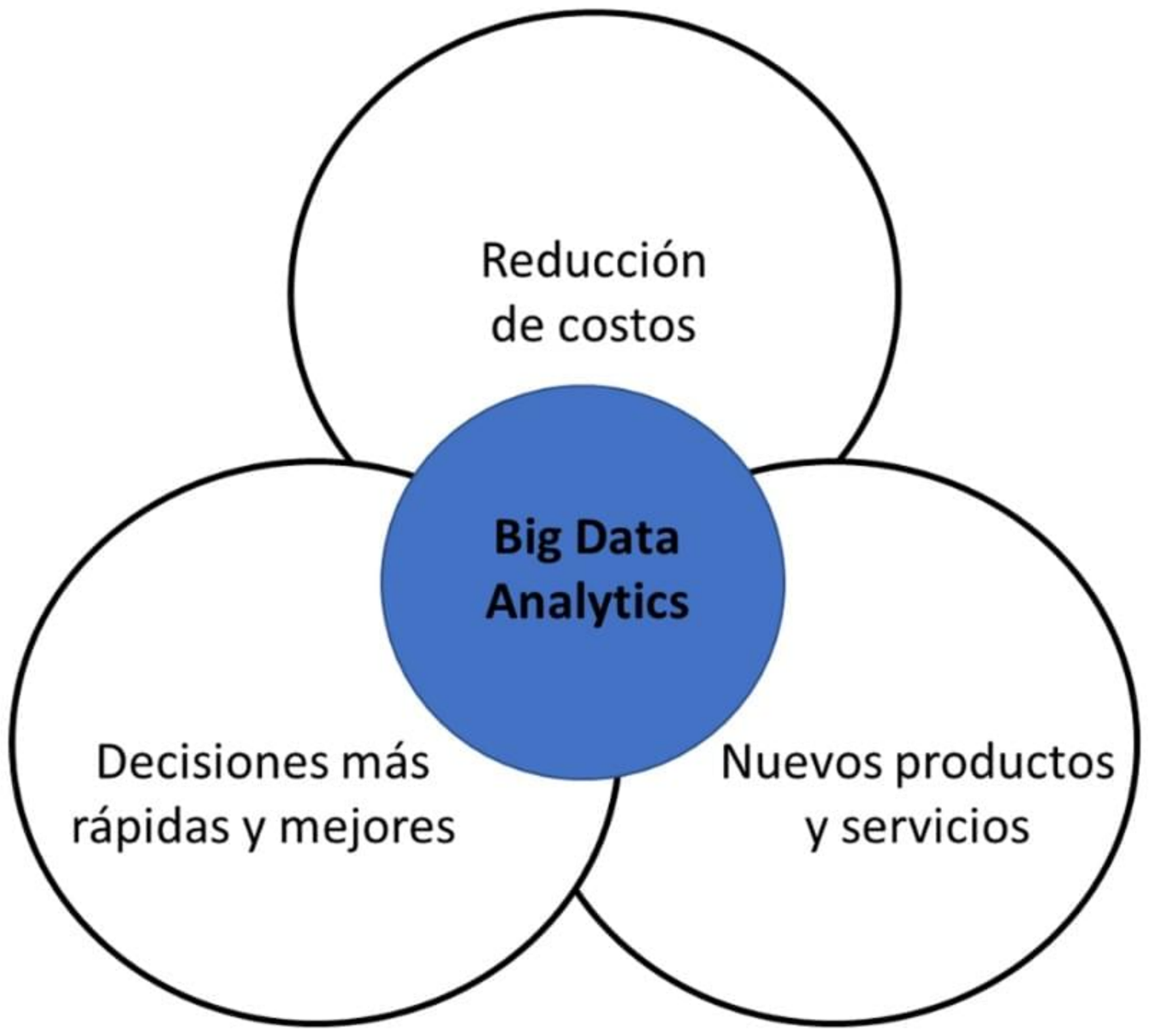


Figura 2.6. Beneficios de la Analítica de *Big Data*. Fuente: Tom Davenport (SAS)

2.9.1 TECNOLOGÍAS DE ANALÍTICA DE *BIG DATA*

SAS, en su página oficial[12] de *Big Data Analytics*, considera que este modelo de Analítica utiliza numerosas tecnologías que funcionan combinadas entre sí. Las tecnologías más relevantes, según SAS, son:

- **Gestión (administración) de datos.** Los datos necesitan ser de alta calidad y bien gobernados antes de ser analizados con fiabilidad.
- **Minería de Datos.** Las técnicas de Minería de Datos ayudan a examinar grandes cantidades de datos para descubrir patrones de datos que se utilicen en una mejor toma de decisiones.
- **Hadoop.** El marco de trabajo, por excelencia, de *Big Data* es Hadoop, que puede almacenar grandes cantidades de datos y ejecutar aplicaciones de *clústeres* de servidores para funcionar eficientemente con grandes volúmenes de datos. Un beneficio adicional es que es una plataforma basada en código abierto.
- **Analítica "en memoria".** El análisis de datos en la memoria del sistema (en lugar de en la unidad de disco duro) se realiza con gran rapidez y eficacia. Posteriormente veremos, en el capítulo 6 dedicado a bases de datos, los fundamentos de esta arquitectura y *software* "en memoria".
- **Analítica predictiva.** La tecnología de analítica predictiva, que ya hemos mencionado, utiliza datos, algoritmos estadísticos y técnicas de aprendizaje automático para identificar y predecir comportamientos basados en los datos históricos.
- **Minería de textos.** Con tecnologías de minería de textos, se pueden analizar datos de la Web, campos comentarios, libros y otras fuentes basadas en texto para descubrir conocimiento (*insights*) no observado anteriormente. La minería de textos (ver capítulo 9) utiliza aprendizaje automático o tecnologías de lenguajes de procesamiento natural para combinar documentos, tales como *e-mails*, *blogs*, entradas de Twitter, informes, etcétera, que ayudarán a analizar grandes cantidades de información y descubrir nuevos tópicos y relaciones entre ellos.

2.10 CIENCIA DE DATOS: EVOLUCIÓN DE LA ANALÍTICA DE NEGOCIOS Y EL ANÁLISIS DE DATOS

Ciencia de Datos es una disciplina científica y de negocios que ha adquirido gran notoriedad a medida que las tendencias de *Big Data* se han ido desarrollando y desplegando. Los términos *Data Science* y *Data Mining* se han utilizado, a veces, de modo intercambiable, aunque la Ciencia de Datos es una evolución gradual de la Minería de Datos y una de sus tecnologías de soporte. Ciencia de Datos, de un modo práctico, es el ámbito de conocimiento que engloba las habilidades asociadas a la extracción de conocimiento de datos, especialmente los *Big Data*. Por su importancia en el sector de Inteligencia de Negocios, dedicaremos el capítulo 13 a tratar las características más notables y en este apartado haremos una pequeña introducción para complementar la información global de analítica y

Descargado en: eybooks.com

su enlace con la visión global de inteligencia de negocios descrita en el capítulo 1.

A alto nivel en la estrategia de negocios, Ciencia de Datos (*Data Science*) es un conjunto de principios fundamentales que dirigen la extracción del conocimiento a partir de los datos y Minería de Datos es la extracción de conocimiento de los datos mediante tecnologías y métodos que incorporan estos principios. En sentido global, Ciencia de Datos se aplica más ampliamente que el uso tradicional de la Minería de Datos, pero las técnicas de minería de datos proporcionan las bases teóricas y prácticas más ilustrativas a los principios de la Ciencia de Datos.

Big Data, como ya conoce el lector, está generado por numerosas fuentes, como datos de empresa, datos públicos, datos de medios sociales, datos de transacciones y, en los últimos tiempos, datos del Internet de las cosas (procedentes de sensores). El paradigma de *Big Data* (véase capítulo 5) es un sistema en el que las grandes velocidades de transmisión de datos, su variedad y el gran volumen de datos (estructurados, no estructurados o semiestructurados) han de ser capturados, almacenados, procesados y analizados, utilizando un conjunto de técnicas y tecnologías que son muy innovadoras comparadas con las utilizadas en décadas pasadas. A la par que el nacimiento de *Big Data*, han nacido también dos ramas importantes: Ciencia de Datos e Ingeniería de datos, que son muy útiles para deducir conocimiento accesible y decisorio a partir de los datos en bruto. Al igual que ya hemos comentado sobre Ciencia de Datos y Minería de Datos, son términos intercambiables, aunque se aplican, normalmente, en dominios de experiencia distintos.

Ciencia de Datos *versus* Ingeniería de Datos

La ciencia, en general, es un método sistemático en el que las personas estudian y explican el fenómeno de un dominio específico que sucede en el mundo natural. La Ciencia de Datos, en particular, es un dominio científico que se centra fundamentalmente en el descubrimiento del conocimiento mediante el análisis de datos. En general, este dominio específico se refiere al sector industrial u otros sectores organizacionales, donde los métodos de Ciencia de Datos se utilizan para explorar y descubrir estos dominios de conocimiento.

Un científico de datos (experto en Ciencia de Datos) es un profesional que utiliza técnicas matemáticas, estadísticas y algorítmicas para encontrar soluciones a problemas complejos científicos y de negocios. En negocios, el propósito de la Ciencia de Datos es potenciar (empoderar) los negocios y las organizaciones con la información de los datos que necesitan para optimizar los procesos organizacionales, para obtener la máxima eficiencia y generación de ingresos; en ciencias, los métodos de Ciencia de Datos se utilizan para obtener resultados y desarrollos de protocolos para conseguir objetivos específicos definidos.

La Ciencia de Datos es multidisciplinar. Para ser un buen científico de datos se necesita tener experiencia en matemáticas y estadística, programación de computadoras y dominio específico del campo o materia donde se vaya a utilizar (figura 2.7). Además de estas experiencias generales, deberá conocer técnicas de aprendizaje automática e investigación tradicional (en el capítulo 14 se ampliará el estudio de Ciencia de Datos y de Científico de Datos).

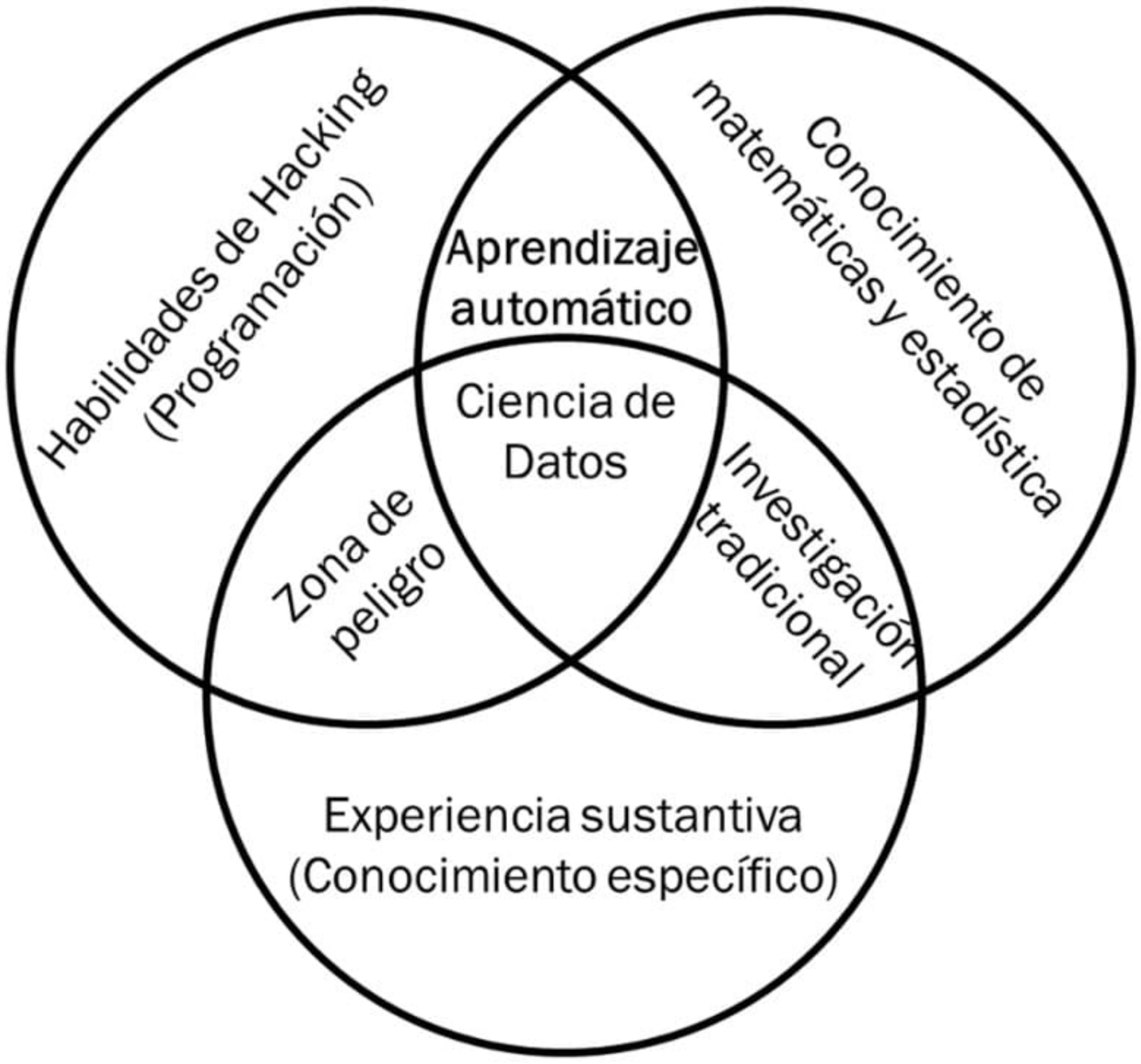


Figura 2.7. Diagrama de Venn de Ciencia de Datos
Fuente: Drew Conway (2013) (adaptado)
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

INGENIERÍA DE DATOS

Ingeniería es la práctica de utilizar la ciencia y la tecnología para diseñar y construir sistemas que resuelvan problemas y la Ingeniería de Datos es el dominio de la ingeniería que se dedica a superar el procesamiento de datos, sus problemas críticos y de manipulación de datos en aplicaciones que utilizan grandes volúmenes de datos (especialmente, *Big Data*).

Los ingenieros de datos utilizan la destreza en ciencias de la computación e ingeniería de *software* para diseñar sistemas y solucionar problemas que surgen en el proceso y manipulación de grandes conjuntos de datos. Los ingenieros de datos han de tener experiencia en las infraestructuras (marcos de trabajo) de procesamiento en tiempo real (como Spark) y plataformas de procesamiento

masivo (bases de datos NoSQL, "en memoria", bases de datos relacionales, MPP). Deben ser capaces de desarrollar código de programación como Java, C++ o Python, y deben ser expertos y conocer infraestructuras como la citada Spark y, sobre todo, Hadoop y sus sistemas de tratamiento de archivos HDFS y sistemas de algoritmos MapReduce para manejar, procesar y refinar los *Big Data* como conjuntos más pequeños y manejables. En resumen, un ingeniero de datos deberá tener formación en ingeniería, especialmente informática o de sistemas, con conocimientos complementarios de electrónica y de telecomunicaciones.

En sentido general, la Ciencia de Datos incorpora técnicas y teorías de muchos campos, incluyendo matemáticas, estadística, computación avanzada, aprendizaje automático, visualización y, según la organización, Ingeniería de Datos e ingeniería de sistemas, con el objeto de extraer el significado de los datos y la creación de productos de datos.

Data Science nace del método científico. Es la evolución de lo que antes y ahora se conoce como Analítica de Datos. Con los métodos tradicionales, la Analítica de Datos se dedicaba a analizar fuentes de datos, en general procedentes de una única fuente. En la actualidad, el científico de datos debe explorar y analizar datos de múltiples fuentes —esencialmente *Big Data*—, que pueden tener formatos diferentes y procesados a grandes velocidades, esencialmente en *streaming*. Además, el científico de datos debe tener una fuerte visión del negocio para ser capaz de extraer y transmitir recomendaciones a los responsables de dirección de la empresa. Este conjunto de datos, como ya se ha comentado, puede provenir de los datos generados por todo tipo de dispositivos electrónicos (un teléfono móvil o celular, sensores de todo tipo, secuenciadores de genoma), redes sociales, informes médicos, páginas web, etcétera, y afectarán de modo muy significativo a numerosos campos como las ciencias geológicas, biológicas, de la salud, ciencias sociales, etcétera.

2.10.1 CASO DE ESTUDIO: CIENCIA DE DATOS

La consultora Gartner, que ya hemos citado en varias ocasiones, suele publicar otro de sus cuadrantes mágicos dedicado, en exclusiva, a las plataformas de Analítica avanzada (además de los ya citados de Inteligencia de Negocios y plataformas de Analítica) para significar la importancia de la Analítica en su visión más avanzada. Dado que la Analítica avanzada ha ido variando, el cuadrante mágico de Gartner ha cambiado su nombre desde el informe de 2017. Desde este año, el cuadrante mágico se denomina, exclusivamente, *Cuadrante mágico de Gartner de Data Science Platforms* (*Advanced Analytics Platforms*). Con este cambio de nombre, Gartner quiere significar la evolución de la Analítica avanzada (en esencia, la fusión de la Analítica de Negocios con la Analítica de Datos) en la nueva disciplina que recomienda a organizaciones y empresas para su concienciación estratégica y táctica: Ciencia de Datos en la empresa.

En este primer cuadrante mágico de Gartner con su nuevo nombre, se evaluaron un conjunto de 16 empresas especializadas en Ciencia de Datos y *Analytics*, con 15 criterios o indicadores.



Figura 2.8. Gartner 2017, Magic Quadrant for Data Science Platforms

El resultado del estudio de *Data Science* de 2017 destaca los siguientes resultados:

- **Líderes (4):** IBM, SAS, RapidMiner, KNIME.
- **Aspirantes o competidores (4):** MathWorks (*new*), Quest (*formerly Dell*), Alteryx, Angoss.
- **Visionarios (5):** Microsoft, H2O.ai (*new*), Dataiku (*new*), Domino Data Lab (*new*), Alpine Data.
- **Actores de nicho (3):** FICO, SAP, Teradata (*new*).

Debemos resaltar que los proveedores líderes se encuentran muy igualados en puntuación, con ligera ventaja de IBM y SAS, seguidos muy de cerca por dos soluciones muy creativas e innovadoras, especializadas en Inteligencia de Negocios, en general, y en Ciencia de Datos, en particular (RapidMiner y KNIME).

2.11 TENDENCIAS DE *ANALYTICS*

Las tendencias en Analytics, previstas por consultoras como Gartner o Ibermática[13], son las siguientes:

El dato no estructurado, omnipresente en el mundo de la Analítica

Se calcula que los datos no estructurados corresponden, aproximadamente, a un 90% de toda la información disponible por las organizaciones y empresas. Como ya se ha comentado, estos datos incluyen videos, fotos, documentos de texto, *e-mails*, redes sociales, registros de conexión, *logs* o información procedente de sensores. Los datos estructurados describen lo que está sucediendo y se asocian, normalmente, a la Analítica descriptiva; mientras que los datos no estructurados proporcionan el porqué o las razones de los hechos.

Los roles profesionales de la Analítica de Datos

El despliegue real de *Big Data* en empresas e industrias ha potenciado la importancia del análisis de datos y la aparición de nuevas disciplinas, como la Ciencia de Datos (*Data Science*), y de roles profesionales, como analista de datos, ingeniero de datos, arquitecto de datos y científico de datos. Según las empresas consultoras, recursos humanos, de empleo, etcétera, los roles y competencias están siempre relacionados con el análisis de datos, aunque esta taxonomía puede variar según las empresas. El científico de datos (*Data Scientist*) sigue siendo la profesión más generalista y que puede contener las competencias de las restantes titulaciones con mayores o menores semejanzas. El ingeniero de datos, desarrollador de las infraestructuras de *Big Data,* será la profesión complementaria del científico de datos, de modo que ambos perfiles seguirán siendo muy demandados en organizaciones y empresas. Ambas profesiones deberán integrar técnicas y herramientas de Inteligencia de Negocios.

Gartner, y muchas otras consultoras, consideran que el futuro del análisis de datos pasa no sólo por la analítica predictiva, sino también por la prescriptiva, es decir, la automatización de toma decisiones de negocio sobre la base de las predicciones hechas sobre los datos disponibles.

Autoservicio BI y acceso inmediato a la Analítica

Los usuarios corporativos y de negocio seguirán demandando un acceso directo a la información, un reto de notable envergadura en la actualidad, y será de vital importancia el poder reducir los tiempos de preparación de los datos para el análisis. Esta tendencia se conoce como **Inteligencia de Negocios de autoservicio**. Se trata de hacer fácil de usar el análisis predictivo y que no dependa de plataformas muy complejas. Es decir, los usuarios con independencia de su nivel jerárquico deben estar en condiciones de desarrollar gráficos o cuadros de mando que les permitan tomar decisiones acertadas. Es fundamental tener la capacidad de preparar en forma de autoservicio los datos y que nos permita mezclar el mundo *Big Data* con el mundo operativo de forma sencilla.

Hoy en día, gracias al crecimiento de las tecnologías "*in-memory*" y de las bases de datos "en memoria", los procesos que conectan directamente a las fuentes de datos con las plataformas de Analítica pueden procesar millones de registros de datos para ser consultados prácticamente al instante. Estas características permiten a las compañías diseñar cuadros de mando en un tiempo **cuasi-real**, proporcionándole al usuario un conocimiento mucho mayor del que nunca tuvo sobre los datos que realmente importan.

Migración de datos a la nube

Empresas de todo tipo están llevando a cabo estrategias para migrar su negocio a la nube (*cloud*), fuera de sus centros de datos. Ante este escenario, muchas organizaciones están optando por desplegar sus plataformas analíticas en arquitecturas de la nube, especialmente el modelo híbrido (datos críticos en los servidores propios de la empresa y el resto de los datos alojados en la nube de sus proveedores de servicios donde sus datos residen), la mayor parte en una arquitectura *cloud* de tipo híbrida.

Calidad de los datos, elemento clave

El incremento masivo de datos en los sistemas de información está complicando la tarea de depurar, de separar, lo que es útil de lo que no lo es. El éxito de los modelos que elaboran los científicos de datos depende, en gran medida, de la calidad de la materia prima con la que tratan. Es por ello por lo que algunas plataformas de *Big Data* están incorporando herramientas de gobierno y calidad del dato en sus distribuciones comerciales.

La explosión de los datos de Internet de las cosas

Los miles de millones de objetos conectados entre sí a través de Internet de las cosas suponen una gran explosión de datos. Los datos procedentes de dispositivos móviles, sensores y tecnologías ponibles (*wearables*) deberán ser recopilados, almacenados, procesados, analizados y, posteriormente, se deberán visualizar sus resultados, en muchas ocasiones en tiempo real. Por esta razón, tecnologías Hadoop y Spark (en tiempo real) proporcionan grandes capacidades para las empresas. Las organizaciones están compitiendo en Analítica, no sólo porque es necesario, sino también porque deben hacerlo.

ANÁLISIS DE DATOS DE FORMATO LIBRE

Es muy importante conseguir que la información disponible en la empresa —como almacenes de datos (*Data Warehouse*) o sistemas de información ERP y CRM— posibilite llegar a mayor tipo de información contenida en los datos externos, como datos geoespaciales, bases de datos externas, información de redes sociales, sensores, etcétera, y poder integrarlos con la información de la compañía, de manera que proporcione a los empleados mayor capacidad de decisión.

Es fundamental tener la posibilidad de analizar los datos de forma libre. Debemos conseguir que la información que tenemos dentro de la empresa, como el *Data Warehouse* o el ERP, nos posibilite llegar a más información con **datos externos** y **mezclarlos con la información de la compañía**, de manera que nos dé una mejor capacidad de decisión. La capacidad de combinar datos de gran volumen con los datos internos de la empresa es una de las tendencias principales existentes en el mercado.

También es muy importante que los resultados de los análisis de datos de la compañía se puedan **embeber en aplicaciones** web, aplicaciones transaccionales, etcétera, de manera que el usuario final pueda actuar como una única aplicación que proporcione información muy relevante para la toma de decisiones, pero embebida en las aplicaciones que usa a diario. Se trata de que los datos utilizados tradicionalmente por los analistas puedan mover estos datos a empleados, clientes y proveedores, de modo que puedan aportar una información muy relevante.

NUEVAS PLATAFORMAS DE ANALÍTICA

Aparecen nuevas plataformas basadas en interfaces de programación de aplicaciones, API, que permiten integraciones que darán lugar a **nuevas herramientas** —tales como asistentes virtuales, *bots* y *chatbots* (ver capítulo 13)—. Así, la utilización de herramientas de *Geo Analytics*, Internet de las cosas o Inteligencia Artificial proporcionarán una información muy importante del comportamiento de los clientes El uso de ***Geo Analytics***, **IoT** (*Internet of Things*) o de herramientas con ***Machine Learning*** nos dará una información muy importante del comportamiento de los clientes y dará respuesta a las nuevas tendencias de industria 4.0, con la integración de aplicaciones del mundo físico y virtual (sistemas ciberfísicos).

2.12 RESUMEN

Analítica de Negocios (*Business Analytics*, BA) es un componente fundamental integrado en la arquitectura de Inteligencia de Negocios tradicional y moderna, aunque su enfoque en el análisis de datos (técnicas y herramientas) se suele

contemplar en ambientes empresariales y de consultorías como disciplina independiente.

BA es el conjunto de técnicas y herramientas que ayudan a los usuarios para aprovechar los datos empresariales con el objetivo de tomar mejores y más acertadas decisiones empresariales, en consonancia con sus líneas de negocio.

La Analítica de Negocios se apoya en la Analítica de Datos y en sus diferentes tipos de análisis.

- Las técnicas y herramientas de BA son muy numerosas: reportes (informes), consultas, Minería de Datos/web/textos, cuadros de mando o tableros de control (*dashboards*) y visualización de datos, entre otras.
- Analítica avanzada es un término que las empresas proveedoras de soluciones de *software* de BI viene utilizando desde el año 2016 para referirse a la evolución de la Analítica en la gestión de *Big Data* y la integración de las nuevas tendencias de Internet de las cosas (IoT) e Inteligencia Artificial (IA).
- La organización, tipos y fuentes de datos son la materia prima de la Analítica de Negocios. Los datos se clasifican en tres categorías: estructurados, no estructurados y semiestructurados.
- La recolección de datos utiliza herramientas de procesamiento por lotes (*batch*) o tiempo real (*streaming*).
- El ciclo de vida de los datos se compone de cuatro etapas: recolección (ingesta), almacenamiento, análisis y visualización de resultados.
- Las fuentes de datos pueden ser internas, externas y alojadas en la nube que, a su vez, pueden ser internas o externas.
- Analítica de Datos es una técnica que consiste en capturar, procesar y analizar los datos con el objetivo de predecir y anticiparse al futuro en la toma de decisiones.
- Existe tres tipos de Analítica: descriptiva, predictiva y prescriptiva. Está tomando gran fuerza una cuarta Analítica: la Analítica de diagnóstico, que trata de diagnosticar el porqué de los resultados descubiertos en la Analítica descriptiva.
- *Big Data Analytics* es un tipo de Analítica que integra los diferentes tipos de Analítica en el análisis de los grandes volúmenes de datos (*Big Data*).
- Ciencia de Datos es una disciplina emergente en el sector de los datos. Es una evolución de la Minería de Datos en el análisis de *Big Data*, al que está incorporando técnicas avanzadas de matemáticas y estadística, Inteligencia Artificial, con algoritmos y conceptos estratégicos de negocios.
- Las tendencias de Analítica del futuro se centran en roles profesionales, datos no estructurados y omnipresentes, Inteligencia de Negocios de autoservicio, migración de datos, calidad de datos, explosión de los datos

de Internet de las cosas, análisis de datos en formatos libres y nuevas plataformas de Analítica.

NOTAS

[1] *Information Builders. www.informationbuikders.es/business-Analytics*

[2] *Dashboard y scorecard, son sinónimos de cuadro de mando o tablero de control; sin embargo Balanced Scorecard, cuadro de mando integral (CMI) es una estrategia de gestión empresarial. Ambas técnicas se verán en el capítulo 7.*

[3] http://www.gartner.com/it-glossary/business-intelligence-bi/

[4] http://www.gartner.com/it-glossary/business-Analytics/

[5] *Ibidem 4*

[6] http://www.gartner.com/it-glossary/Analytics/

[7] https://www.gartner.com/doc/reprints?ct=160204&id=1-2XXET8P

[8] http://www.iic.uam.es/innovacion/analitica-prescriptiva-tercera-fase-del-big-data/

[9] http://www-03.ibm.com/software/products/es/category/prescriptive-Analytics

[10] ***Keith Collins.*** *Vicepresidente senior y director de sistemas de información, SAS. Big Data Analytics. Qué es y por qué es importante.*
https://www.sas.com/es_es/insights/Analytics/big-data-Analytics.html

[11] https://www.sas.com/es_es/insights/analytics/big-data-analytics.html

[12] *Ibidem 11*

[13] *Carlos Antón, Ibermática, 11 enero 2017.*
http://ibermaticadigital.com/tendencias-analytics-para-2017/

CAPÍTULO 3

TRANSFORMACIÓN DIGITAL EN ORGANIZACIONES Y EMPRESAS (ECONOMÍA COLABORATIVA, EXPERIENCIA DE CLIENTE Y BLOCKCHAIN)

CONTENIDO

3.1	Introducción
3.2	¿Qué es Transformación Digital?
3.3	Tecnologías facilitadoras de la Transformación Digital
3.4	La empresa digital
3.5	La Transformación Digital en la industria y en la empresa
3.6	El proceso de Transformación Digital
3.7	Fábrica inteligente: la Transformación Digital en la Industria 4.0
3.8	Economía Colaborativa
3.9	Experiencia de Cliente
3.10	Blockchain (cadena de bloques)
3.11	Blockchain en Inteligencia de Negocios
3.12	RESUMEN

OBJETIVOS

- Conocer y comprender la tendencia de Transformación Digital de la empresa.
- Definir y conocer los fundamentos de la empresa digital.
- Identificar y conocer el proceso de Transformación Digital.
- Introducción a los principios de la Economía Colaborativa, así como aplicaciones de impacto en Inteligencia de Negocios.
- Conocer las nuevas tendencias de experiencia de cliente y la necesidad de integración en IN.
- Introducción a la nueva y revolucionaria tendencia tecnológica de Blockchain, nuevo y definitivo soporte de seguridad en las transacciones y la necesidad de su integración en Inteligencia de Negocios y Analítica de Negocios.

3.1 INTRODUCCIÓN

Actualmente, casi todas las empresas están digitalizadas en mayor o menor medida. Utilizan correo electrónico, disponen de un sitio web con diferentes páginas web para sus distintos departamentos –en el caso de grandes empresas–, *blogs* para sus empleados o sus líneas de negocio, disponen de cuentas en redes sociales –en las más implantadas como Facebook, Twitter, LinkedIn, Instagram–, contabilizan sus facturas y presupuestos mediante herramientas informáticas, etcétera. Sin embargo, no todas utilizan herramientas de gestión empresarial como CRM, SCM, ERP, ni todas las empresas que se han digitalizado se pueden considerar empresas digitales, ya que esta característica requiere el cumplimiento de una serie de normas.

La digitalización de la empresa requiere de una **Transformación Digital,** que obligará a la misma a una gran reorganización y, sobre todo, concienciación de la necesidad de digitalizarse y transformarse digitalmente. En principio se analizarán los principios fundamentales de una empresa digital y el proceso de Transformación Digital, tanto para las empresas tradicionales como para las empresas industriales que conformarán la Industria 4.0 (cuarta revolución industrial).

La empresa digital ha dado lugar a un nuevo modelo de economía, **economía digital** que, a su vez, se ha sustentado en otros tipos de economías, como es el caso de la **Economía Colaborativa** y nuevos modelos de innovación: *cocreación, crowdsourcing* y *crowdfunding*, así como nuevas tendencias de gestión de relaciones con los clientes, como experiencia de usuario y experiencia de cliente.

Se estudiará el proceso de Transformación Digital y las etapas para su consecución y se analizarán los diferentes modelos de Transformación Digital en

los sectores de mayor impacto en la sociedad, la economía y la industria, así como en los sectores académicos y de investigación.

Una de las tendencias tecnológicas de impacto que está teniendo un gran impacto en los negocios es *Blockchain* (cadena de bloques) soporte de las criptomonedas o criptodivisas, como *Bitcoin* y *Ethereum*.

Por último, se describirán diferentes casos de éxito de empresas que realizaron **procesos de Transformación Digital,** que han conducido a la conversión de las compañías en verdaderas empresas digitales.

3.2 ¿QUÉ ES TRANSFORMACIÓN DIGITAL?

La Transformación Digital va más allá de que la empresa tenga presencia en Google, en las redes sociales y en la Web a través de páginas web, y naturalmente, del correo electrónico corporativo y de sus empleados. La Transformación Digital (**DX**, *Digital Transformation*) recibe su denominación por el cambio en la naturaleza de productos y servicios de las empresas tradicionales.

Existen numerosas definiciones de Transformación Digital. Margaret Rouse[1] define la Transformación Digital como "la reinvención de una organización a través de la utilización de la tecnología digital para mejorar la forma en que la organización se desempeña y sirve a quienes la constituyen. Digital se refiere al uso de las tecnologías que generan, almacenan y procesan datos". La mera implementación de tecnología por sí sola, según Rouse, no produce Transformación Digital, pero sí consigue cambiar una organización para aprovechar el potencial de las tecnologías.

Otra definición acertada es: "La Transformación Digital es la oportunidad estratégica de incorporar nuevas tecnologías, pero sobre todo nuevas lógicas, para que el negocio sea más eficiente y permita nuevas oportunidades" (Rocasalvatella, 2016)[2].

Una definición de la empresa consultora Flat 101 Digital Business[3], centrada en negocios digitales, declara que la Transformación Digital es un proceso necesario en todas las empresas actuales si quieren competir en un mercado cada vez más globalizado, interconectado y digital. Asimismo, considera que "«la Transformación Digital implica un cambio en la manera de hacer las cosas en una empresa y en la forma en que ésta se relaciona con sus clientes y *stakeholders*, adaptando el entorno digital a lo largo de toda su estructura".

La Transformación Digital ha comenzado en muchas empresas y se está llevando a cabo en un proceso que será de vital importancia para el futuro de las mismas; pero no es sólo tecnología la Transformación Digital, ya que, si bien el despliegue de las nuevas tecnologías es indispensable, son sólo un ingrediente

necesario, pero no el sujeto ni el objeto del cambio Delgado [4] (2016) considera que: "se está produciendo un cambio a una nueva sociedad que denomina ***hipersociedad*** ya que se está digitalizando la sociedad, la cultura, la política, los trabajadores y los consumidores". La Transformación Digital no sólo es tecnológica sino social y económica. Los cambios que se están produciendo son generacionales y, por consiguiente, inevitables (la generación de los "*millennials*" prácticamente viven en una Transformación Digital que han hecho suya en su manera de vivir). Está en marcha una revolución social para la digitalización de la sociedad, como en su momento hablamos de **cibersociedad** a finales de la década de los noventa, con el advenimiento de la nueva sociedad impregnada del ciberespacio y del nacimiento de la Web.

La Transformación Digital es un proceso necesario en todas las empresas actuales si quieren competir en un mercado cada vez más globalizado, interconectado, digital y omnicanal (acceso a Internet y a las organizaciones por múltiples canales).

Una última definición de Transformación Digital, que también aporta grandes ideas prácticas, es la de Lombardero (2015)[5], quien considera que "la Transformación Digital se basa fundamentalmente en la hibridación de productos físicos y servicios virtuales o en la conexión de productos, servicios, cosas y personas de forma inteligente, de manera que modifica la naturaleza de los productos y servicios". La convergencia de los productos físicos y lo virtual es uno de los objetivos fundamentales de la Transformación Digital, que implica transformar la cadena de valor para transformar, a su vez, a la empresa.

3.2.1 LOS MODELOS DE NEGOCIO EN LA TRANSFORMACIÓN DIGITAL

La Transformación Digital no se trata sólo de tecnología sino de transformar los modelos de negocios y, a su vez, ha de ser la base de la estrategia corporativa. La Transformación Digital requerirá nuevas destrezas y un desplazamiento en inversiones TIC.

La consultora IDC define la Transformación Digital como "el proceso continuo por el que las empresas se adaptan para conducir y controlar los cambios disruptivos en sus clientes y mercados (ecosistema externo) mediante la potenciación (*leveraging*) de competencias digitales para crear nuevos **modelos de negocio**, productos y servicios". La TD facilita a las empresas combinar la perfección digital, el negocio físico y la **experiencia de cliente** mientras mejora la eficiencia operativa y el desempeño organizacional.

Una fórmula muy utilizada por los expertos en DX es:
TD = Experiencia de cliente = Modelo de negocio

3.3 TECNOLOGÍAS FACILITADORAS DE LA TRANSFORMACIÓN DIGITAL

Las tecnologías facilitadoras de la Transformación Digital que impactan en los sistemas de Inteligencia de Negocios están ayudando a transformar digitalmente a las empresas. Los pilares fundamentales son: **Computación en la nube** (*Cloud Computing*), ***Big Data,*** **Internet de las cosas** –y sus variantes *Internet de todas las cosas* e *Internet industrial de las cosas*–, **Ciudades Inteligentes** (*Smart Cities*) y, como espina dorsal de todo el sistema, la **Ciberseguridad** y el soporte de la **Inteligencia Artificial** y la **Robótica**. Naturalmente, no todas estas tecnologías impactarán por igual a todas las empresas, ya que ello dependerá del tamaño de las mismas y del sector al que pertenezcan, así como de sus líneas de negocio fundamentales.

Es necesario que las empresas tradicionales de servicios e industriales de todos los sectores se transformen digitalmente con el objetivo de competir. El único límite es la capacidad de sus directivos para la innovación y gestión de nuevos modelos de negocio. De cualquier forma, es necesario tener presente que existirán empresas cuyas líneas de negocio serán las tradicionales y la digitalización sólo les afectará en cuanto a la gestión y productividad comecial.

3.3.1 TENDENCIAS DE TRANSFORMACIÓN DIGITAL

Al igual que sucede con las tendencias tecnológicas generalistas, también la Transformación Digital presenta sus tendencias para cada año. En este apartado vamos a recoger tres tendencias de Transformación Digital, dos publicadas en la revista *Forbes* a finales de 2015 (tendencias 2016) y otra a mediados de 2016 (tendencias 2017), y una propuesta de Transformación Digital de Telefónica, realizada en octubre de 2016. Con el estudio de estas tres informaciones solventes podremos hacer una síntesis para considerar nuestra estrategia en la Transformación Digital de la empresa y de sus empleados.

3.3.2 PREDICCIONES SOBRE EL FUTURO DE LA TRANSFORMACIÓN DIGITAL: TENDENCIAS 2016 (FORBES)

La prestigiosa revista *Forbes* –cuyas ediciones en inglés y en español llegan a numerosos países, entre ellos México y España– publicó el 6 de diciembre de 2016 un artículo de Gil Press[6], un reconocido experto en asuntos tecnológicos, en el que hizo seis predicciones sobre el futuro de la Transformación Digital, y donde proponía tres tendencias estratégicas y tres tecnologías que tendrían, según él, gran impacto en la Transformación Digital a lo largo de 2016.

Gil Press estudió tres informes de tres de las grandes consultoras multinacionales:

1. IDC predijo la emergencia de la Transformación Digital, a la cual llama **Economía DX**.
2. Gartner hace énfasis en la irrupción del negocio de los algoritmos o economía de los algoritmos (*algorithmic business)* y la economía programable *(programmable economy*).
3. Forrester, la tercera consultora, dibuja una hoja de ruta para empresas que proporcionen servicios a clientes y consumidores digitales.

Basado en estas predicciones, Press propone tres tendencias estratégicas y tres conjuntos de tecnologías –**Analítica de *Big Data*, Internet de las cosas (IoT) e Inteligencia Artificial**–, que impulsarán la Transformación Digital como resultado de la actual década.

Estrategias

- La Transformación Digital se convertirá en el principal impulso estratégico para la mayoría de los CEO.
- Las iniciativas de Transformación Digital se consolidarán y se crearán estrategias de negocios con una visión muy amplia durante 2016.
- La Transformación Digital requerirá nuevas habilidades y competencias con un desplazamiento hacia inversiones en TI.

Tecnologías

- *Big Data Analytics* será el soporte fundamental de la Transformación Digital.
- **Internet de las cosas** será un catalizador en la expansión de la Transformación Digital a todos los campos de la economía.
- La **Inteligencia Artificial (AI)** impulsará nuevos flujos de ingresos de Transformación Digital.

3.3.3 LAS 10 TENDENCIAS MÁS IMPORTANTES PARA LA TRANSFORMACIÓN DIGITAL EN 2017 *(FORBES)*

Daniel Newman[7], publicó en *Forbes* las siguientes tendencias para 2017:

1. **La adaptabilidad es más vital para el éxito que nunca.** Cita el caso del éxito de la computación en la nube a lo largo de 2016. Como ya hemos visto al comentar las características básicas de la Transformación Digital, Newman también considera que la Transformación Digital no se trata sólo de

tecnología, sino de reunir el poder de la tecnología con una cultura que abrace el cambio que puede conducir a la organización.

2. **La importancia del crecimiento de la experiencia de usuario.** La **experiencia del cliente** (incluyendo empleados) es el último objetivo de cualquier Transformación Digital.

3. **La innovación debe suceder rápidamente.** La innovación proactiva es una de las mejores maneras de mantenerse competitivo en un mercado en evolución.

4. **Abrazar a la fuerza de trabajo remota.** Los jóvenes profesionales prefieren la flexibilidad a la compensación. La tecnología móvil (celular) y la proliferación de la banda ancha permiten a las empresas conectarse con (y retener) los mejores talentos en cualquier parte del mundo.

5. **El advenimiento de la Realidad Aumentada (RA) y la Realidad Virtual (RV).** El éxito del *boom* de la aplicación Pokémon GO AR lo demuestra.

6. **Las interfaces de programas de aplicaciones (API).** Las API son un arma secreta para abrazar la Transformación Digital. eBay y PayPal son dos compañías que hacen uso muy significativo de estas tecnologías.

7. ***Big Data and Analytics*.** En la era de la Transformación Digital todo gira en torno a los datos que permiten medir todo o casi todo.

8. **La Transformación Digital está conducida por la Internet de las cosas.** Con unas estimaciones de 50 mil millones de sensores (cosas) para 2020 y más de 200 mil millones para 2030, el IoT no sólo será transformador sino disruptivo de los negocios.

9. **Las máquinas inteligentes y la Inteligencia Artificial (IA) están despegando en gran medida.** Nuestras relaciones con la tecnología continúan evolucionando. Pronto las máquinas podrán aprender y adaptarse a sus entornos.

10. **Destrucción de silos.** El rol de los CIO ha cambiado radicalmente en los años pasados. Con el advenimiento de los nuevos roles del CDO (*Chief Digital Officer*) y el CCO (*Chief Customer Officer*) estamos viendo la importancia de la Transformación Digital, no sólo en la tecnología de una compañía, sino en toda la organización completa. Las soluciones tradicionales son más multidimensionales y la tecnología no se puede utilizar como una muleta. Un enfoque en romper los silos, dará a la innovación más espacio para florecer y la colaboración se hará más fácil.

CASO DE ESTUDIO: LA TRANSFORMACIÓN DIGITAL EN LA VISIÓN DE TELEFÓNICA

Para impulsar en forma global la Transformación Digital de las grandes empresas y ayudarlas a adaptarse al nuevo contexto de mercado, Telefónica ha desarrollado ***Living Cloud*** (noviembre 2016)[8], una propuesta basada en su propia experiencia de transformación como compañía, y que se apoya en los siguientes pilares:

1. **La conectividad** que, lejos de ser una *commodity*, **es la clave de la Transformación Digital**. Su evolución aporta los tiempos de respuesta y los anchos de banda que permiten consumir los nuevos servicios digitales con una **experiencia de usuario excelente** (en este punto la latencia de la red es fundamental). Telefónica apuesta por una nueva generación de redes capaces de aportar la agilidad, seguridad y flexibilidad que reclaman los procesos de negocio.

2. **Una estrategia de servicios *cloud* vertebrada sobre un** modelo híbrido junto con la nube pública y un modelo de gestión y facturación único. Se trata de llevar todos los servicios a la nube: aquellos que habilitan las capacidades de comunicación, colaboración y productividad de las personas, así como los que permiten ofrecer servicios de infraestructura, plataformas e incluso *software* como servicio, y también los escritorios.

3. **El *digital workplace*:** el puesto de trabajo del empleado evoluciona y se hace más móvil, personalizado, virtualizado. Más allá del equipamiento hablamos del *software* que está en la nube, de las herramientas de seguridad y de gestión, en definitiva, de todo aquello que aporta a las empresas flexibilidad y eficiencia.

4. **La *omnicanalidad* y el *marketing* digital** permiten unir la vivencia física y *on-line* del usuario para ofrecerle una experiencia única, diferencial y personalizada.

5. ***Big Data*** e **IoT** son elementos diferenciales para hacer más competitivas a las empresas. Telefónica facilita *Big Data* desde tres vertientes diferentes:

   - Las plataformas en *cloud* ofrecen a las empresas la infraestructura necesaria para acometer sus proyectos de *Big Data* con modelos de pago por uso.
   - Los datos anonimizados que acumula como operador se cruzan con los de sus clientes para generar la mayor variedad de fuentes.
   - La experiencia en la elaboración de modelos predictivos y analítica de datos de sus *Data Scientists* (Synergic Partners).

6. La **seguridad** es clave para garantizar la confianza en el nuevo entorno digital. En los proyectos de Telefónica se incorpora desde el principio, y la propia red se convierte en un elemento activo de la seguridad, que va adquiriendo conocimiento de los ataques para bloquearlos.

3.4 LA EMPRESA DIGITAL

La Transformación Digital de la sociedad está conduciendo a la digitalización de las empresas. Prácticamente, la totalidad de las empresas, grandes, medianas y pequeñas, están en mayor o menor proporción digitalizadas: utilizan muchas tecnologías habilitadoras de la cuarta revolución industrial. Sin embargo, utilizar correo electrónico, disponer de página web, ser usuaria de una o varias redes sociales e, incluso, utilizar herramientas de *software* de gestión empresarial (como CRM, ERP o SCM) no significa que la empresa pueda ser considerada una empresa digital.

Delgado (2016) señala que una empresa digital es "la que usa intensamente las TIC para competir", y continúa: "Una empresa digital es aquella que ha realizado un esfuerzo consciente y sistemático para, gracias a las tecnologías de la información, ser más ágil, conocer y tener una mejor relación con sus clientes, reducir sus costes mediante la automatización extrema de sus procesos, incorporar tecnología a sus productos o servicios o facilitar la colaboración digital de sus empleados". Existe una gran unanimidad en los directivos, consultores y expertos en que la digitalización que entraña la Transformación Digital hoy en día es una necesidad para competir en el actual mundo digital, y máxime en los comienzos de la cuarta revolución industrial.

3.4.1 LOS DOCE DOMINIOS DE DIGITALIZACIÓN DE LA EMPRESA

La digitalización de la empresa se asocia tradicionalmente al *marketing* digital, comercio electrónico o digital, redes sociales, etcétera; sin embargo, en la actualidad la empresa digital implica la actuación en diversos ámbitos y sectores. Delgado (2016; 26-30)[9] define en su obra un modelo para realizar la digitalización de una empresa que se compone de doce ámbitos o dominios en los que la digitalización impacta a las empresas, y que son los siguientes:

- Marketing digital.
- Comercio digital.
- Redes sociales.
- Analíticas.
- Procesos digitales.

- Colaboración digital.
- Innovación digital.
- Habilidades digitales.
- Conocimiento digital.
- Roles digitales.
- Plataforma digital.
- Puesto de trabajo digital.

CASO DE ESTUDIO: "LAS TECNOLOGÍAS MÁS RELEVANTES Y DISRUPTIVAS DE LA ECONOMÍA DIGITAL" (ICEMD DE ESIC)

El ICEMD de la ESIC Business & Marketing School, publicó en el segundo trimestre de 2017 el estudio *Digital Technology for Business*[10], un informe exhaustivo sobre las tecnologías que ICEMD considera más relevantes y disruptivas en 2017, de impacto en la economía digital. Las conclusiones más significativas son:

- **Tecnologías más relevantes en la actualidad** son *Big Data*, *Cloud Computing* y Virtualización, Ciberseguridad, lenguajes de programación relacionados con *Big Data* (Hadoop, Cassandra, Hbase, Mongodb, Python, Hive, Spark, Impala, R), tecnologías de API y microservicios, IoT, y Realidad Aumentada y Virtual.
- **Tecnologías más disruptivas y con mayor impacto en la sociedad,** donde este impacto es considerado desde diferentes puntos de vista: impacto en la forma de gestionar la información del cliente, y en la que se comunican, relacionan, venden y atienden estos clientes; impacto en la optimización y eficiencia de procesos de toda la cadena de valor de la empresa; e impacto en la aparición y desarrollo de nuevos modelos de negocio y transformación de industrias, sectores y actividades. Así, estas tecnologías son la Inteligencia Artificial y cognitiva; *Deep Learning*, Chatbots y asistentes virtuales; reconocimiento de imagen, analíticas de texto, reconocimiento de voz; procesos de integración a través de *software* MOM; Industria 4.0; MOM: *Manufacturing Operations Management Software*; *Blockchain*; y tecnología *weareable*.
- **Tecnologías asociadas a tendencias que aparecerán o desarrollarán más su importancia en los próximos años**, tales como Robótica de servicios, vehículo autónomo, Inteligencia Artificial, Realidad Virtual/Ampliada[11], *Blockchain,* impresión 3D, nanotecnología, biotecnología (ADN, medicina preventiva, *cyborgs*), todo lo que tenga que ver con tecnologías que comprendan el comportamiento de los usuarios a través de huellas digitales, desempeños de

creatividad (humana) para hacer funcionar la Inteligencia Artificial, conexión máquina/humano, convivencia de la logística física y la digital, y *Smart Cities*.

3.5 LA TRANSFORMACIÓN DIGITAL EN LA INDUSTRIA Y EN LA EMPRESA

La digitalización de las organizaciones requiere de la puesta en marcha de un proceso de Transformación Digital con las estrategias adecuadas. Existen numerosas propuestas para llevar a cabo el proceso de Transformación Digital, dependiendo de los sectores a los que pertenecen las organizaciones, así como el tamaño, el estado real de digitalización, líneas de negocio, etcétera. Tenemos como ejemplo dos modelos: a) ***España 4.0: La Transformación Digital de la empresa***, realizado por la consultora Ronald Berger para Siemens España, muy centrado en el proyecto de Industria 4.0, y, en consecuencia, en empresas con este perfil; b) un segundo informe que hemos extraído del ***I Estudio de Transformación Digital de la Empresa***, realizado por la empresa especializada Territorio Creativo en colaboración con The London School of Economic Enteprise (**LSEE**), y que presentaron a finales de noviembre de 2015, el cual está más orientado a empresas de cualquier tamaño y realizado como una encuesta sobre el tema y consultas a numerosos directivos de todo tipo de empresas españolas, multinacionales y nacionales.

3.5.1 ESPAÑA 4.0: LA TRANSFORMACIÓN DIGITAL DE LAS EMPRESAS (UNA VISIÓN DESDE INDUSTRIA 4.0)

El estudio España 4.0 fue presentado en mayo de 2016 en Madrid por Rosa García, presidenta de Siemens España, Patrick Biecheler, director de Roland Berger y Juan Rosell, presidente de la CEOE (Central Española de Organizaciones Empresariales). La Transformación Digital, según Roland Berger/Siemens[12], es "la adaptación de las cadenas de valor de los distintos sectores de la economía a ese efecto disruptivo que comienza con el consumidor digital. La Transformación Digital es también la conexión integral de las distintas áreas de la economía y la manera en la que los *players* (actores) de cada sector se adaptarán a las nuevas condiciones que imperan en la economía digital". La Transformación Digital, según Siemens, podría implicar un incremento de 120.000 millones de euros en los sectores analizados para el año 2025.

Las grandes disciplinas consideradas en el estudio fueron:

- Economía digital.
- Empresa digital.

- Transformación Digital.
- Proceso de Transformación Digital.
- Herramientas para iniciar la Transformación Digital.

Y los procesos considerados fueron:

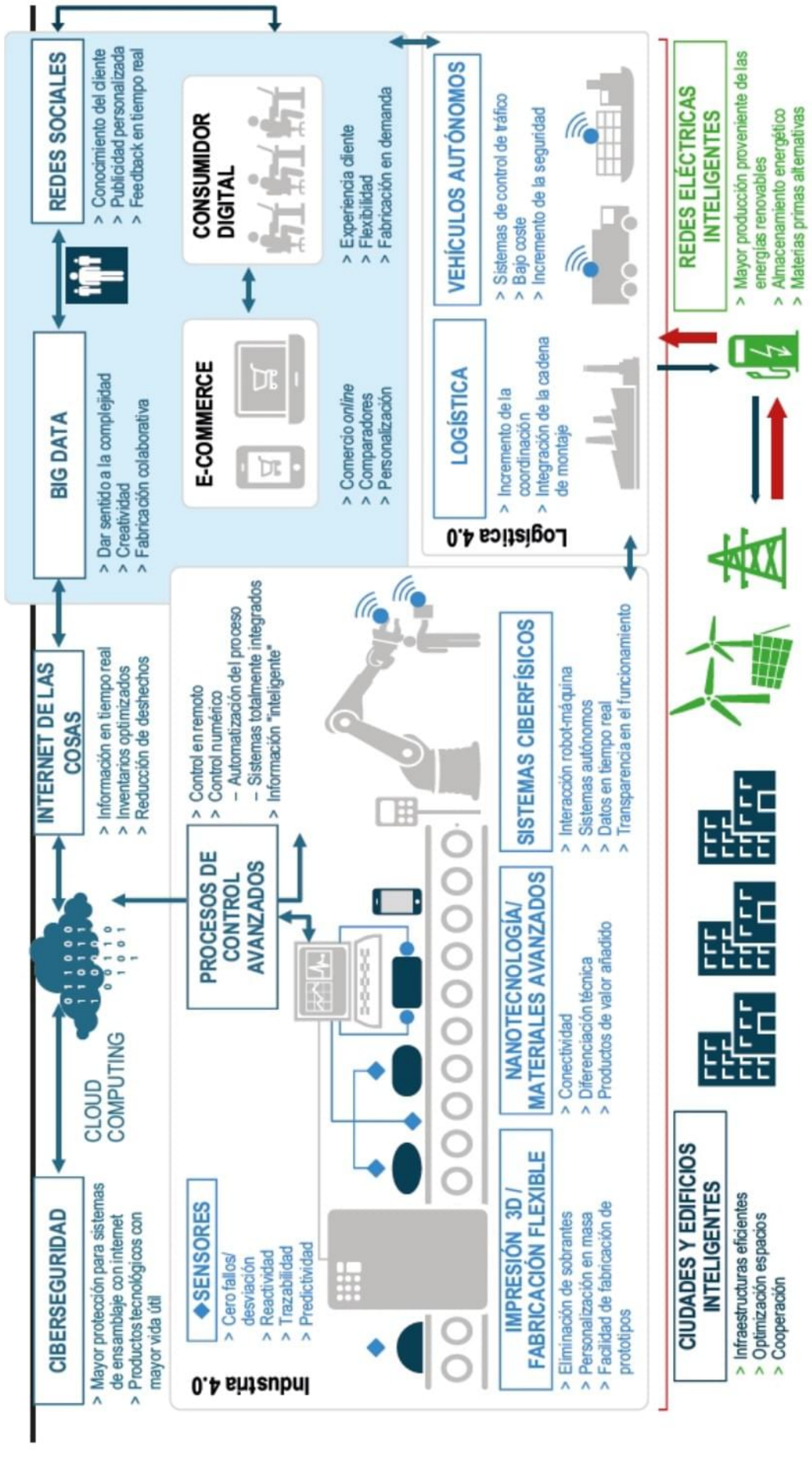
- Acceso a todo tipo de información digital.
- Automatización de procesos.
- Acceso digital al cliente.
- Conectividad en el ámbito interno.

Siemens fue una de las grandes empresas alemanas —junto con Bosch— que impulsaron la iniciativa Industria 4.0, por lo cual el ecosistema digital que proponen en el estudio se basa fundamentalmente en las tecnologías y estrategias propuestas en Industria 4.0.

El ecosistema digital propuesto se apoya en los pilares clásicos de Industria 4.0: Ciberseguridad, Internet de las cosas, *Cloud Computing* y *Big Data*, al que añade redes sociales con el soporte del comercio electrónico (*e-commerce*) y el rol del consumidor digital. El ecosistema se sustenta en Industria 4.0 y Logística 4.0 (disciplina lógicamente unida a la fabricación digital, con Siemens como líder mundial), cada una de las cuales se apoya, a su vez, en tecnologías y materiales de la industria y de la fabricación.

Industria 4.0

- Sensores.
- Procesos de control avanzado.
- Impresión 3D/Fabricación flexible.
- Nanotecnología/Materiales avanzados.
- Sistemas ciberfísicos.



Fuente: Roland Berger

Figura 3.1. Ejemplo de ecosistema digital

Fuente: Roland Berger/Siemens (Figura 1, p. 11).España 4.0: El reto de la transformación digital de la economía, 2016
https//w5.siemens.com/spain/web/es/estudio_digitalizacion_Espana40 Siemens.pdf

Logística 4.0

- Logística.
- Vehículos autónomos.

El ecosistema se sustenta en las **ciudades y edificios inteligentes** con la conectividad de las **redes eléctricas inteligentes**.

PALANCAS Y FACILITADORES DE LA TRANSFORMACIÓN DIGITAL

En el estudio se definieron cuatro palancas de la Transformación Digital (figura 3.2): información digital, automatización de los procesos, conectividad y acceso digital al cliente. Cada palanca ofrece unas propuestas y unas tecnologías facilitadoras.

El proceso de Transformación Digital de una empresa se concreta en los cuatro ejes o palancas antes citados.

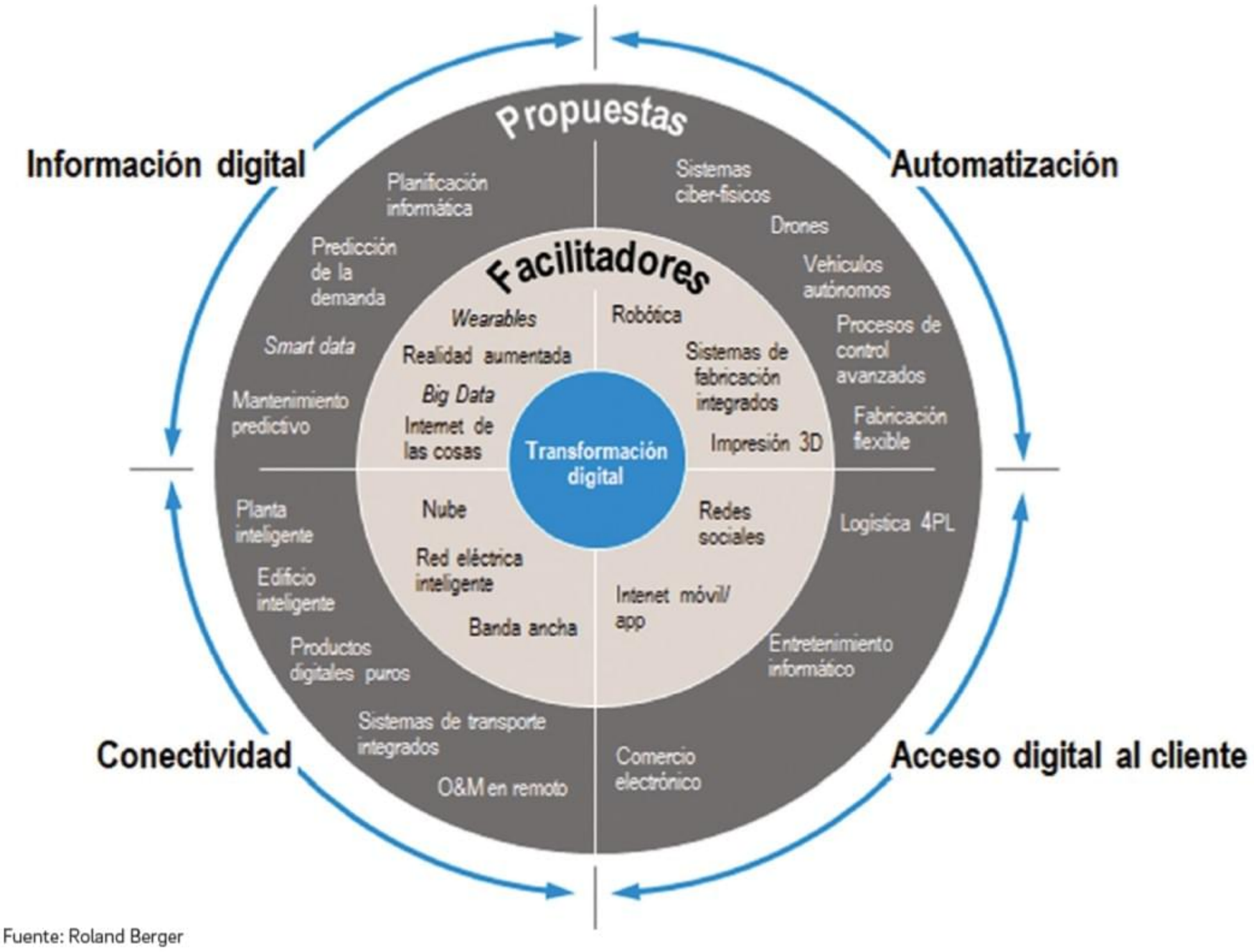
1. **Información digital**: "Disponibilidad de la información digital en la empresa en las etapas: captura, procesamiento y análisis de la información digital, que permite mejorar las predicciones y toma de decisiones" (Roland Berger/Siemens, 2016:11-13):. Este pilar **implica acceso a todo tipo de información** en forma instantánea, accesible y masiva, que condiciona todo el proceso de toma de decisiones, y que debe ser correctamente analizada para asegurar el éxito y adaptarse a las tendencias que condicionan el mercado. Se realiza una medición constante de resultados.

2. **Automatización de procesos**: "La combinación de la tecnología tradicional y la Inteligencia Artificial genera sistemas que pueden trabajar en forma autónoma". "La automatización de procesos en la organización de tareas y en los planes de negocio permite ganar en autonomía y flexibilidad en el funcionamiento interno y ofrecer un servicio y un producto flexible, adaptado y escalable a nuestro entorno".

3. **Conectividad**: "Interconexión de toda la cadena de valor vía móvil o banda ancha, permite sincronizar cadenas logísticas, acortar plazos de entrega y ciclos de innovación. Conectividad en el ámbito interno de la empresa en donde todos sus miembros y procesos están conectados, lo que permite agilizar procesos y acortar plazos de entrega".

4. **Acceso digital al cliente**: "Internet móvil y alta velocidad (redes 4G y las futuras 5G) permite a nuevos intermediarios dirigirse a consumidores a los que pueden ofrecer transparencia total y muchos servicios". Al cliente se llega con una vista en la *omnicanalidad* (cualquier canal de comunicación existente) en que se ofrece un servicio que va mucho más allá de la venta de un producto y que también exige transparencia en la relación con el consumidor. **La experiencia de cliente** se ha convertido en una necesidad

ineludible ya que si el cliente cambia su forma de comprar, las empresas tendrán que cambiar la forma de vender y, sobre todo, retener a los clientes".



Fuente: Roland Berger

Figura 3.2. Palancas propuestas y facilitadores de Transformación Digital. Fuente: Roland Berger/Siemens, p. 12

Como puede verse en la figura 3.2, en el caso de la automatización, las propuestas van desde los ya mencionados sistemas ciberfísicos, pasando por drones, vehículos autónomos, procesos de control avanzado y fabricación flexible. En síntesis, los principios de la fabricación inteligente, y en el caso de la conectividad, se consideran las plantas y edificios inteligentes, junto con los productos digitales puros y los sistemas de transporte integrados y la O&M (operación y mantenimiento) en remoto. Las otras dos propuestas se centran en la optimización de la información digital y en el acceso digital al cliente, punto fuerte en todos los estudios que hemos analizado de Transformación Digital.

Por último, las tecnologías facilitadoras coinciden con todas las ya descritas, aunque también hay que mencionar la red eléctrica inteligente y los sistemas de fabricación integrados, por el hecho diferenciador de la fabricación inteligente que tienen como soporte básico ambos sistemas.

El estudio de Siemens cuestiona **por qué todavía no se ha concretado esa Transformación Digital,** lo que plantea para el caso de la empresa española,

principalmente, una cuestión de mentalidad que se sustenta en la **resistencia y el miedo al cambio** de muchas organizaciones, aunque también señala los **costos para afrontar la digitalización** de las estructuras, así como la **capacitación de los empleados** o la falta de un **producto o servicio adaptado** a las nuevas necesidades. La solución pasa por una colaboración estrecha entre empresa, organizaciones y administraciones en un proceso que necesariamente, deben afrontar juntas.

3.5.2 I ESTUDIO DE TRANSFORMACIÓN DIGITAL DE LA EMPRESA (TERRITORIO CREATIVO & LSEE)

La empresa especializada en economía y *marketing* digital Territorio Creativo y The London School of Economic Enterprise (LSEE), con el apoyo de las empresas Brandwatch e Informática, publicaron a finales de noviembre de 2015 un estudio que realizaron acerca de la Transformación Digital[13] en la empresa española, apoyándose en estudios y encuestas de 900 directivos de medianas y grandes empresas, el cual recoge la visión de estos directivos sobre el impacto de la digitalización en sus organizaciones. Este estudio se realizó también en Colombia y fue presentado en julio de 2016 enfocado hacia la empresa colombiana.

En el estudio se define Transformación Digital como "un proceso de gestión que orienta la estrategia, la cultura, los procesos y las capacidades de una organización para canalizar la disrupción creada por el contexto digital".

El estudio destaca el impacto de las nuevas tecnologías en el negocio (todas ellas ya consideradas en Industria 4.0); las de mayor impacto son: *Business Analytics*, *Big Data*, movilidad, *omnicanalidad*, herramientas colaborativas, Internet de las cosas y tecnologías *wearables*, mientras que las de menor impacto son tecnologías disruptivas tales como Computación Cognitiva, Robótica, Realidad Virtual e Impresión 3D.

Las tendencias más destacadas en el estudio resultaron ser:

- Social Business Intelligence (Inteligencia de Negocio Social, con el enfoque en las redes sociales).
- Metodología Design Thinking (facilita el diseño de procesos y experiencias).
- Innovación abierta para significar la apertura de procesos de innovación a actores y plataformas tecnológicas.
- Analítica de Big Data.

Las recomendaciones a las empresas se agruparon en cuatro grandes categorías:

- Transformación Digital.

- Empresa digital.
- Negocio digital.
- Cliente digital.

Los motores de la Transformación Digital, según mostraron las estadísticas de los encuestados, fueron:

- Experiencia de cliente (44%).
- Mejoras operacionales (30%).
- Cambio del modelo de negocio (25%).

3.6 EL PROCESO DE TRANSFORMACIÓN DIGITAL

La Transformación Digital requiere de un proceso que puede resultar más o menos complejo en función de la empresa: su tamaño, su sector de influencia, sus líneas de negocio, etcétera. De igual forma, existen numerosos modelos de recomendación para la Transformación Digital de una empresa tradicional hacia un nuevo modelo de empresa digital. Hemos seleccionado dos métodos para recomendar a nuestros lectores: uno realizado por la prestigiosa revista ***The Economist*** y otro realizado por **Delgado** en su obra *Digitalízate*, referencia obligada en transformación digital, que consideramos muy práctica y eficiente, y cuyo contenido se basa en la experiencia de un experto profesional en digitalización de empresas y dirigido a directivos y emprendedores.

3.6.1 EL PROCESO DE TRANSFORMACIÓN DIGITAL DE LA EMPRESA (MODELO DELGADO)

Alberto Delgado[14], en su obra *Digitalízate*, propone un proceso de Transformación Digital para impulsar la digitalización en las organizaciones. Se apoya en las teorías de Kotter en su libro *Leading Change*, que señala los pasos necesarios para impulsar un cambio y los adapta a las transformaciones que implica la digitalización de la empresa.

El proceso de Delgado de Transformación Digital de la empresa consta de siete etapas o pasos, en cada uno de los cuales explica los conceptos fundamentales y las acciones que se han de realizar:

1. **Concienciación. Creación de urgencia: conciencia sobre la importancia de la digitalización.** Destaca la importancia de crear sentido de urgencia en el comité de dirección.

2. **Coalición. Creación de una coalición para la Transformación Digital.** Propone una gran coalición para el cambio. La mejor recomendación es la de coaligarse para crear un grupo de trabajo (*task force*) que lidere el proceso.
3. **Visión. Definición de la visión de la digitalización.** Para impulsar un cambio duradero y sostenible es necesario que exista una clara visión inspiradora de lo que la organización quiere conseguir con esta transformación.
4. **Comunicación. Comunicación de la visión al comité de dirección y al resto de la organización.** Comunicar la visión en dos pasos: en primer lugar, se trata de extender la coalición detrás de la visión al comité de dirección y, a continuación, al resto de la organización.
5. **Plan. Asignación de recursos y calendario: el plan estratégico digital.** Una vez que tenemos la visión comunicada a la organización, se trata de definir iniciativas que se desplegarán en proyectos concretos y que supondrán esfuerzos e inversiones.
6. **Piloto. Pilotear iniciativas de retorno rápido.** Identificar proyectos y actuaciones que tengan un retorno rápido.
7. **Implantación. Monitorización de la implantación del plan.** Monitorear la implantación de los proyectos que se incluyen en el plan para evaluar su retorno. Comunicación periódica al Comité de Dirección y también comunicación corporativa.

3.7 FÁBRICA INTELIGENTE: LA TRANSFORMACIÓN DIGITAL DE LA INDUSTRIA 4.0

La tendencia Industria 4.0, como venimos analizando, ha sido y será el soporte de la Transformación Digital de la industria, cualquiera que sea el sector al que pertenezca. La digitalización de las factorías convencionales y su conversión en factorías inteligentes es la espina dorsal del proceso de transformación. En este apartado hacemos una reflexión práctica sobre el cómo utilizar y sacar rendimiento para una correcta toma de decisiones de algunos de los pilares tecnológicos -citados anteriormente- y que nos sirva de introducción a las siguientes partes de esta obra.

La cuarta revolución industrial está llegando ya —lentamente en algunos casos y países— a la mayoría de las industrias y sectores, tales como el motor, la automoción, el metal, la aeronáutica, naval, ferrocarril, agricultura, ganadería o la pesca. Las innovaciones y tendencias tecnológicas (los pilares de la 4RI), que hemos visto anteriormente y que ampliaremos en la parte II y parte III del libro, son ya una realidad y constituyen herramientas de trabajo de muchas compañías.

La Transformación Digital ha llegado a la fábrica inteligente del futuro; la mayoría de las innovaciones tecnológicas —incluidas las más disruptivas— hoy son herramientas de trabajo y constituyen el armazón de la fábrica inteligente 4.0. La transformación que se está produciendo en todos los sectores industriales es posible gracias al uso de las tecnologías que, si bien no todas son nuevas, muchas veces se habían utilizado de forma aislada y no de forma integrada, como comienza a ser un hecho en la industria 4.0.

La **Transformación Digital en la industria** compete hoy a todos los sectores industriales que están sometidos a la presión de adaptarse a los cambios tecnológicos y sociales, con el objetivo de ser competitivos, y no sólo a las grandes multinacionales como Google, Facebook, Microsoft o IBM.

El **uso de sensores** en las máquinas y el análisis en tiempo real de los datos que generan permiten conocer los patrones habituales de funcionamiento y, a continuación, detectar cambios que señalan si una máquina requiere una reparación o predecir cuándo una máquina o una pieza va a fallar, es decir, planificar el mantenimiento.

Mediante **la simulación**, los fabricantes pueden usar datos reales de productos en el mercado y luego ejecutar simulaciones basadas en lo que sucedería si cambia un componente o un aspecto de diseño del producto o reproducciones del mundo real a un mundo virtual, para realizar pruebas y optimizar la programación y desarrollo de las máquinas. Las técnicas de IA y *Big Data* permitirán deducir, mediante las simulaciones, las reacciones de los clientes ante estos cambios.

La simulación, el uso de sensores en las máquinas y el análisis en tiempo real de los datos son hoy la base de lo que se denominarán los **gemelos digitales** (*digital twins*). Si bien era un concepto antiguo, a finales del año 2016, merced a la publicidad generada por Gartner con sus famosos informes de tendencias tecnológicas y la propaganda proporcionada por los grandes resultados en el sector industrial de la feria HANNOVER MESSE, fueron puestos de moda como un elemento clave en la composición de la fábrica inteligente.

Todos los sectores industriales producen grandes cantidades de datos que se pueden utilizar para mejorar los procesos productivos y la venta de productos y servicios. Esta situación exigirá el conocimiento de técnicas de análisis de Big Data. El aprovechamiento de toda la información que se genera en el proceso de fabricación, expedición y venta de un producto redundará en su aplicación y en el aumento de su rendimiento, optimización y valor añadido.

Las fábricas de automóviles, por ejemplo Citroën en la planta de Vigo en España, fue la primera de su grupo industrial en adoptar y probar los robots colaborativos **Cobots**, así como la impresión 3D en procesos industriales de la fabricación de automóviles. Una impresora 3D es capaz de fabricar piezas en tres dimensiones, lo que conlleva reducciones de costes, almacenaje y logística.

La sincronización de los flujos logísticos y el **uso de algoritmos** para optimizar las cadenas de suministro es otro gran avance. La automoción, el metal, la aeronáutica y la naval están impulsando el uso de simuladores para simplificar y reducir los costes de diseño y las pruebas de un nuevo producto, desde un componente para un automóvil o un avión a un buque.

La transformación digital de la industria se sustenta también en:

1. La nube (*cloud*) como centro de almacenamiento y seguridad de datos industriales que proporciona, además, un escalado eficiente de los recursos.
2. La ciberseguridad que garantiza la protección de los sistemas industriales y las infraestructuras críticas.

La feria de HANNOVER MESSE 2017 ha propuesto un modelo de fábrica inteligente 4.0. Los ingredientes o componentes que constituyen la Factoría Inteligente 4.0 son: una mezcla de automatización innovadora, gemelos digitales, Cobots e Inteligencia Artificial. Los distintos métodos en que se pueden combinar estas tecnologías en el proceso de fabricación constituirán el futuro científico de la fabricación industrial.

3.8 ECONOMÍA COLABORATIVA

Una de las tendencias de mayor impacto que están trayendo las tecnologías de la cuarta revolución industrial y, en particular, los algoritmos de búsqueda de información y conocimiento, es la Economía Colaborativa. Datos estimados por la consultora PWC en 2016 prevén un mercado global potencial para la Economía Colaborativa de 570.000 millones de dólares para 2025.

La Economía Colaborativa es una de las grandes tendencias del milenio en la medida en que cambia la relación con los clientes, la manera de consumir y la de producir. Como modelo emergente que es, a medida que se expansiona, trae un gran número de controversias, incomprensiones de los modelos tradicionales, competencia desleal, grandes aceptaciones o grandes rechazos.

El informe *Conclusiones preliminares sobre los nuevos modelos de prestación de servicios y la Economía Colaborativa*, de la Comisión Nacional de los Mercados y la Competencia (CNMC)[15] de España plantea que la Economía Colaborativa no sólo modifica la relación con el cliente, el modo de producir y el de consumir, sino también cómo compartimos activos, bienes o servicios mediante intercambio monetario —o no— a través de la Red. La Economía Colaborativa proporciona nuevos negocios en Internet que interconectan clientes y proveedores para servicios muy diversos. Estimaciones fiables hablan de que la Economía Colaborativa facturará en 2025 unos 300.000 millones de euros, 10 veces más que en la actualidad.

Existen numerosas definiciones y tipos de Economía Colaborativa, aunque hay cierta unanimidad en algunos criterios básicos: "Economía Colaborativa es el sistema económico en el cual se comparten e intercambian bienes y servicios a través de plataformas digitales" (Casas 2017)[16]. Las plataformas colaborativas tienen su origen en las primeras plataformas **P2P** (*peer to peer*), nacidas en el año 1999, y, como referente, la primera y más popular, Napster.

Existen miles de plataformas colaborativas desplegadas por todo el mundo y dedicadas a la compartición de bienes y servicios en infinidad de sectores, aunque han sido los dedicados al transporte y al alojamiento los que han impulsado la popularidad de esta nueva disciplina y, en particular, las plataformas Uber, Cabify para transporte y Airbnb para el sector de apartamentos privados y turísticos. Son casi innumerables las plataformas o aplicaciones en los servicios de alquiler de pisos, servicios de automóviles con o sin conductor, barcos, aviones, gastronomía, libros, ropas, joyas, bicicletas, intercambios en especies, etcétera. Numerosas empresas investigan sobre el automóvil compartido, con y sin conductor, como Toyota, Volkswagen, Amazon. Igualmente existen plataformas de *crowdfunding, coworking*, conocimiento abierto, cocreación, etcétera.

3.8.1 ¿QUÉ ES ECONOMÍA COLABORATIVA?

Como se ha comentado anteriormente existen numerosas definiciones de Economía Colaborativa y, en consecuencia, muchas respuestas posibles. Es un proceso en formación y evolución y es difícil y complicado llegar a una definición clara. Hemos estudiado varias referencias e informes notables relacionados y publicados en 2016 y 2017, que pasaremos a describir.

Alonso y Miranda (2017: 45-66)[17], autores de un libro de referencia sobre la empresa colaborativa y expertos en Economía Colaborativa, basándose en sus trabajos de investigación y en su experiencia profesional, definen Economía Colaborativa como:

"Un modelo de relaciones y conjuntos de iniciativas basadas en redes horizontales con la participación de una comunidad y que tiene efectos en todos los ámbitos de la sociedad".

El nuevo modelo **CoCoMaFiGoSi** —creado por los autores— está en constante evolución, pero los citados autores engloban todas las iniciativas que pueden presentarse en seis bloques diferentes:

- **Consumo colaborativo**. Es el consumo que se lleva haciendo toda la vida, pero aprovechando las nuevas tecnologías y las redes P2P se consigue la conexión necesaria entre pares para que lleguen a buen fin. Citan cinco plataformas correspondientes a cinco sectores diferentes: Airbnb (alojamientos en todo el mundo), BlaBlaCar (transporte basado en la confianza, que conecta a

conductores y pasajeros), TaskRabbit (resolución de tareas que necesita un usuario), Etsy (mercado de todo el mundo que se conecta en línea y localmente para compra y venta de productos únicos) y Screenly (consumo de cultura mediante una plataforma basada en un algoritmo, que calcula el número mínimo de espectadores para ofrecer visionado de películas, documentales, etcétera).

- **Conocimiento abierto (*open knowledge*).** Abarca una serie de principios e iniciativas encaminadas al acceso libre a la información, la investigación y la producción de aprendizaje. Comprende actividades tales como publicaciones científicas, *software* libre, contenido educativo abierto (plataformas MOOC).
- ***Makers.*** Producción colaborativa mediante herramientas y aplicaciones (Raspberry Pi, Arduino, revista *Make*, Ma Makers).
- **Finanzas.** *Crowdfunding* (KickStarter, TransferWise), colaboración en el mundo de las divisas, nuevos modelos, criptomonedas basadas en *Blockchain* (ver apartado 3.10) tales como *Bitcoin* y *Ethereum*.
- **Gobierno abierto.** Las administraciones públicas deben ofrecer a los ciudadanos la información necesaria de la que disponen para conseguir su confianza y ofrecerle buenas soluciones a sus preocupaciones y necesidades (Holacracy, FixMyStreet).
- **Sistemas de intercambio de valores.** Los ciudadanos buscamos diferentes modos de intercambiar valor. (TIMERPUBLI, ADBdT).

3.8.2 SECTORES DE IMPACTO DE LA ECONOMÍA COLABORATIVA

Como ya se ha comentado anteriormente, los sectores de impacto de la Economía Colaborativa son muy numerosos y las plataformas colaborativas se contabilizan en miles de unidades, tanto a nivel global como local, y en los sectores más diversos. Una síntesis de los sectores más afectados y las plataformas más sobresalientes, extraídos de las referencias bibliográficas de consulta utilizadas, es la siguiente:

1. **Transporte público**

Este sector recoge el mayor número de plataformas populares y que han adquirido mayor notoriedad, lideradas por Uber.

- **Uber.** Servicio profesional de transporte. Ofrece diferentes servicios de conductor con mayor o menor especialización: UberX, Taxi, Black, SUV, LUX, Uber Pop (servicio muy polémico por el cumplimiento de requisitos de regulación, que depende del país donde se establezca).
- **BlaBlaCar.** Es una plataforma de compartición de gastos de transporte.

- **Cabify**. Empresa de servicios profesionales de transporte, tanto a particulares como a empresas —normalmente vehículos de alta gama (Mercedes, Audi, BMW) —, con chofer profesional.
- **EasyTaxi, MyTaxi, Hallo.** Servicios de contacto de intermediación entre profesionales y clientes.
- **Avancar.** Empresa de compartición de automóviles (*carsharing*) global.

2. Sector Inmobiliario
   - **Coworking.** Compartición de espacios de oficinas, salas de reuniones, etcétera (**CoworkingSpain, CoworkingON, Copas).**
   - Compartición de plazas de garajes. En España, **Parkfy, Wayra, AparcaYa, Aparcalia**.
3. Sector financiero
   - ***Crowdfunding.*** Aportación de recursos monetarios/no monetarios para el desarrollo de una iniciativa privada (plataformas **Kickstarter, Gotea, Verkami, Crowdfundre, Bihoop**).
   - **Divisas.** Transferencias, cambios de divisas, nuevas monedas físicas y virtuales, criptomonedas, ***Blockchain, Fintech***, envío de dinero (**Transferwise**).
4. Turismo y ocio
   - **AirBnb.** Referente mundial en alojamiento.
   - **BeMate.** Alojamiento colaborativo: viajes, hoteles, apartamentos.
   - **Hundredrroms.** Metabuscador global de apartamentos y casas de vacaciones.
   - **Glovo.** Aplicación especializada en reparto de comida a domicilio. Creada en Barcelona y con gran difusión en la actualidad.
5. Educación
   - **Blabelia**. Idiomas.
   - **Erasmus.** Página web que ayuda a los estudiantes de Erasmus.
   - **Tutellus.** Plataforma colaborativa de gran difusión en el mundo hispano.
   - **Plataformas MOOC** (*Massive Open Online Courses*): Coursera, edX, MiriadaX, Udacity.
   - **Sahring Academy.** Una academia de iniciativa privada con gran penetración a nivel mundial.
   - **Khan Academy**. Iniciativa privada sin fines de lucro, con gran número de ofertas y opciones.

3.8.3 REGULACIÓN NORMATIVA DE LA ECONOMÍA COLABORATIVA

La Economía Colaborativa ha traído nuevos modelos de negocio con grandes oportunidades para las empresas, los trabajadores y los gobiernos. Sin embargo, también ha traído consigo grandes retos de regulación en el marco regulatorio del funcionamiento de las nuevas empresas colaborativas, en el marco de los impuestos y en el marco del empleo, además del cumplimiento de las normativas legales de privacidad y protección de datos[18].

La situación ha dependido, y depende, de cada estado, por lo que la situación en Estados Unidos, México, Colombia, Argentina, Gran Bretaña, Francia o España ha sido, y es, diferente, y cada estado considera su regulación de acuerdo con sus normativas legales.

AGENDA EUROPEA PARA LA ECONOMÍA COLABORATIVA

Con el objetivo de dar a conocer su posicionamiento sobre la Economía Colaborativa, la Comisión Europea publicó el 2 de junio de 2016 un primer informe: *An European agenda for the collaborative economy*[19]. El informe comienza reconociendo que la Economía Colaborativa está creciendo rápidamente y que los nuevos modelos de negocio crean incertidumbre en los operadores tradicionales, nuevos proveedores de servicios y consumidores, pero a la vez los nuevos modelos de negocio suponen una contribución importante a la innovación, al empleo y al crecimiento de la Unión Europea, por lo cual alienta a su desarrollo de un modo responsable.

El informe se publicó a modo de una guía legal para aconsejar a los consumidores, los negocios y las autoridades públicas, con normas y directrices a los países miembros sobre cómo afrontar el reto de la Economía Colaborativa. Sin embargo, la Comisión Europea se posiciona claramente a favor de impulsar plataformas colaborativas (como Uber, BlaBlaCar o Airbnb). Según la UE, las prohibiciones de este tipo de actividades deben ser sólo una medida de último recurso. Una economía europea competitiva requiere innovación, por lo que se ha de crear un marco regulatorio que permita desarrollarse a los nuevos modelos de negocio, proteger a los consumidores y asegurar políticas impositivas y de empleo justo.

La Comisión Europea recomienda que, en vez de prohibir totalmente la innovación asociada a la Economía Colaborativa, se deberían llevar a cabo formas de asociación entre estas nuevas empresas y los gobiernos, a fin de lograr su regulación y permitir su funcionamiento. De esta forma, la oferta sería variada y mejor, a favor de los consumidores y la innovación en general.

La CE trata de afrontar en su estudio los tres frentes vitales de la economía: el marco regulatorio para permitir (o no) el funcionamiento de servicios de Economía

Colaborativa, el marco impositivo y el del empleo. El periódico *El Confidencial*[20], de España, realizó un amplio dosier sobre el informe de la Comisión Europea, cuya lectura recomendamos.

El 22 de diciembre de 2016, la Comisión Europea dio forma de proyecto al informe preliminar *sobre una la creación de una Agenda Europea para la Economía Colaborativa*. Sobre la base del informe preliminar de junio del mismo año, ya se propone un **proyecto de informe**[21]. En el proyecto se destaca la necesidad de considerar la Economía Colaborativa no sólo como un modelo empresarial, sino también como una nueva forma de integración de la economía y la sociedad, capaz de imbricar las relaciones económicas en el entramado social y crear nuevos modelos comunitarios.

3.9 EXPERIENCIA DE CLIENTE

Experiencia de cliente (**CX**, *Customer Experience*) es una de las tendencias en el sector de la economía digital que se ha convertido en el foco de atención de todas las marcas en los últimos años, especialmente en los sistemas de información, y, en particular, en CRM —gestión de relaciones con los clientes—. Algunos expertos consideran a la experiencia de cliente como la evolución natural del CRM y una herramienta de gran impacto en la Inteligencia de Negocios.

Básicamente, la experiencia de cliente es una ventaja competitiva para alcanzar la Transformación Digital de la empresa y es el conjunto de experiencias que tiene un cliente con una marca durante el tiempo en que haya una relación de servicio. Los expertos en *marketing* y en economía digital señalan que existe una diferencia grande entre atención/servicio al cliente y experiencia de cliente. CX es una visión muy amplia y es la suma de todas las interacciones que tiene un cliente con una empresa.

Existen numerosas definiciones de experiencia de cliente. Una sintesis de las que hemos consultado (receñadas en la bibliografía de la obra) podría ser la siguiente: **Experiencia de cliente es el conjunto de experiencias que tiene un cliente con una marca durante el tiempo que haya una relación de servicio. Mientras mejor sea la experiencia de cliente, mayor es la probabilidad de incrementar la retención de los clientes, generar nuevos *leads*, incrementar el tiquete promedio, etcétera. La medición de la experiencia de cliente requiere de varias métricas, como es el caso de "satisfacción del cliente positiva". Los indicadores clave de rendimiento, KPI, son el elemento central para medir cada punto de interacción entre una marca y el cliente que contribuye a la experiencia global del producto o servicio.**

La Asociación DEC para el Desarrollo de la Experiencia de Cliente (`asociaciondec.org`), de España, define la experiencia de cliente como un factor emocional: "Es el recuerdo que se genera en la mente del consumidor

como consecuencia de su relación con la marca"[22]. Sus características principales son:

- Recuerdo (es algo emocional, un recuerdo subjetivo).
- Relación del consumidor con la marca antes, durante y después de la compra.
- Asociada a la marca. La experiencia **es** la marca.

La experiencia de cliente también se suele definir como el producto de las percepciones de un cliente después de interactuar racional, física, emocional y/o psicológicamente con cualquier punto de una organización. Esta percepción afecta a los comportamientos del cliente y genera recuerdos que impulsan la lealtad al cliente y afectan al valor económico que genera una organización. La experiencia de cliente alcanza a todas las interacciones, desde la atracción y el interés que generó el *marketing*, el descubrimiento, la compra, el uso, el servicio al cliente y hasta la baja, sin importar el canal por el que suceda. Su alcance es integrador, uniendo el negocio interno a la visión al cliente. De ahí que la integración con las herramientas de Inteligencia de Negocios, como más adelante comentaremos, se considera vital.

Bruce Jones[23] *en HBR* define CX como la suma de todas las interacciones que tiene un cliente con una empresa. Jones considera que puede incluir todo, desde la sensibilización o conocimiento inicial de un cliente (*awareness*), el descubrimiento de una empresa, producto o servicio, la compra del producto o servicio, el uso e, incluso, el servicio posventa. En conjunto estos se suman a los momentos críticos de la verdad que crean la experiencia de cliente.

3.9.1 PROCESO DE EXPERIENCIA DE CLIENTE

El proceso de CX consiste en diversas interacciones o "puntos de contacto". El proceso implica *omnicanalidad*, tanto en dispositivos como en otros diferentes canales de comunicación, *marketing*, distribución, departamento, factores internos o externos. Cada una de las experiencias de un cliente está compuesta por un número de interacciones que se tienen con un consumidor.

El proceso de experiencia de cliente se suele asociar en la actualidad a la *omnicanalidad*. Los diferentes canales de comunicación (visita física, dispositivos móviles, visitas al sitio web, mensajes instantáneos "*chat*", mensajes de texto SMS, asistentes virtuales por voz (*chatbots*), correo electrónico, etcétera) han de configurarse de modo que en la práctica sean un único canal de comunicación. La gestión de experiencia de cliente requiere que todas las conversaciones tengan un único discurso coherente, ordenado y con un trato y exquisitez —como señalan los expertos en *marketing*-—que ha de ser igual en todos los canales.

Experiencia de cliente en Inteligencia de Negocios

La recopilación, proceso y análisis de los datos extraídos durante el recorrido de un cliente en la gestión empresarial requiere la integración de la Inteligencia de Negocios para mejorar la experiencia de cliente. Se hace necesario un conocimiento exhaustivo, automatizado y en continua actualización, para generar un ciclo de personalización constante. La recopilación de datos en cada punto de interacción del recorrido del cliente aporta una foto instantánea del rendimiento integral. La utilización de toda esta información es vital en el proceso de ingesta y posterior proceso y análisis de datos en el sistema de Inteligencia de Negocios. El análisis de los datos extraídos de la gestión de la experiencia de cliente facilitará la toma de decisiones, tales como las mejoras que se podrán implementar rápidamente y cuales requerirán de más tiempo.

Qmatic[24], una consultora especializada en experiencia de cliente y en Inteligencia de Negocios en sus objetivos estratégicos, plantea que:

> *"El primer aspecto al que debemos prestar atención es que cada punto de contacto con el cliente es una oportunidad para reunir información y, por lo tanto, para aplicar el BI. Desde el momento en el que conectan con tu empresa y organización, ya sea por teléfono, online o en una oficina o punto de atención físico, los clientes son una fuente de información sobre lo bien que el personal, los procesos y tu marca se están desempeñando y cómo es la percepción de ese desempeño por parte de los clientes. Lo que comenzó como una forma de obtener feedback posterior al servicio o post-venta se ha convertido en una fuente de conocimiento útil que puedes utilizar para simplificar, conectar, agilizar y personalizar las interacciones desde el momento en el que un potencial cliente te encuentra online".*

Las herramientas de Inteligencia de Negocio se han de integrar en las plataformas de experiencia de cliente y gestión de relaciones con los clientes. Alcaide y Hernández (2016)[25] hacen una revisión general de las tendencias en Inteligencia de Negocios que se han de considerar para la integración de la experiencia de cliente de la empresa. Las tendencias propuestas son:

- **Visión única del cliente.** El cliente es un todo: todo lo que sepamos de él tiene que estar en una ficha accesible a toda la organización.
- **Microsegmentación orientada a la *Customer Experience*.** La microsegmentación es entendida como agrupación de clientes con características similares y suficientemente distintos de otros.
- ***Data Mining* orientado a la visión de conjunto.** Proponen valorar al cliente, agrupándolo a otros clientes similares para personalizar al máximo la relación. En esencia, determinar perfiles de comportamiento de cliente.
- **El cliente en tiempo real.** La tecnología de *Big Data*, como ya se ha comentado exhaustivamente a lo largo de nuestra obra, ayudan al tratamiento del cliente en tiempo real.

- **Microsegmentación geográfica.** El *geomarketing*, apoyado en la geolocalización (información geográfica), ayudará a la interactuación con los clientes junto con las herramientas GIS (Sistema de Información Geográfica).
- **Microsegmentación sociodemográfica.**
- **Identificando *influencers*.** Se debe identificar a nuestros clientes influyentes, aquellos que desempeñan una labor activa en la comunidad y pueden ser el complemento a nuestras actividades comerciales. Alcaide y Hernández recomiendan el uso de aplicaciones gratuitas, como Netvizz y Gephi, para obtener información de comunicades y formas de relaciones entre ellas, fundamentalmente con Facebook.

3.10 *BLOCKCHAIN* (CADENA DE BLOQUES)

La tecnología *Blockchain* (cadena de bloques) es el soporte de las criptomonedas y, de modo muy especial, de la moneda *Bitcoin*. Su despertar como tecnología innovadora y disruptiva comenzó a finales de 2016 (aunque su nacimiento, en 2009, es anterior, incluso, al despegue de *Bitcoin*), y es en 2017 cuando su impacto comenzó a llegar no sólo a los medios de comunicación, sino, y sobre todo, a numerosos sectores industriales, económicos y sociales.

Es dificil encontrar una definición común a *Blockchain*, pero dado que sus primeras aplicaciones comerciales han llegado al sector financiero, vamos a optar, en primer lugar, por una definición dada en una publicación del banco español BBVA[26]: **"Blockchain es una contabilidad pública entre pares que se mantiene mediante una red de computadores y que no requiere ninguna autoridad central ni terceras partes que actuén como inermediarios".**

De acuerdo con la definición, el documento del BBVA considera que *Blockchain* consta de tres componentes fundamentales:

- Una transacción.
- Un registro de transacciones.
- Un sistema que verifica y almacena la transacción.

Los bloques se generan a través de un *software* de código abierto y registran la información sobre cuándo y en qué secuencia ha tenido lugar la transacción. Este "bloque" almacena cronológicamente información de todas las transacciones que tienen lugar en la cadena (de ahí el nombre de cadena de bloques o *Blockchain*).

Un *Blockchain*, o cadena de bloques, se conoce también como libro de contabilidad distribuido (*distributed ledger).* En esencia, es una base de datos distribuida que registra bloques de información y los enlaza mediante

apuntadores aleatorios, para facilitar la recuperación posterior de la información y verificar que no se han modificado.

Blockchain es una base de datos distribuida, formada por cadenas de bloques diseñadas para evitar su modificación una vez creadas. Es un protocolo o combinación de tecnologías que actúan conjuntamente:

- Red P2P.
- Criptografía.
- Sellado de tiempo (*time stamping*).

La combinación de estas tecnologías permite construir un registro de hechos digitales, operaciones o bloques de información, distribuido, compartido y sincronizado entre muchas computadoras, y cuyo contenido no puede deshacerse, modificarse o alterarse sin el consenso de todos los participantes de esa red. Esta combinación de tecnologías también hace a las redes más seguras, transparentes y confiables sin necesidad de intermediarios.

Los archivos o bloques de información no se guardan en una base de datos ni en un servidor centralizado, como ocurre en bases de datos tradicionales, sino que se reparten en las computadoras de todos los usuarios de la red que pertenecen a esa cadena de bloques. Una vez que se encapsula el dato en un bloque, éste es inmutable y es visible para todos los usuarios que están en la misma cadena, por lo que nadie puede cambiar esa información sin la intervención de todos los participantes de la cadena.

Las comunicaciones y transacciones entre computadoras se encriptan (cifran) e incluyen un sello de tiempo, en fracciones de segundo, para evitar que puedan hacerse dos operaciones iguales a la vez (como hacer dos pagos con el mismo dinero o que dos personas compren un mismo objeto). El sellado de tiempo detecta la primera transacción y la segunda se rechaza.

La cadena de bloques es una especie de registro de transacciones que, en lugar de guardarse en una base de datos única, se comparte con todos los usuarios de la red, para que todos ellos autentifiquen que esos datos son válidos. En la práctica, *Blockchain* es un registro contable distribuido y digital que ofrece un gran nivel de seguridad y confidencialidad a las transacciones financieras, Esta cadena de bloques está vinculada a las tecnologías de registro distribuido (*distributed ledger technologies*), muy diferente al funcionamiento de las tradicionales bases de datos, y garantizan la seguridad de los datos contenidos en la cadena de bloques.

Las propiedades fundamentales del funcionamiento de las cadenas de bloques son:

- *Sistema descentralizado* (no se requiere una entidad central de confianza).

- *Tecnología abierta* (cualquier usuario puede hacer uso del sistema con sólo descargarse el *software*; a partir del registro se pueden realizar transacciones con criptomonedas y acceder a los datos registrados en su cadena de bloques).
- *Sistema distribuido* (reglas claras y consenso sobre la validez de las transacciones y el estado de la cadena de bloques).
- *Segura* (verificación criptográfica).
- *Privacidad*, autentificación y rapidez.

Por todas estas propiedades, una de las características más significativas de la cadena de bloques es la reducción significativa de costes, al permitir interactuar y hacer transacciones sin intermediarios, por lo que se simplifican y se abaratan los procesos.

Blockchain permite realizar transacciones de cualquier tipo de forma fiable y segura, sin necesidad de que haya un intermediario, facilitando que un registro de datos inalterable y descentralizado en redes P2P (seudoanónimas). La cadena de bloques es una tecnología con un enorme potencial ya que simplifica procesos, reduce costes y elimina intermediarios.

3.10.1 EL ALGORITMO DE *BLOCKCHAIN*

Cuando se une un usuario a la red, el contenido no se sube a los servidores centralizados de una empresa (o a la nube), sino que el archivo se trocea y se reparte mediante un algoritmo (similar al caso de los *clústeres* de servidores de Hadoop de *Big Data*) en diversas computadoras y, cuando se desea descargar, se recupera desde dichas computadoras.

El algoritmo de *Blockchain* es un algoritmo de cifrado o encriptación que se desarrolló, específicamente, para el sistema *Bitcoin,* pero que se puede aplicar a cualquier otro caso donde se requiera establecer un consenso distribuido. Los algoritmos criptográficos permiten que una base de datos descentralizada y distribuida entre diferentes participantes ofrezca unos niveles muy elevados de seguridad y privacidad para sus usuarios.

Cada vez que se ejecuta una operación, los algoritmos de encriptación forman una cadena de bloques interconectados, que no se pueden manipular sin la intervención de los otros agentes.

3.10.2 APLICACIONES DE LAS TECNOLOGÍAS *BLOCKCHAIN*

La tecnología de cadenas de bloques tiene un potencial enorme, debido a las razones antes enunciadas de simplificación de procesos, reducción de costes,

eliminación de intermediarios y la gran seguridad que ofrece a las transferencias de datos. El sector financiero es el más rápido que se ha lanzado a investigar los posibles usos del *Blockchain*, pero los expertos y empresas especializadas en la tecnología consideran que será útil en todos los sectores y se acabará imponiendo en todos los negocios digitales. Existen en la actualidad un gran número de empresas que utilizan tecnologías de *Blockchain* además de las aplicaciones financieras (*fintech*) como en la banca, donde numerosos bancos de diferentes países están investigando en ellas, como es el caso de los bancos Santander, BBVA o Caixa Bank, en España o Bancolombia en Colombia. Además del sector financiero, donde su aplicación ha sido pionera, numerosas empresas ya utilizan la tecnología. Algunos sectores con casos de éxito son:

- Contratos inteligentes (*smart contracts*). Realización de contratos entre partes sin necesidad de un intermediario.
- Empresas de alquiler de automóviles para realizar el seguimiento de los vehículos.
- Seguimiento de joyas, obras de arte y diamantes para realizar el seguimiento de los dispositivos asegurados.
- Redes colaborativas de aseguradoras, bancas, compañías eléctricas, estaciones de servicio de combustibles, recarga de automóviles eléctricos y la automatización de los pagos.
- Colaboración con las empresas eléctricas para pagar la recarga de un automóvil eléctrico sin necesidad de emplear una tarjeta de crédito. En las empresas de automóviles eléctricos, como Car2Go (de la casa Daimler Benz), Emov (del grupo PSA, Citroën), etcétera, podrían tener una gran aplicación.
- Trazabilidad (seguimiento) de productos comercializados por agricultores, ganaderos, etcétera, para seguir el camino de frutas y hortalizas hasta la llegada al consumidor final.
- Trazabilidad en los procesos de fabricación. El sistema hace posible la trazabilidad total de un producto fabricado en una factoría industrial, desde el proceso industrial del producto hasta que sale de la línea de producción y se envía a un distribuidor y un consumidor final lo compra.
- Realización de micropagos para eliminar las comisiones.
- En los sistemas de voto electrónico para el recuento de votos en elecciones. Los registros enviados a través de *Blockchain* son inalterables, por eso aportan seguridad y confianza a cualquier intercambio y envío de votos de las mesas electorales.

3.10.3 *BLOCKCHAIN* EN INDUSTRIA 4.0

En el sector industrial, las tecnologías *Blockchain* tienen una gran aplicación en el despliegue de la Industria 4.0. La mayoría de las industrias (automotriz, de aviación, pesquera, farmacéutica, armamentística) ofrecen grandes oportunidades para las aplicaciones de tecnologías de cadenas de bloques.

Una de las grandes aplicaciones de la tecnología es la trazabilidad total de una pieza industrial. El sistema permite la **trazabilidad total de un producto**, desde que se fabrica hasta que lo compra un consumidor final. Grandes fabricantes, como Airbus, Daimler o Siemens, han comenzado a trabajar en la trazabilidad de las piezas a lo largo de todo su ciclo de vida. Así, en un proceso de fabricación en una planta industrial, se sigue el ciclo de vida de un producto mientras se está fabricando en el interior de la planta, pero las empresas industriales necesitan seguir la traza del producto desde que éste sale de la planta y se envía al distribuidor para su venta al consumidor final. Las empresas industriales pueden solicitar a los proveedores que registren en *Blockchain* el código de cada pieza y le añadan todos los datos asociados a la misma, así como a las modificaciones que tuvieran lugar durante todo el proceso de actualización o mantenimiento de la pieza.

3.11. *BLOCKCHAIN* EN INTELIGENCIA DE NEGOCIOS

La tecnología *blockchain* aporta grandes oportunidades a negocios de numerosos sectores, desde seguros, hasta salud o turismo, y los grandes beneficios se apoyan en propiedades tan solidas como la confianza y la autoridad, claves, también en los sistemas de inteligencia de negocios.

La integración de la tecnología de cadena de bloques en la “nube” (cloud computing) ayudará a las empresas a mantener la ventaja competitiva al proteger la valiosa información almacenada en ella y la naturaleza encriptada de blockchain se puede utilizar para proteger las redes de dispositivos interconectados de posibles interferencias externas. De igual forma, la integración de blockchain con la inteligencia artificial y el internet de las cosas conforman también un triángulo que mejora la experiencia del usuario en el entorno empresarial y, por consiguiente, los resultados de negocios.

La importancia de blockchain es que permite generar confianza sin la necesidad de un intermediario. Al ser un modelo descentralizado puede tener gran impacto en las diferentes etapas y componentes de un sistema de inteligencia de negocios. Esta tecnología aporta transparencia en todo el ciclo de vida, lo que conlleva ahorrar tiempo y costes al consumidor y a las corporaciones que lo utilizan y otorgan seguridad de un modo más disruptivo al no requerir la existencia de intermediarios. La posibilidad de hacer transacciones de forma

segura efectiva y fiable es otra de las características importantes de la cadena de bloques

Sin lugar a dudas, una de las propiedades más notables y más beneficiosas, especialmente en la cadena de suministros, es la trazabilidad, de gran impacto en la inteligencia de negocios. La trazabilidad de las operaciones que permite eliminar intermediarios en operaciones de negocio y en la cadena de suministros es otra característica importante de impacto en los negocios. Con la trazabilidad asegurada, las empresas de servicios y de suministro realizaran los contratos con clientes de forma directa y automática, la confirmación de la recepción de un pedido será suficiente para iniciar el proceso de facturación y pago al proveedor.

La tecnología de blockchain facilitará nuevos modelos de negocio descentralizados. Actúa como una plataforma descentralizada y distribuida donde se puede compartir valor sin intermediarios y donde el poder y donde el poder es de la comunidad que impulsa la confianza entre los usuarios. Es una alternativa viable desde un punto de vista operativo y económico sobre el que construir procesos automatizados y nuevos modelos de negocio.

La cadena de bloques posibilitará automatizar los procesos, transacciones y relaciones entre diferentes partes de la cadena de valor; almacenar todos los datos y archivos que se disponen para fomentar la trazabilidad y regularidad, confidencialidad de los datos y frecuencias de registro de estos para protección de los intereses de los usuarios.

Analítica de blockchain

Las cadenas de bloques son, esencialmente, bases de datos y eso implica que los datos que contienen deben requerir su análisis. A medida que las organizaciones y empresas continúan explotando la tecnología blockchain, nuevos tipos de datos van apareciendo lo que requiere nuevas técnicas de análisis de datos -principalmente de big data-. Blockchain aumenta la precisión de los procesos analíticos consiguiendo que los patrones de comportamiento del cliente sean mucho más precisos de lo que se había logrado hasta ahora. En esencia, blockchain añade una capa a la capa segura existente de análisis de big data, lo que aumenta la seguridad de los datos.

La analítica de blockchain se debe integrar en el análisis de datos y su impacto será en las siguientes etapas:

- Accesibilidad de herramientas.
- Intercambio de datos.
- Monetización de los datos.
- Blockchain de datos.

Empresas especializadas en analítica como SAS han comenzado a distribuir herramientas de analítica de *blockchain* integradas en sus herramientas clásicas de analítica. A medida que el uso de *blockchain* aumenta, también aumentarán las organizaciones que necesitaran acceder a analizar los datos a medida que éstos crecen en complejidad y volumen. El análisis de datos en los procesos de *blockchain* se ha convertido en una necesidad vital y esta situación está conduciendo a la nueva analítica de blockchain (*blockchain analytics*).

3.12 RESUMEN

La primera parte está centrada en el concepto de transformación digital. Se describen las tecnologías facilitadoras, la empresa digital y el proceso de transformación digital.

La **Transformación Digital** (***DX,*** *Digital Transformation*) está suponiendo un cambio profundo en la naturaleza de los productos y servicios de las empresas tradicionales. Margaret Rouse[27] define la Transformación Digital como "la reinvención de una organización a través de la utilización de la tecnología digital para mejorar la forma en que la organización se desempeña y sirve a quienes la constituyen. Digital se refiere al uso de las tecnologías que generan, almacenan y procesan datos".

¿Cuáles son las tecnologías facilitadoras de la Transformación Digital? Los pilares fundamentales son **Computación en la nube** (*Cloud Computing*), ***Big Data,*** **Internet de las cosas** —y sus variantes *Internet de todas las cosas* e *Internet industrial de las cosas*—, **Ciudades Inteligentes** (*Smart Cities*) y, como espina dorsal de todo el sistema, la **Ciberseguridad** y el soporte de la **Inteligencia Artificial** y la **Robótica.**

En la segunda parte se estudia la Economía Colaborativa, una de las tendencias de mayor impacto que están trayendo las tecnologías de la cuarta revolución industrial y, en particular, los algoritmos de búsqueda de información y conocimiento.

La definición de referencia que hemos utilizado (Alonso y Miranda, 2017) determina que la Economía Colaborativa es **"un modelo de relaciones y conjuntos de iniciativas basadas en redes horizontales con la participación de una comunidad y que tiene efectos en todos los ámbitos de la sociedad".**

El modelo de Economía Colaborativa se apoya esencialmente en los siguientes bloques temáticos:

- Consumo colaborativo.
- Conocimiento abierto (open knowledge): plataformas MOOC.
- Makers.

- Finanzas. Crowdfunding.
- Gobierno abierto.
- Sistemas de intercambio de valores

La Economía Colaborativa ha traído nuevos modelos de negocio con grandes oportunidades para las empresas, los trabajadores y los gobiernos, pero también grandes retos de regulación de normas (temas legales) en el marco del funcionamiento de las nuevas empresas colaborativas.

La tecnología *Blockchain* (cadena de bloques), es una de las innovaciones tecnológicas disruptivas que están potenciando el despliegue de la Industria 4.0. Si bien no estaba contemplada en las tecnologías facilitadoras de la Industria 4.0 y la cuarta revolución industrial, sus fundamentos técnicos ya estaban presentes, dado que soportan la criptomoneda digital *Bitcoin*, que sÍ lleva varios años en el mercado.

Numerosas fuentes y gurús de reconocido prestigio en el campo económico, industrial y social consideran que la tecnología *Blockchain* puede llegar a tener un impacto tal como lo tuvo Internet y la Web, a principios de los 90 (impacto que sigue vigente hoy en día).

La tecnología de cadena de bloques es una combinación de tecnologías de registro de datos, que actúan como una base de datos distribuida que se apoya, a su vez, en tres grandes tecnologías: redes P2P “igual a igual”, criptografía y sellado de tiempo (*time stamping*).

Las cadenas de bloques son muy seguras, transparentes y confiables, sin necesidad de intermediarios. Sus grandes ventajas en la transmisión de datos son la seguridad, privacidad, autentificación y rapidez.

En el sector industrial se convertirán en una de las tecnologías más disruptivas de la Industria 4.0 y la fabricación inteligente. Una de sus aplicaciones más notables es la posibilidad de realizar la trazabilidad total de un producto, desde que se fabrica hasta su consumidor final.

El *Blockchain* añade seguridad y eficiencia a un gran número de transferencias de información y está abriendo nuevos modelos de negocio, no sólo para el sector industrial sino en muchos otros sectores, tales como los negocios digitales, el sector de la salud, el sector farmacéutico, el sector agrícola y pesquero, etcétera.

NOTAS

[1] *Margaret Rouse.* www.searchdatacenter.techtarget.com/es/definicion/Transformacion-digital

[2] *Roca Salvatella.* www.rocasalvatella.com/acompanamos-la-transformacion-digital-de-los-negocios

[3] http://www.flat101.es/servicios/transformacion-digital/

[4] *Delgado, Alberto (2016) Digitalizate. Cómo digitalizar tu empresa.* Barcelona. Libros de cabecera.

[5] *Luis Lombardero. Trabajar en la era digital. Tecnología y competencias para la transformación digital. Madrid: LID Editorial, 2015, p. 135.*

[6] *Gil Press. «6 Predictions About The Future Of Digital Transformation». Forbes (edición USA), 6 de diciembre, 2015.* http://www.forbes.com/sites/gilpress/2015/12/06/6-predictions-about-the-future-of-digital-transformation/3/#40823e046279.

[7] *Daniel Newman, «Top 10 Trends For Digital Transformation In 2017», Forbes, 30 de agosto, 2016.*

http://www.forbes.com/sites/danielnewman/2016/08/30/top-10-trends-for-digital-transformation-in-2017/#251e13d36567

[8] *En abril de 2016 la operadora española Telefónica presentó la plataforma Living Cloud, su propuesta para transformar de forma global la transformación digital de las organizaciones y que tiene como núcleo la nube (cloud computing)* www.movistar.es/grandes-empresas/soluciones/livingcloud

[9] *Op.cit. pp. 26-30*

[10] http://www.icemd.com/digital-knowledge/digital-business/

http://www.icemd.com/digital-knowledge/infografias/tecnologias-disruptivas

[11] *En el año 2017 y con gran despliegue en 2018 ha comenzado a tomar fuerza una nueva tecnología denominada Realidad Mixta (o Realidad Fusionada, así la ha definido el fabricante de electrónica Intel), que como su nombre indica es una combinación o mezcla de ambas tecnologías,* BBC Mundo. bbc.com/mundo/noticias-42455209

[12] *Roland Berger/Siemens. España 4.0: El reto de la transformación digital de la economía, 2016.* https://w5.siemens.com/spain/web/es/estudiodigitalizacion/Documents/Estudio_Digitalizacion_Espana40_Siemens.pdf

[13] *Territorio Creativo y The London School of Economic Enterprise (LSEE), 2015. I Estudio de Transformación Digital de la Empresa.* www.territoriocreativo.es/tcresearch

[14] *Alberto Delgado, Digitalízate (pp. 35-42). Dedica todo el capítulo 3 a proponer su modelo de proceso de Transformación Digital de la empresa. Además de describir su modelo de proceso y sus siete pasos para llevar a efecto dicha Transformación Digital, Delgado proporciona un excelente conjunto de consejos prácticos de: "cómo preparo a mi empresa para ser digital".*

[15] www.cnmc.es

[16] *Joan B. Casas. ¿Cómo se ha de regular la economía colaborativa? Mundo Empresarial, 2ºT, 2017. Editorial Komunica Kit, monempresarial.com/es, 2017.*

[17] *Manuel Alonso y Adrián Miranda. La empresa colaborativa. La nueva revolución económica. Madrid: Pearson, 2017.*

[18] *Jaume Esteve y Manuel Ángel Méndez. Informe de la UE sobre economía colaborativa. 2 de junio, 2016. [en línea]* http://www.elconfidencial.com/tecnologia/2016-06-02/economia-colaborativa-uber-airbnb-cabify-taxi_1210018/

[19] http://europa.eu/rapid/press-release_IP-16-2001_en.htm?utm_campaign=SMEs+H2020&utm_source=twitterfeed&utm_medium=twitter

[20] *Jaume Esteve y Miguel Ángel Méndez. Informe de la UE sobre economía colaborativa, en El Economista, 2 de junio 2016.*
http://www.elconfidencial.com/tecnologia/2016-06-02/economia-colaborativa-uber-airbnb-cabify-taxi_1210018/

[21] *Nicola Danti (ponente): PROYECTO DE INFORME sobre una Agenda europea para la economía colaborativa. COM (2016)0356 – (2016/0000(INI)). [en línea]*
http://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/PR/2017/03-20/1113122ES.pdf.

[22] *Sofía Medem de la Torriente (coordinadora). ¿Qué es la experiencia de cliente? White Paper DEC#0*

[23] *Bruce Jones de Disney Institute. Know the difference between Customer Service and Customer Experience . Harvard Business Review (HBR). 1 march, 2016*
https://bit.ly/1LUvGfR citado por Nelson García enel blog: customertrigger.com

[24] http://www.qmatic.com/es-es/conoce-qmatic/sala-de-prensa/2016/julio/business-intelligence-experiencia-de-cliente/

[25] *Juan Carlos ALCAIDE y Raúl HERNÁNDEZ. Experiencia de cliente. Cómo diferenciarse en la era digital. Madrid: The Valley Digital Business School, 2016. Págs. 96-103.*

[26] *BBVA Innovation Center. Ebook. Tecnología Blockchain. Centro de Innovación BBVA, 2016.* http://www.centrodeinnovacionbbva.com/ebook/ebook-tecnologia-blockchain

[27] *Margaret Rouse.*
www.searchdatacenter.techtarget.com/es/definicion/Transformacion-digital

PARTE II
INFRAESTRUCTURAS Y ARQUITECTURA DE INTELIGENCIA DE NEGOCIOS

CAPÍTULO 4

ALMACENES DE DATOS: *DATA WAREHOUSE*, OLAP Y *DATA LAKE*

CONTENIDO

4.1 Introducción
4.2 Datos: gestión, gobierno, calidad e integridad
4.3 Administración de archivos
4.4 Bases de datos
4.5 *Data Warehouse*
4.6 *Data Mart*
4.7 Marco de trabajo (*framework*) de un sistema de almacenamiento de datos
4.8 Metadatos, calidad y gobierno de un *Data Warehouse*
4.9 Herramientas **ETL**
4.10 Desarrollo de un sistema de *Data Warehouse*
4.11 Enfoques de desarrollo (modelos) de un sistema de *Data Warehouse*
4.12 **OLAP** (Procesamiento analítico en línea)
4.13 *Data Lakes* (Lagos de Datos): Los nuevos depósitos de almacenamiento de datos
4.14 *Data Lake* versus *Data Warehouse*
4.15 Proveedores de soluciones de *Data Warehouse*
4.16 RESUMEN

OBJETIVOS

- Identificar y analizar los datos como activo estratégico de la empresa.
- Examinar y comprender la gestión de los datos como proceso de gran valor en la empresa.
- Conocer y comprender la importancia de la calidad y de la integridad de los datos.
- Examinar y analizar las características fundamentales de las bases de datos y de los almacenes de datos (*Data Warehouse* y *Data Mart*).
- Conocer e identificar los dos enfoques principales de un sistema de datos Warehouse: Inmon y Kiwball.
- Conocer y comprender las características fundamentales de un *Data Warehouse*, así como sus componentes principales.
- Conocer los nuevos depósitos, "lagos de datos" (*data lakes*).

4.1 INTRODUCCIÓN

Los departamentos de TI enfrentan enormes retos debido al crecimiento continuo de los datos procesados por las bases de datos y las aplicaciones empresariales, junto con el crecimiento exponencial de los datos sociales procedentes de los medios sociales (redes sociales, blogs, wikis, etcétera). La explosión de los datos, en realidad, es el resultado de la existencia de múltiples canales de comunicaciones que los transportan. Ahora bien, la infraestructura de datos se refiere a la estructura fundamental de un sistema de información y requiere su gestión eficaz a fin de mejorar el rendimiento y la productividad, tanto de las empresas como de sus gerentes, directivos, y trabajadores.

En estos tiempos, las empresas se enfrentan a una espectacular "revolución industrial de los datos", que ha acuñado el término Big Data (macrodatos o grandes volúmenes de datos), debido a su superabundancia en el mundo actual. Los Big Data están trayendo grandes oportunidades a los negocios, aunque también están creando numerosos problemas. Esta nueva infraestructura de datos requiere de nuevas herramientas y aplicaciones de software, que sobrepasan a las herramientas tradicionales de bases y almacenes de datos. En ese sentido, nos centraremos en la infraestructura y la gestión de los datos, y cómo se implementan en el planeamiento, diseño y construcción de negocios.

4.2 DATOS: GESTIÓN, GOBIERNO, CALIDAD E INTEGRIDAD

Los datos son uno de los activos estratégicos más importantes de las organizaciones. Son elementos discretos sin valor por sí solos, porque su valor reside en una gestión adecuada para convertirse en información y, luego, en conocimiento. Esto significa que es necesario desarrollar una capacidad para

extraer valor de la información y, por eso, hace falta tener modelos que permitan evaluar la realidad de los diferentes tipos de datos.

La información y el conocimiento son recursos organizacionales valiosos que pueden proporcionar una ventaja competitiva. En consecuencia, se tratará la primera parte del proceso de conversión de los datos en información y, a continuación, de la información en conocimiento para su conversión en toma de decisiones que conducirá a la sabiduría. Este proceso se conoce como pirámide del conocimiento: datos→información →conocimiento → sabiduría.

La gestión de los datos y del conocimiento son una pieza clave para las organizaciones y empresas modernas. Por esta razón, analizaremos en detalle la gestión de los datos y las herramientas más utilizadas: las bases de datos y los almacenes de datos (*Data Warehouses* y *Data Marts*).

Los datos se caracterizan por su formato (estructura de datos), su contenido (tipos de datos) y procedencia de origen (de dónde vienen). El formato donde se almacenan los datos es de dos tipos: estructurado y no estructurado. El formato estructurado determina la longitud del campo y la clase de caracteres; es el formato tradicional de los archivos y base de datos (direcciones, nombre, edad, ciudad). El formato no estructurado no contiene campos definidos ni en tipo ni en tamaño (mensajes de correo electrónico, audio, video, fotografía). Estadísticas fiables dan estimaciones del 80 al 85%, e incluso superior, para la información no estructurada que actualmente se maneja en las organizaciones.

Estos números manifiestan que los datos que una empresa recoge mayoritariamente de sus fuentes internas, de su entorno próximo, de los medios sociales (redes sociales, blogs) no serán datos perfectamente estructurados y normalizados (se pueden obtener de sus procesos internos y operacionales), sino todo lo contrario: no estructurados o desestructurados. La conversión de datos es información y conocimiento; ambos conforman una cadena de valor real.

La empresa realiza y ejecuta un conjunto de procesos operacionales que conforman la cadena de valor que, a su vez, proporciona una cadena de valor virtual directa, que son los datos. Sin embargo, en el interior de la empresa tiene lugar un conjunto de procesos no operacionales, no normalizados, que dependen más de activos intelectuales y que dificultan su transformación en el binomio información-conocimiento.

La gestión de los datos requiere en primer lugar, y como herramienta clásica, su almacenamiento en las bases de datos de las empresas, utilizando productos de software, tales como Oracle, Microsoft, IBM, etcétera. Después de dicho almacenamiento, los datos deben ser accesibles a los usuarios en formatos para permitir una fácil y rápida toma de decisiones.

Las empresas utilizan los datos de sus bases de datos y los datos históricos, creando almacenes de datos (*Data Warehouses*), que son herramientas muy valiosas para la toma de decisiones. En los siguientes apartados, examinaremos las ventajas e inconvenientes implicados en la gestión de datos y el enfoque de

las bases de datos que utilizan las organizaciones para resolver los problemas que puedan surgir y potenciar sus ventajas tradicionales.

Después que los datos se almacenan en las bases de datos de su organización, éstos deben quedar accesibles a los usuarios, de forma que ellos ayuden en la toma de decisiones. Las organizaciones cumplen estos objetivos desarrollando almacenes de datos (*Data Warehouses*). Por esta razón, es necesario familiarizarse con el concepto de almacenes de datos, dado que son herramientas imprescindibles en la gestión diaria de la empresa.

Se examinan las características fundamentales implicadas en la gestión de los datos y el enfoque de base de datos que las organizaciones utilizan para un uso eficiente de los mismos. Se analiza también cómo los sistemas de gestión de datos facilitan a las organizaciones el acceso y uso de los datos almacenados en los archivos y bases de datos. A continuación, se estudian los *Data Warehouses* y *Data Mart* y cómo se utilizan para la toma de decisiones.

4.2.1 GESTIÓN DE LOS DATOS

Las aplicaciones de TI requieren datos de alta calidad para su uso eficaz. ¿Qué se entiende por datos de calidad? Son los datos precisos, completos, consistentes, accesibles, relevantes, concisos y oportunos. Los datos de pobre calidad pueden producir tomas de decisiones pobres (ineficientes), servicios de baja calidad al cliente y diseño de productos inadecuados. Por estas razones, se requiere una eficaz gestión de los datos, es decir, una buena administración de datos, textos, documentos, contenidos de audio, video, fotografía, etcétera.

"El objetivo de la gestión de datos es proporcionar la infraestructura y herramientas para transformar los datos en bruto (rawdata) en información corporativa usable (utilizable) de la más alta calidad" (Laudon, 2014). Al igual que se gestionan los activos financieros, se deben gestionar los datos, ya que son activos informacionales de la organización.

La gestión de los datos es un proceso complejo que requiere:

- Comprensión de los datos (perfiles).
- Calidad de los datos y su mejora continua.
- Integración de los datos, combinando datos similares procedentes de fuentes diferentes.
- Aumento de los datos con la mejora continua de su valor.

Los directivos y restantes empleados de la empresa que deben tomar decisiones necesitan de un rápido acceso a los datos completos y correctos de la empresa para mejorar sus procesos de negocio y su desempeño profesional. Su toma de decisiones, con el consiguiente servicio a los clientes, depende, en gran medida, de la disponibilidad y calidad de los datos, es decir, de una base de

datos, un *Data Warehouse* o un *Data Mart*. Una base de datos organiza datos de la empresa en un formato (filas y columnas), de modo que resulte sencillo acceder a ellos, almacenarlos y recuperarlos. un *Data Warehouse* agrega datos de bases de datos transaccionales para que puedan ser analizados y puedan servir, por ejemplo, para identificar y examinar tendencias de los negocios y ayudar en la planificación y toma de decisiones.

La gestión de los datos es un enfoque estructurado para la captura, almacenamiento, procesamiento, integración, distribución, aseguramiento y archivado de los datos de un modo eficiente a lo largo de su ciclo de vida. El ciclo de vida de los datos se identifica con los modos en que viajan a través de una organización, desde su captura o creación hasta su uso operativo en soluciones (aplicaciones) controladas por los datos, como son aplicaciones de ERP, CRM, SCM, o de comercio y negocios electrónicos.



Figura 4.1. Ciclo de vida de los datos.

El ciclo de vida de los datos está muy relacionado con la gestión del conocimiento, ya que la transformación de los datos en conocimiento (las aplicaciones o soluciones que utilizan eficazmente la información) se realiza en varias etapas (figura 4.1). Comienza con la captura de colecciones de datos que proceden de distintas fuentes (datos internos, externos y personales o de usuarios finales), y su almacenamiento en bases de datos. A continuación, serán preprocesados para adaptarse al formato de un *Data Warehouse* o un *Data Mart*, donde se alojarán definitivamente. Luego, Los usuarios acceden allí y recogen los datos, una muestra, que necesitan para su análisis. El análisis de datos se realiza con técnicas de análisis y herramientas de Minería de Datos, que obtienen patrones para corregir la interpretación mediante resultados (visualización de datos, apoyo a la decisión y herramientas de gestión del conocimiento) y el uso de herramientas o soluciones (sistemas de información) como CRM, ERP, SCM o comercio electrónico.

Las dificultades de la gestión de los datos

El proceso de adquirir, mantener, actualizar y gestionar los datos se ha convertido en una tarea compleja, no sólo por su gran volumen, sino también por sus diferentes formatos, especialmente los datos no estructurados (aunque, por ahora, la gestión más eficiente y de calidad está representada por no estructurados). Como se procesan en diferentes etapas, lugares y momentos, con distintos dispositivos fijos (PC, servidores) y móviles (tabletas, teléfonos

inteligentes), la gestión de los datos en las organizaciones presentan grandes problemas y dificultades.

Esta dificultad de la gestión de los datos se produce por varias causas (Rainer et al., 2013):

1. La cantidad de datos aumenta exponencialmente con el tiempo. El asentamiento de los Big Data en las organizaciones y empresas es un caso significativo. Sin embargo, sólo una pequeña parte de estos grandes volúmenes de datos son relevantes para las diferentes aplicaciones y esos datos deben identificarse y encontrarse para poder ser útiles. Muchos datos históricos se deben mantener durante largos períodos, y además nuevos datos se añaden con gran rapidez. Actualmente, numerosas grandes empresas o almacenes pueden gestionar decenas o centenas de terabytes, incluso algunas de petabytes.

2. Los datos están dispersos a través de las organizaciones y de las innumerables bases de datos de la Web y son recolectados por muchas personas, utilizando diferentes métodos y dispositivos. Los datos se almacenan, frecuentemente, en numerosos servidores y lugares y en distintos sistemas de información, bases de datos, infraestructuras de datos, así como se construyen con diferentes lenguajes de programación y lenguajes humanos ordinarios.

3. Los datos se obtienen de múltiples y diferentes fuentes: fuentes internas (bases de datos corporativas, documentos institucionales), fuentes personales (experiencias, currículum, destrezas) y fuentes externas (informes, consultoras, auditoras, gobiernos, sitios web, blogs). Una fuente origen de grandes oportunidades, y también de grandes problemas, son las descargas de contenido de datos y páginas web. Los datos se descargan mediante los flujos sucesivos de clics de los usuarios, generando una huella o ratio digital que permite conocer las direcciones URL de las páginas web visitadas. Dichas huellas tienen las dos características positivas o negativas, según se utilicen o interpreten.

4. Los datos se degradan con el tiempo. Los clientes cambian de dirección de correo electrónico, los contenidos de las páginas web se actualizan o se quedan obsoletos, las empresas inician nuevas líneas de negocio, se desarrollan nuevos productos, las empresas se extienden por ciudades, países, etcétera, o se reduce su presencia de igual forma.

5. Los datos sociales dependen en gran medida de los soportes en que se almacenan y, de igual forma, los diferentes tipos de datos almacenados en sus soportes pueden impactar de maneras muy diferentes: por ejemplo, imaginemos tipos de datos como la temperatura, la humedad o la presión atmosférica, que son muy diferentes en su almacenamiento de análisis de datos financieros, económicos o de comercio electrónico.

6. La seguridad, la calidad y la integridad de los datos son críticos y pueden ser fácilmente atacados. Además, los requerimientos legales difieren entre países, así como las industrias que trabajan con los datos. En España, existen la Agencia de Protección de Datos o el INCIBE (Instituto Nacional de Ciber-seguridad de España) y el antiguo instituto de tecnologías de las comunicaciones (INTECO); asimismo, Europa cuenta con la Agencia Europea de Seguridad (ENISA). Son organizaciones que se dedican a analizar, explorar, proteger los datos de usuarios y organizaciones y empresas, así como elaborar normativas, velar por su aplicación o proporcionar documentación para facilitar la seguridad de los datos.
7. Los datos en una organización son redundantes y con frecuencia están desactualizados, creando un enorme problema de mantenimiento para los gerentes de datos.

Estas dificultades, y la necesidad de disponer de información precisa y oportuna, exige eficaces y efectivas soluciones de gestión de datos por parte de organizaciones y empresas.

4.2.2 GOBIERNO DE LOS DATOS

El gobierno de los datos es un enfoque para gestionar la información a través de una organización completa; implica un conjunto formal de procesos de negocios y prácticas para asegurar que los datos puedan ser manejados de un modo bien definido. Es decir, la organización sigue reglas no ambiguas para creación, recolección, manipulación y protección de la información. El objetivo es poner la información disponible, transparente y útil para las personas autorizadas a su acceso a lo largo de todo el proceso.

El Business Inteligence Institute (the BI Institute), de México, considera que el gobierno de datos es una de las mayores áreas de oportunidades de las organizaciones, y lo define como una serie de políticas, procedimientos y gente necesaria para establecer un marco de consistencias de datos dentro de la organización. Este gobierno de datos está presidido por algún alto ejecutivo, que tiene influencia sobre las áreas de negocio y las áreas de tecnología, y su finalidad es instrumentar un control efectivo sobre la información apoyado por toda la organización.

El citado BI Institute considera que todas las organizaciones se encuentran actualmente ocupadas en preservar y explotar uno de sus activos más valiosos: la información.

Considera también la necesidad de añadir al gobierno de datos un marco regulatorio nacional e internacional con una gestión eficiente. Los datos deben tener los siguientes atributos para garantizar un uso competente:

- El dato debe ser único.

- El dato debe ser correcto y exacto.
- El dato debe ser oportuno (estar a tiempo).
- El dato debe ser consistente.
- El dato debe ser completo.
- El dato debe estar protegido (confidencialidad, acceso, respaldo, etcétera).

Una estrategia para la implementación del gobierno de datos es la gestión de datos maestros (*Master Data Management*), que abarca todos los procesos y aplicaciones de una organización. Proporciona a las empresas la capacidad de almacenar, mantener, intercambiar y sincronizar los datos maestros, de modo que sean consistentes, precisos (correctos) y oportunos.

Los datos maestros son un conjunto de datos, tales como cliente, vendedor, empleado, producto, posición geográfica, que se extiende en todos los sistemas de información de la empresa. Es preciso diferenciar entre datos transaccionales y datos maestros. Los datos transaccionales son los datos que se generan y son capturados por los sistemas operacionales; describen las actividades o las transacciones de los negocios. Los datos maestros implican transacciones múltiples y se utilizan para categorizar, agregar y evaluar los datos transaccionales. Por ejemplo, un dato transaccional puede ser: Luis Macko ha comprado en un gran almacén un televisor LED Smart TV de 42 pulgadas, el 5 de septiembre de 2014. En este caso, los datos maestros son “vendedor”, “número de serie”, “precio de la compra”, “fecha de la compra”, etcétera. La aplicación de los valores específicos a un dato maestro representa una transacción.

4.2.3 CALIDAD DE INTEGRIDAD DE LOS DATOS

Los datos son el activo intangible más preciado de las organizaciones. La calidad de datos es una medida importante que las empresas han de tener presente para su uso en la planificación de estrategias y tomas de decisiones tácticas. La característica de calidad de datos es uno de los principales problemas de muchas organizaciones. Las cualidades del dato son la integridad, la calidad y la veracidad, que son indispensables para su correcta integración.

Es preciso centrarse en la integración de los datos teniendo presente su rol en la estrategia de la empresa, así como en la calidad de los propios datos, y han de ser componentes fundamentales de cualquier proceso.

Calidad de los datos

La calidad de los datos (*DQ, Data Quality*) es una propiedad muy importante, dado que determina la utilidad de los datos así como la calidad de las decisiones que se basan en ellos. La recolección de datos es un proceso complejo que puede crear problemas relativos a su calidad. Por consiguiente, con

independencia de cómo se recolectan los datos, necesitan ser validados para que se pueda confiar en ellos.

La calidad de los datos tiene las siguientes dimensiones: precisión, accesibilidad, relevancia, oportunidad y completitud (lo completo). Los datos que se encuentran, con frecuencia, son imprecisos, incompletos o ambiguos, especialmente, en bases de datos centralizadas. Los daños producidos por datos de pobre calidad ocasionan grandes daños a los procesos de negocios; por eso, son la piedra angular de una Inteligencia de Negocios eficiente. Las diferentes categorías de calidad de datos son: estandarización (consistencia), correspondencia (matching), verificación (frente a las fuentes) y mejora (añadido de datos para incrementar su utilidad). Un área de importancia creciente es la calidad de los datos que se procesan en tiempo real y de modo muy rápido.

Los problemas típicos son: datos incorrectos, datos redundantes, datos robados, datos irrelevantes o pérdidas de datos; las causas más comunes proceden de malas entradas de datos, diseño de bases de datos pobre, seguridad pobre, datos recogidos incorrectos o la inexistencia de los datos requeridos por el negocio.

Integridad de los datos

Aunque existen numerosas definiciones, tal vez, la definición más simple, y que recoge la idea central del concepto es: Integridad de datos. Esto significa que los datos deben ser precisos, correctos y válidos.

Es preciso que los sistemas de información proporcionen herramientas para conseguir la mayor calidad de datos posibles, y la integridad y fiabilidad de los datos de la empresa.

4.3 ADMINISTRACIÓN DE ARCHIVOS

Durante las primeras décadas de la era de la computación, las aplicaciones de negocios gestionaban sus datos en entornos de gestión de archivos. Las aplicaciones requerían sus propios datos que, a su vez, se organizaban en un archivo de datos. Por consiguiente, en un entorno o sistema de gestión de archivos, cada aplicación tenía un archivo específico de datos relacionados con ellos, y que contenía todos los registros necesarios en la aplicación (contabilidad, facturación o nóminas). A medida que las empresas evolucionaban, se desarrollaban numerosas aplicaciones, donde cada una de las cuales se asociaban a un archivo de datos específico de la aplicación.

A medida que las bases de datos se iban implantando en las organizaciones y empresas, convivían los archivos y sistemas de gestión de archivos, las bases de datos y los sistemas de gestión de bases de datos. Esta convivencia fue produciendo inconsistencias e imprecisiones, aislamientos de datos, etcétera.

Ejemplo de esto son: la profusión de archivos (nombres de archivos repetidos, olvido de archivos, datos correctos en un archivo e incorrectos en otros, la no actualización simultánea del mismo dato en distintos archivos).

Los sistemas de bases de datos eliminan muchos de los problemas que se producen en los métodos de almacenamiento y acceso y recuperación de los datos en los archivos. Las bases de datos se instalan y ejecutan mediante un conjunto de programas de software denominados sistemas de gestión (administración) de bases de datos, que proporciona a los usuarios la manipulación de todos los datos.

La mayoría de las aplicaciones de una empresa en un sistema de información organiza los datos en una estructura jerárquica de información. La jerarquía de datos de una computadora se organiza en entidades de información: bits, bytes, campos, registros, archivos y bases de datos (figura 4.2). Un bit representa la unidad más pequeña de datos (0 ó 1) que puede procesar una computadora. Un byte (conjunto de ocho bits) representa un único carácter (letra, dígito, símbolo). Un campo es una palabra, un grupo de palabras o un número (“Pintor Murillo”, “Maestría”, “64.549”). Un registro es un conjunto de campos relacionados entre sí (nombre, edad, dirección, curso que estudia, etcétera.). Un archivo o archivo de datos es un conjunto de registros. Una base de datos es un conjunto de archivos relacionados (por ejemplo, una base de datos de un banco puede contener un archivo de clientes, de empleados o de productos).

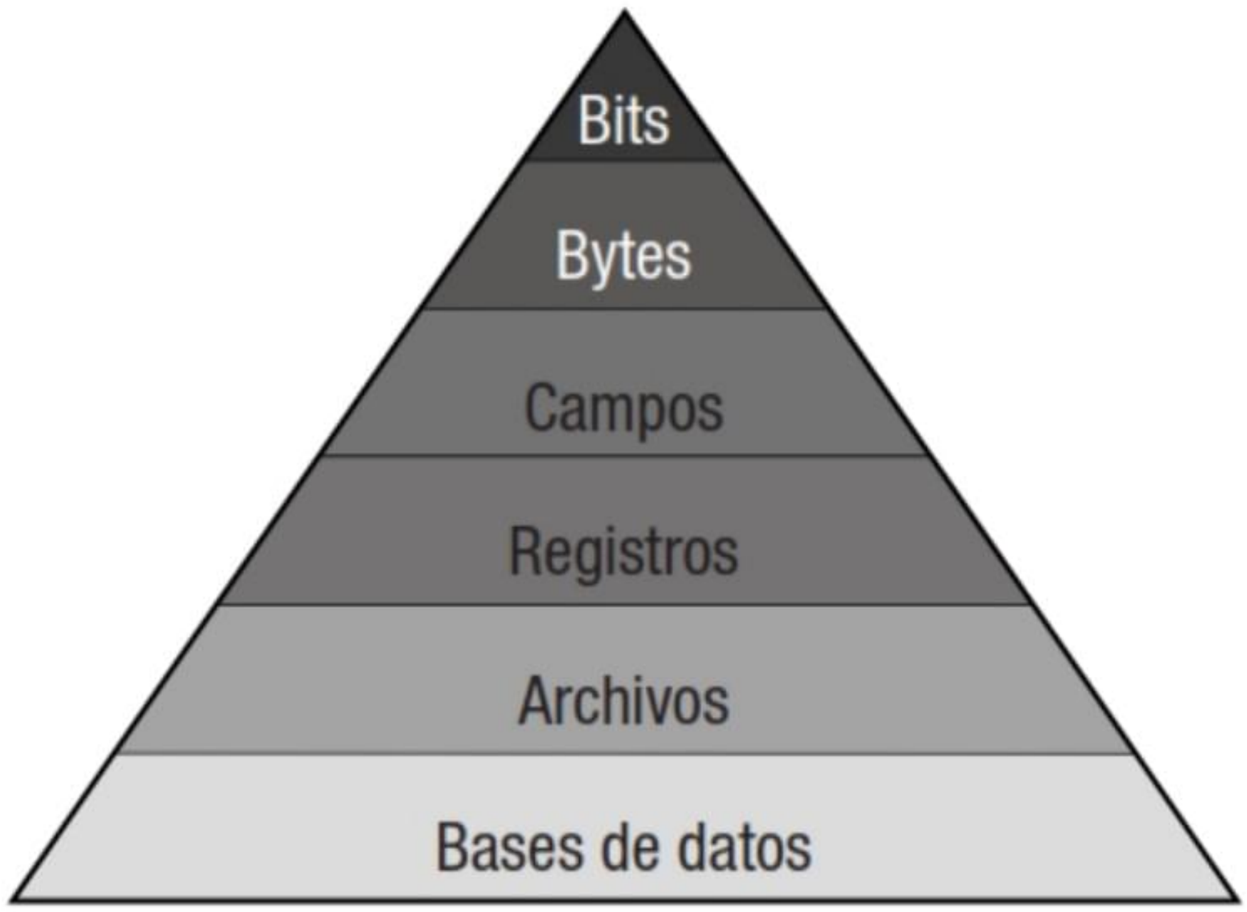


Figura 4.2. Jerarquía de datos.

4.3.1 PROBLEMAS EN LA GESTIÓN DE ARCHIVOS

Los sistemas de archivos, cuando se utilizan de modo independiente con sus respectivas aplicaciones, presentan los siguientes problemas:

- Redundancia de datos. Los mismos datos se almacenan en muchos lugares.
- Aislamiento de datos. La organización por archivos crea silos (islas) de datos, que dificultan un acceso desde aplicaciones diferentes; es decir, determinadas aplicaciones no pueden acceder a datos asociados en otras aplicaciones.
- Inconsistencia de datos. Versiones diferentes de datos no concuerdan; en la práctica, significa que los valores reales de los datos no están sincronizados en diferentes versiones de un archivo.
- Seguridad de los datos. Se dificulta en los sistemas de archivos, debido a que el aumento de aplicaciones que gestionan los datos aumenta también el número potencial de personas que pueden acceder a ellos, con diferentes perfiles e identificaciones, y mayores riesgos para la seguridad.

Por estas razones, aunque el uso de archivos independientes o conectados sigue siendo la práctica habitual en muchas empresas, cada día está ya más implantando el uso de las bases de datos, especialmente, relacionadas, y, a un nivel superior, los almacenes de datos (*Data Warehouse* y *Data Marts*).

4.4 BASES DE DATOS

Una base de datos es una agrupación lógica organizada de archivos relacionados entre sí. Una base de datos de estudiantes de una universidad se puede crear agrupando un archivo de los cursos, un archivo de los alumnos matriculados en los cursos, un archivo de los costes de la matrícula de cada alumno, etcétera. En una base de datos, los datos se integran y se relacionan de modo que existe un conjunto de programas de software que proporcionan acceso a todos los datos (que pueden ser compartidos). La ventaja de la base de datos es que la seguridad de integridad de los datos es mucho más alta que en los sistemas de archivos y que las aplicaciones y los datos son independientes unos de otros, además de resolver prácticamente casi todos los problemas que presentan los sistemas de gestión de archivos.

Existen dos grandes tipos básicos de bases de datos: centralizadas y distribuidas.

4.4.1 BASES DE DATOS CENTRALIZADAS

La posibilidad de proporcionar un control centralizado de los recursos de información de una empresa en su conjunto fue considerada, durante décadas, como una de las más fuertes motivaciones para la introducción de las bases de datos en los primitivos sistemas de información.

Una base de datos centralizada (figura 4.3) almacena todos los archivos relacionados en una única posición lógica. Las grandes computadoras (mainframes) soportaban bases de datos en este tipo, debido a los enormes costes de implantación y mantenimiento. Las bases de datos centralizadas ofrecen muchos beneficios a la organización y empresa, como sucede con cualquier sistema centralizado, derivado de la centralización de los servicios, pero también son más vulnerables, ya que cuando la computadora central falla, todos los usuarios se ven afectados.

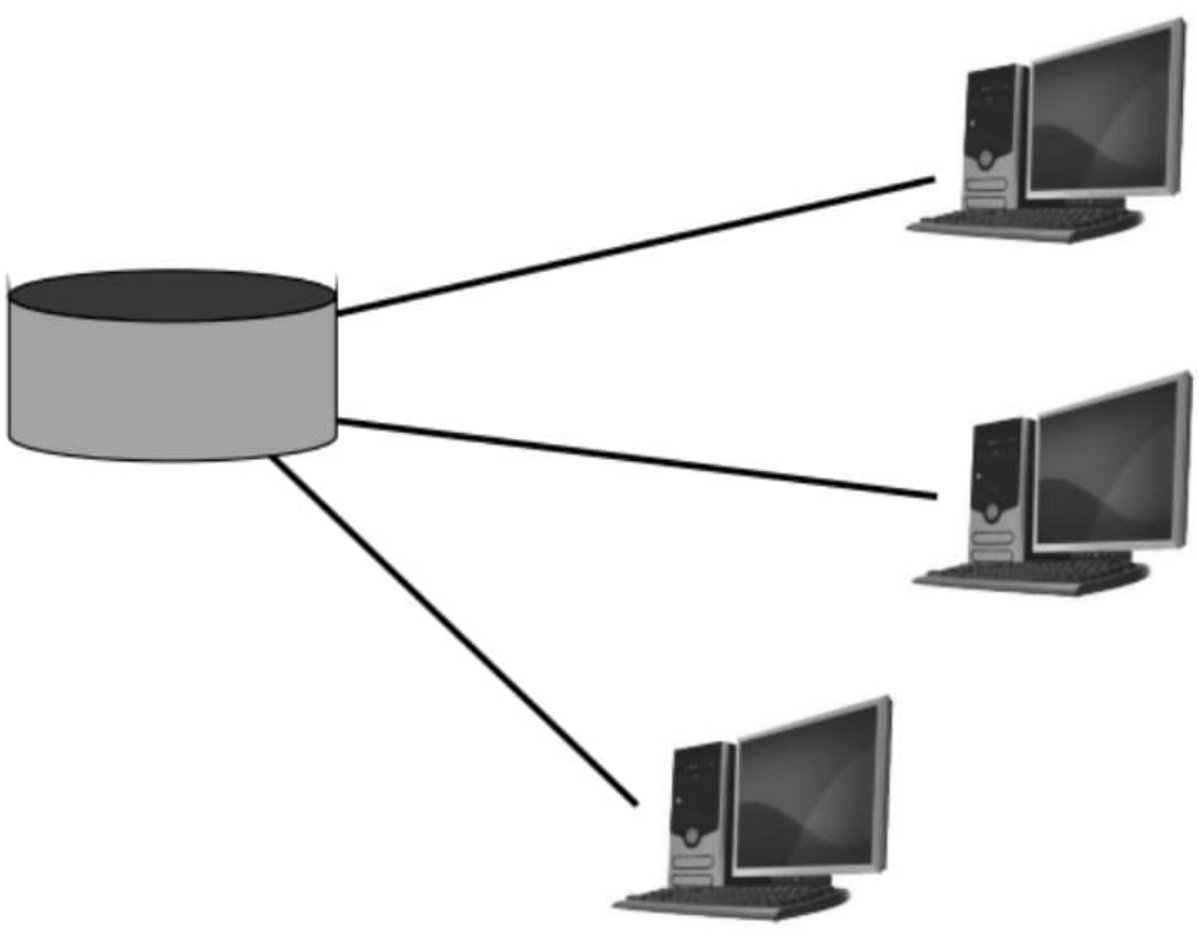


Figura 4.3. Bases de datos centralizadas.

4.4.2 BASES DE DATOS DISTRIBUIDAS

Una base de datos distribuida tiene copias completas de una base de datos, o parte de una base de datos, en más de una posición, que, normalmente, está próxima al usuario.

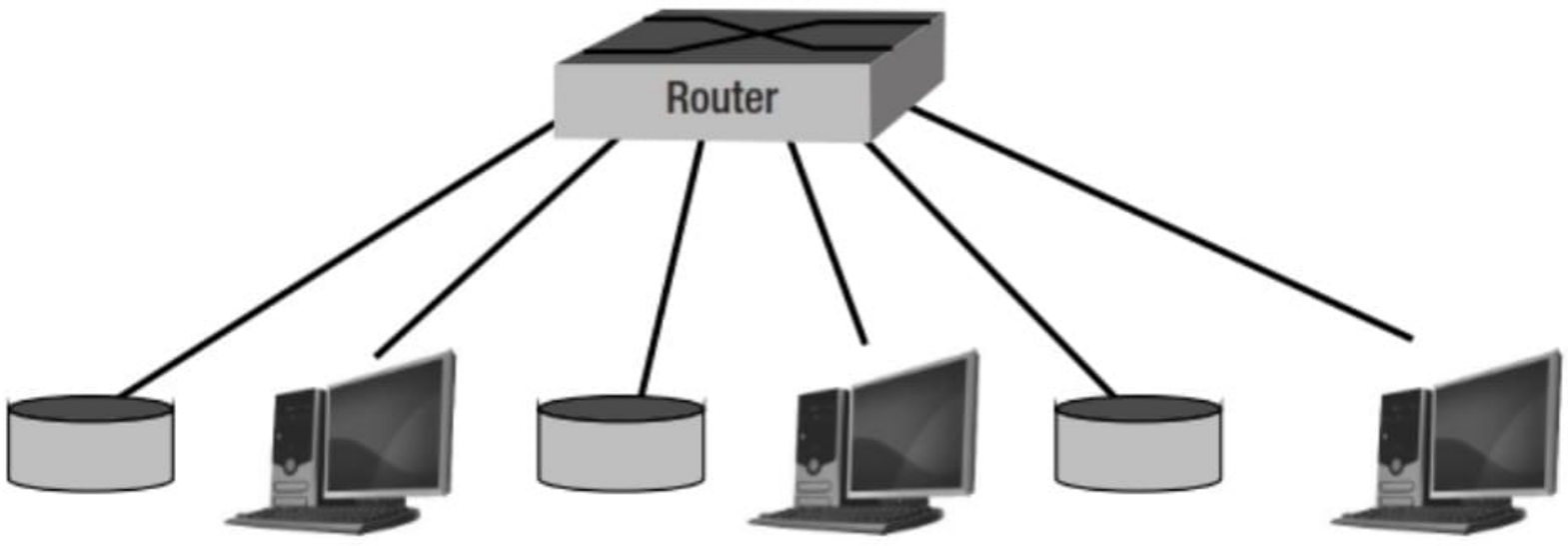


Figura 4.4. Bases de datos distribuidas.

Los datos de una base de datos distribuida pertenecen lógicamente a un solo sistema, pero se encuentran físicamente esparcidos en diferentes sitios de la red. Constan de una colección de sitios mediante algún tipo de red de comunicaciones, en el cual cada sitio es, a su vez, un sistema de bases de datos completo, pero todos los sitios trabajan juntos, a fin de que un usuario pueda acceder a los datos de la red desde cualquier lugar, de igual forma que si los datos estuvieran guardados en el propio sitio del usuario.

Las bases de datos distribuidas se dividen, a su vez, en otras dos categorías: replicadas y particionadas.

- Una base de datos replicada tiene copias de la base de datos completa en muchas posiciones, con el objeto de resolver los fallos que se pueden presentar en una base de datos centralizada, así como aumentar la capacidad de respuesta del acceso del usuario. Sin embargo, tiene la desventaja del mantenimiento de la consistencia de los datos cuando se actualizan los registros con las operaciones de añadir, modificar y eliminar.
- Una base de datos particionada está subdividida, de modo que cada posición tiene una parte de la base de datos completa, que cubre las necesidades del usuario local. Estas bases de datos proporcionan la necesaria velocidad de respuesta de archivos localizados sin la necesidad de replicar todos los cambios en múltiples posiciones. Otra ventaja significativa es que los datos en los archivos se pueden introducir más rápidamente, manteniendo la exactitud de los datos.

4.5 *DATA WAREHOUSE*

Actualmente, las empresas que tienen más éxito en sus estrategias empresariales son aquellas que responden con mayor flexibilidad y rapidez a los cambios y oportunidades que ofrecen los mercados. Uno de los factores clave en esas estrategias es el uso eficiente y efectivo de los datos y de la información por los analistas y gerentes. El reto en las organizaciones es proporcionar a los usuarios acceso a los datos corporativos, de modo que se puedan analizar con precisión y fidelidad, a fin de tomar mejores decisiones.

Las empresas, además de los sistemas de gestión de archivos y de bases de datos, utilizan una variedad de herramientas denominadas *Data Warehouses* (almacenes de datos) y *Data Marts* (almacenes de datos departamentales), que pretenden proporcionar a los usuarios, del modo más fácil y rápido posible, el acceso, el análisis y la consulta de datos.

Los *Data Warehouses* y los *Data Marts* son el soporte fundamental de las aplicaciones de Inteligencia de Negocios (*Business Intelligence*). La Inteligencia de Negocios es una amplia categoría de aplicaciones, procesos y tecnologías que

se encarga de reunir, almacenar, acceder y analizar datos que ayuden a los usuarios de los negocios a una mejor toma de decisiones.

Un *Data Warehouse* es un gran almacén o depósito de datos, donde se integran datos procedentes de varias fuentes: internas (procedentes de los sistemas transaccionales de los diferentes departamentos de la empresa, tales como recursos humanos, marketing, ingenierías, etcétera), externas y personales.

El concepto de *Data Warehouse* fue un concepto introducido por Bill Inmon, a principios de la década de los noventa, que lo definió como "una colección de datos orientado a temas, integrado, variable con el tiempo y no volátil para ayudar al proceso de gestión en toma de decisiones en una organización". Esta definición clásica viene a decir que un *Data Warehouse* es un depósito o repositorio (repository) de datos históricos que se organizan por temas para ayudar a la toma de decisiones en una gran organización. Posteriormente, Ralph Kimball propuso una definición más simple: "un *Data Warehouse* es una copia de los datos transaccionales específicamente estructurados para consultas y análisis".

Desde el punto de vista práctico, un *Data Warehouse* es una gran base de datos orientada al análisis de la información histórica; es decir, un depósito (repositorio) de datos históricos que se organizan por temas para el apoyo en la toma de decisiones. Por ese motivo, debe disponer de una gran capacidad de almacenamiento (decenas y centenas de gigabytes y terabytes en las grandes empresas, aunque el petabyte comienza a ser la unidad de almacenamiento en grandes establecimientos tipo Walmart, en Estados Unidos), ya que los datos deben permanecer, por largos períodos de tiempo, "no volátiles" o "históricos".

4.5.1 CARACTERÍSTICAS DE UN *DATA WAREHOUSE*

Las características fundamentales de un *Data Warehouse* son las recogidas por Bill Inmon en su definición clásica de *Data Warehouse* (y algunas otras que se han ido introduciendo en estas dos décadas de edad del término):

- **Orientado a temas (entidades).** Los datos contenidos en un *Data Warehouse* se organizan por temas (entidades): cliente, vendedor, producto, precios, región, que contienen sólo información relevante para la toma de decisiones.

  Un *Data Warehouse* difiere de una base de datos operacional en que la mayoría de las bases de datos operacionales tiene una orientación al producto y se sintoniza para manipular transacciones que actualizan la base de datos. La orientación a temas proporciona una visión más completa y comprensiva de la organización. Durante el análisis de Inteligencia de Negocios, la orientación hacia temas permite que el desempeño

(rendimiento) de una compañía se pueda evaluar más fácilmente y también se pueda detectar cualquier fuente de ineficiencia.

- **Integrado**. Los datos que se producen en las diferentes fuentes se integran y homogenizan a medida que se cargan en un *Data Warehouse*. Por ejemplo, los datos de los clientes se pueden extraer de los sistemas internos y externos, y son integrados alrededor de un identificador de cliente para crear una visión completa y exhaustiva del mismo.

- **Variable con el tiempo**. Un *Data Warehouse* mantiene datos históricos. A diferencia de los sistemas transaccionales, que sólo mantienen datos recientes (día, mes, año actuales), un *Data Warehouse* puede almacenar años de datos, es decir, datos históricos de varios años. Cada *Data Warehouse* tiene una calidad temporal. El tiempo es una dimensión importante que deben soportar todos los *Data Warehouses*.

- **No volátil**. Después de que los datos se han introducido en un Data Warehouse, los usuarios no pueden cambiar o actualizar los datos. Los datos obsoletos se eliminan y los cambios se registran como nuevos datos. Los datos en los *Data Warehouses* sólo pueden ser cambiados o actualizados por profesionales de TI —responsables de los sistemas de información—, es decir que los *Data Warehouses* se actualizan, pero sólo a través de procesos de carga (entrada) en lugar de los usuarios.

Algunas características adicionales a las cuatro fundamentales de la definición de Bill Inmon son:

- **Multidimensional**. Un *Data Warehouse* puede utilizar una estructura relacional o multidimensional. Las bases de datos relacionales almacenan datos en tablas bidimensionales. En contraste, los *Data Warehouse* almacenan datos en una estructura multidimensional que consta de más de dos dimensiones. Una representación típica de esta estructura son los cubos de datos, que se describirán más adelante. Los datos en los almacenes de datos se organizan por dimensiones del negocio, temas o entidades (productos, ciudades, empleos o períodos de tiempo). Las dimensiones del negocio o temas son aristas del cubo de datos (figura 4.5).

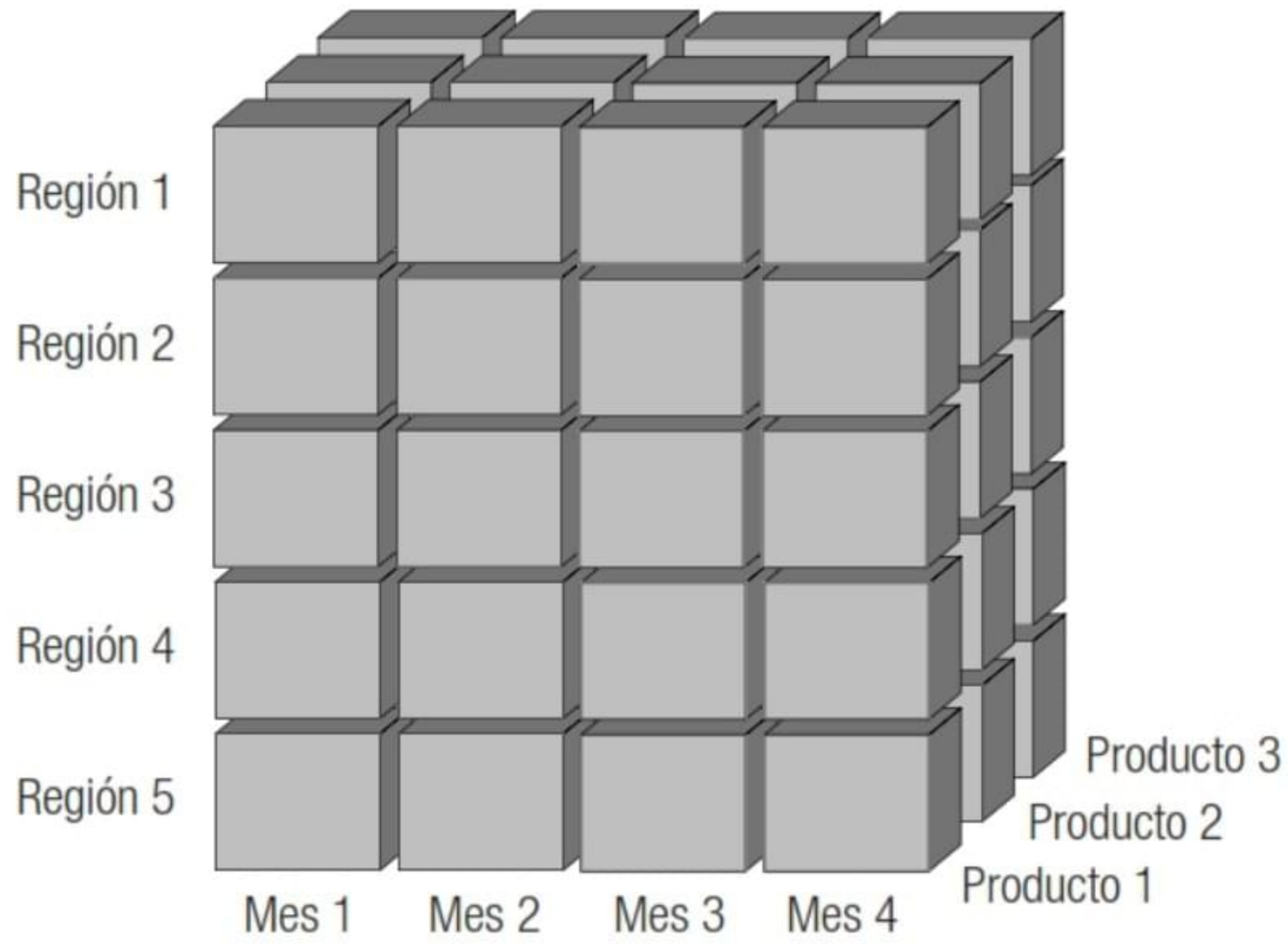


Figura 4.5. Cubo de datos.

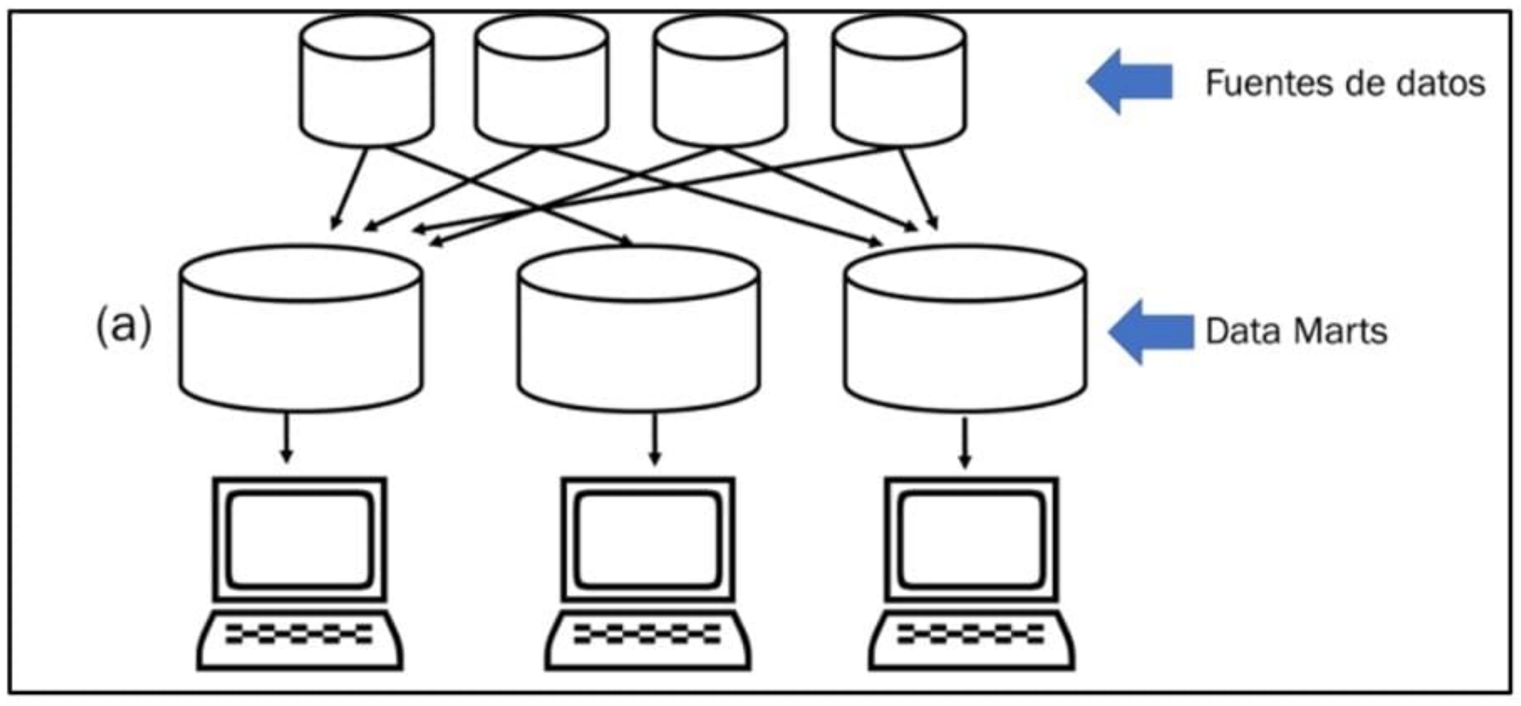
- Basados en la Web. Los *Data Warehouse* se diseñan, normalmente, para proporcionar un entorno de computación eficaz para aplicaciones basadas en la Web.
- Cliente/servidor. Un *Data Warehouse* utiliza la arquitectura cliente/servidor para proporcionar acceso fácil a los usuarios finales.
- Tiempo real. Los nuevos *Data Warehouse* permiten gestión de datos en tiempo real.
- Metadatos. Un *Data Warehouse* contiene metadatos (datos que generan datos), que facilitan la gestión de datos complejos.
- Consolidados. Los datos almacenados en un *Data Warehouse* se obtienen, normalmente, como resúmenes parciales de datos principales que pertenecen a los sistemas operacionales de los cuales se originaron. Por ejemplo, un operador de telefonía celular puede almacenar en un *Data Warehouse* el costo total de las llamadas realizadas por cada cliente en una semana, subdividida por rutas de tráfico y por el tipo de servicio seleccionado (llamada, mensaje SMS, datos, etcétera.), en lugar de almacenar las llamadas individuales registradas por el sistema operacional. La razón de esta característica de consolidación puede ser doble: por un lado, reducir el espacio requerido para almacenar en el *Data Warehouse* los datos acumulados durante los años de antigüedad del cliente, y, por otro, la información consolidada puede mejorar las necesidades de los sistemas de negocio.

4.6 *DATA MART*

Los *Data Warehouses* suelen ser unas herramientas muy caras y difíciles de implantar, por lo que se utilizan, normalmente, en grandes organizaciones y empresas. Un *Data Martes* un sistema que reúne todos los datos requeridos por un departamento específico de una empresa, tales como marketing, recursos humanos, logística o administración, con el objeto de realizar análisis de Inteligencia de Negocios y ejecución de aplicaciones de apoyo a las decisiones de funciones específicas concretas. Por consiguiente, un *Data Mart* puede ser considerado como un almacén de datos departamental o funcional, de un tamaño más pequeño y aplicado a un departamento específico, en lugar del *Data Warehouse* global de la empresa.

Un *Data Mart* contiene un subconjunto de los datos almacenados del *Data Warehouse* de la compañía, que están, normalmente, integrados con otros datos que pueda tener el departamento responsable del *Data Mart* de cualquier otro tipo. Por ejemplo, un *Data Mart* de un departamento de marketing tendrá datos extraídos del *Data Warehouse* central de la compañía (información de clientes y operaciones de ventas), pero también tendrá datos adicionales específicos del departamento de marketing, tales como resultados de campañas de marketing pasadas tanto en negocios tradicionales como en Internet.

Existen dos categorías de *Data Mart*: dependiente o independiente. Un *Data Mart* dependiente es un subconjunto de datos que se crea directamente del *Data Warehouse*. Un *Data Mart* independiente es un pequeño *warehouse* diseñado para un departamento o unidad de negocio estratégica. Un *Data Mart* dependiente tiene las ventajas de utilizar un modelo de datos consistente y proporcionar calidad de datos. El problema que se presenta es la necesidad de construir, en primer lugar, el *Data Warehouse* y, a continuación, el *Data Mart*. La ventaja del *Data Mart* dependiente es que el usuario final visualiza la misma versión de datos a la que acceden todos los restantes usuarios del *Data Warehouse*.



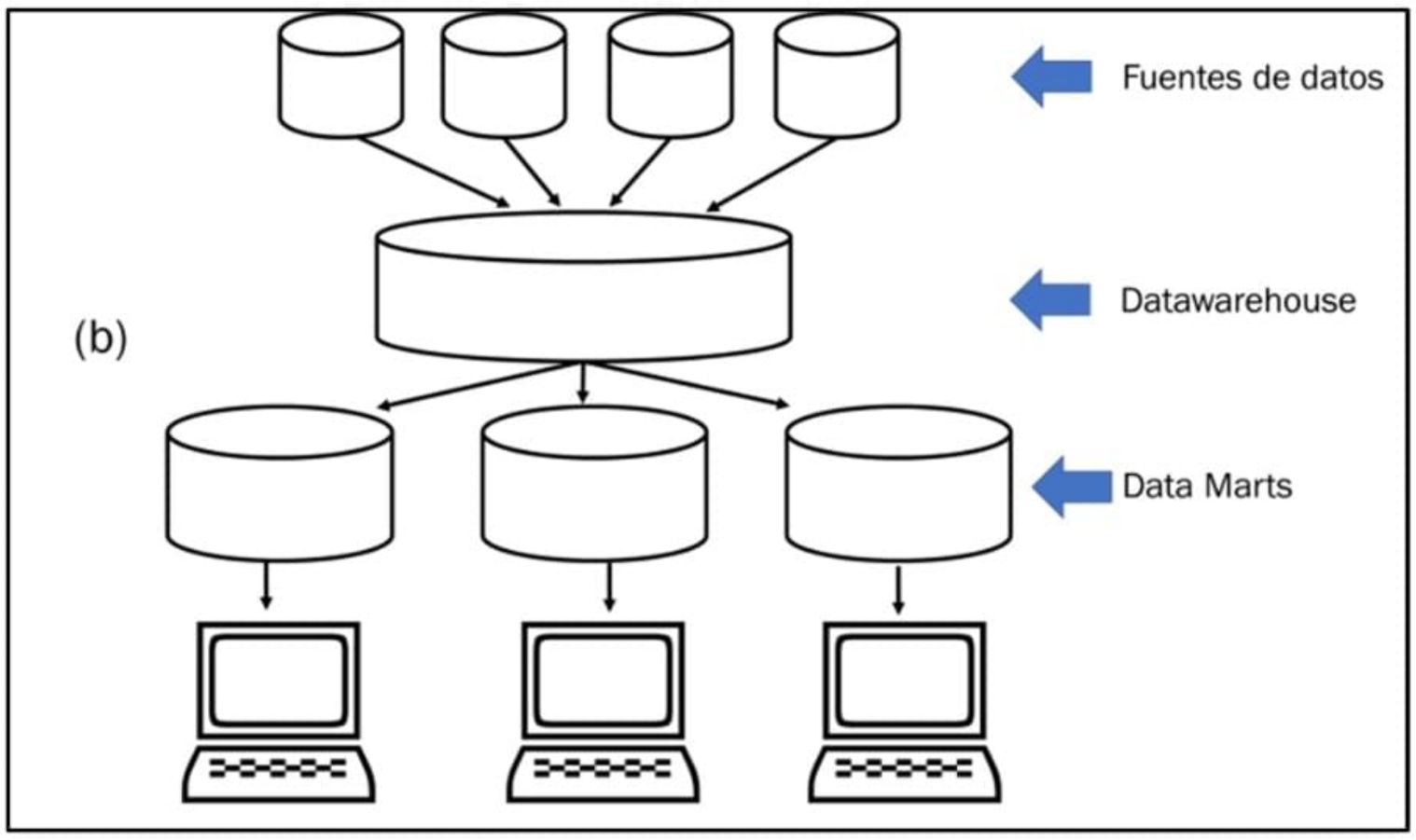


Figura 4.6. Data Mart: (a) Independientes; (b) Dependientes.

Uno de los problemas importantes de la implementación de un *Data Warehouse* es su alto costo de desarrollo y mantenimiento, lo que reduce la construcción de almacenes de datos a grandes organizaciones y empresas. Por esta razón, han surgido alternativas a los *Data Warehouse* y se construyeron versiones más reducidas, diseñadas para resolver los objetivos de departamentos, unidades funcionales o pequeñas empresas. Estos subconjuntos de *Data Warehouse* son los *Data Mart* independientes y, normalmente, pueden ser construidos más rápidamente y a un menor coste que un *Data Warehouse* completo de una empresa.

Los *Data Warehouse* y los *Data Mart* comparten el mismo marco de trabajo tecnológico. Según el tipo de empresa, la implementación de aplicaciones de Inteligencia de Negocios se puede realizar mediante el diseño y desarrollo de modo incremental de un conjunto de *Data Mart* integrados, en lugar de un gran *Data Warehouse* central. La razón fundamental es temporal y económica. Un *Data Mart* se puede implantar más rápidamente y más económicamente que un *Data Warehouse*; cifras de especialistas hablan de dos a tres meses para un *Data Mart*, y no menos de seis a nueve meses para un *Data Warehouse*. De acuerdo con el tamaño y las estrategias de la empresa, puede ser recomendable utilizar *Data Mart* y dar el paso de la centralización cuando las circunstancias lo exijan por volumen de transacciones, número de datos, clientes, etcétera. Probablemente, a la empresa que sólo requiere pocas aplicaciones de Inteligencia de Negocios le será suficiente la implantación de uno o varios *Data Marts.*

4.6.1 *DATA WAREHOUSE* VERSUS *DATA MART*

Un *Data Warehouse* o, con mayor precisión, un almacén de datos corporativos es un almacén o repositorio de datos que recolecta y consolida los datos de múltiples sistemas de fuentes, tanto internas de la organización como externas, con el propósito de facilitar su análisis. Un *Data Warehouse* normalmente tiene las siguientes características:

- Grande en tamaño (terabytes y, cada día, más petabytes).
- Grande en ámbito (recogen información de una amplia variedad de fuentes).
- Facilitan la integración de datos (compilan y recogen datos de múltiples fuentes, asegurando que los datos sean precisos y actuales).
- Diseñados para Analítica (centrados en el análisis y en la Analítica, para lo que están optimizados).

Un *Data Mart*es una versión reducida de un *Data Warehouse*, que se centra en las necesidades de un departamento o audiencia específica. Al igual que un *Data Warehouse*, un *Data Mart* es un repositorio de información, construido explícitamente para facilitar el análisis de los datos. El *Data Mart* está diseñado para cumplir las necesidades específicas de un grupo, comunidad o departamento y su gran ventaja es el tamaño más pequeño con relación al *Data Warehouse* y, por consiguiente, es más fácil de construir. La figura 4.6 muestra las dos arquitecturas de *Data Mart* típicas de una organización.

4.7 MARCO DE TRABAJO (*FRAMEWORK*) DE UN SISTEMA DE ALMACENAMIENTO DE DATOS

Un *Data Warehouse* de una empresa (*EDW, Enterprise Data Warehouse*) es un almacén de datos que se utiliza en la empresa y es el depósito más importante de datos disponibles para el desarrollo de arquitecturas de Inteligencia de Negocios y sistemas de apoyo a la decisión. El término *Data Warehousing* (almacenamiento de datos) indica el conjunto completo de actividades interrelacionadas implicadas en el diseño, implementación y uso de un *Data Warehouse*.

Las organizaciones, públicas y privadas, recogen datos, información y conocimiento de modo continuo que se almacenan en sistemas computarizados. El mantenimiento y uso de esos datos y la información se vuelven complejos, sobre todo, cuanto se necesitan escalar esos datos a volúmenes de datos mayores. El trabajo y la gestión con múltiples bases de datos integradas o no en un *Data Warehouse* se vuelven tareas complejas y extremadamente difíciles, que cada día más requieren sólidos conocimientos técnicos y experiencias; pero una

buena manipulación de estas herramientas proporciona grandes beneficios que compensarán los excesivos costes que supone su implementación y actualización continua.

El marco de trabajo o entorno (framework) de un sistema de *Data Warehouse* y de *Data Mart* comienza extrayendo datos desde fuentes de datos (*data sources*), mediante un software denominado ETL (Extract, Transform y Load: extraer, transformar y cargar), y se carga en un área provisional, donde se transforman y limpian.

Una vez leídos los datos, se cargan en el depósito de datos (*Data Warehouse*). A continuación, los datos son analizados, minados, presentados y visualizados, utilizando herramientas de negocio final como OLAP, Minería de Datos y de Texto, paneles o tableros de control (*dashboards*), etcétera.

4.7.1 COMPONENTES DE UN *DATA WAREHOUSE*

El marco o entorno de trabajo de un *Data Warehouse* incluye, al menos, los siguientes componentes fundamentales (Turban 2011):

- Fuentes del sistema. Internas, externas y personales, que proporcionan datos al *Data Warehouse* o *Data Mart*, según su caso.
- Integración de datos. Tecnologías y procesos que se necesitan para preparar los datos para su uso (sistemas ETL: procesos de "extracción", "transformación" y "carga" de los datos); es decir, se extraen los datos utilizando después un software comercial denominado ETL o software escrito por el usuario (organización o empresa).
- Arquitectura de almacenamiento de datos para su almacenamiento en el *Data Warehouse* o *Data Mart* de una organización.
- Herramientas y aplicaciones para los diferentes tipos de usuarios, que deberán aprender a utilizar o, en su caso, desarrollar.
- Acceso a los datos (middleware). Herramientas que facilitan el acceso al *Data Warehouse*. Actúan de intermediación entre los dispositivos de almacenamiento y los usuarios que utilizan los datos con las aplicaciones y herramientas adecuadas.
- Metadatos, calidad de datos y procesos de gobierno, que deberán cumplir con sus especificaciones y asegurar que los almacenes de datos cumplen sus objetivos. Debido a sus características especiales entre tecnologías y políticas de gobierno de TI, les dedicaremos unos apartados especiales.

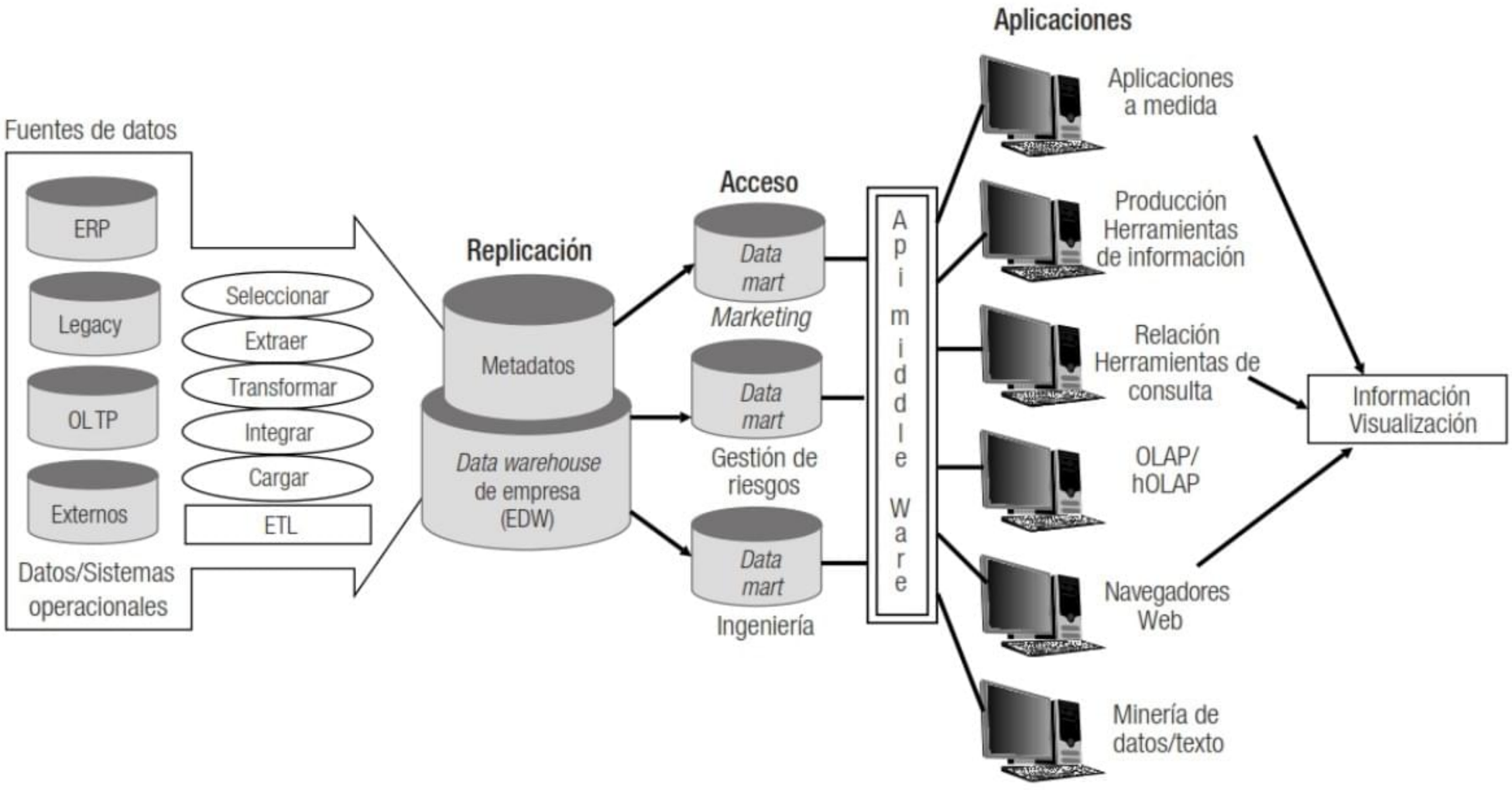


Figura 4.7. Marco de trabajo de un sistema de *Data Warehouse*.
Fuente: (Turban 2011:334) adaptada.

4.7.2 FUENTES DE DATOS

Normalmente, los datos procederán de múltiples fuentes: sistemas operacionales/transaccionales OLTP (*On Line Transaction Processing*: procesamiento de transacciones en línea), sistemas de planificación de recursos ERP, datos de sitios web, datos propios o ajenos heredados, *legacy* (datos de estadísticas del gobierno, de la Unión Europea, de Eurostat, de la oficina del Censo de Argentina), datos de terceras partes, etcétera. La tendencia actual de Big Data es incluir datos procedentes de sensores, etiquetas (tags), chips, RFID y NFC, geolocalización, etcétera. Estos sistemas de fuentes utilizarán diferentes sistemas de software, tales como Oracle, SAP, IBM, y almacenarán los datos en diferentes formatos (bases de datos relacionales, multidimensionales, jerárquicas, etcétera).

Las fuentes de datos más típicas de una empresa son las bases de datos operacionales, que son bases de datos relacionales (bidimensionales), y los *Data Warehouse*; además, utilizarán sistemas multidimensionales (bases de datos multidimensionales y OLAP).

Un tema que deben afrontar las empresas es la redundancia de los datos: tienen muchos sistemas que contienen los mismos datos; en estos casos, las empresas deben seleccionar cuál es la mejor fuente o el mejor sistema de datos.

Otro concepto importante para tener en cuenta es la granularidad, es decir, el nivel de detalle con el que se necesitan los datos; no se requiere el mismo detalle en las transacciones diarias de un cliente que en un banco, o al momento de solicitar un préstamo personal o una hipoteca para comprar una casa.

4.7.3 CATEGORÍAS DE DATOS

En lo relativo a la alimentación de datos al *Data Warehouse*, los datos se pueden agrupar en tres grandes categorías: datos internos, datos externos y datos personales (Vercellis, 2009).

Datos internos

Los datos internos almacenados en la mayoría de las bases de datos se conocen como sistemas transaccionales o sistemas operacionales y constituyen la espina dorsal (*the backbone*) de un sistema de información de empresa. Los datos internos se recolectan de aplicaciones transaccionales en las operaciones rutinarias de una empresa (administración, contabilidad, logística, etcétera). Este conjunto de aplicaciones de software se denomina ERP (*Enterprise Resources Planning*).

Los datos almacenados en los sistemas operacionales, normalmente, tratan con las entidades principales implicadas en los procesos de negocio de las empresas: clientes, productos, ventas, empleados, proveedores y socios. Estos datos provienen de diferentes componentes de un sistema de información:

- Sistemas de *back office*. Reúnen registros transaccionales básicos (pedidos, inventarios, producción, factura, y procesos logísticos).
- Sistemas de *front office*. Contienen datos que proceden de atención al cliente, actividades de centros de llamadas (*call centers*) y compañías de marketing.
- Sistemas basados en la Web. Reúnen datos de ventas de sitios web en operaciones de comercio electrónico, visitas a sitios web, historial de visitas, etcétera.

Datos externos

Los datos externos pueden proceder también de diferentes fuentes y se utilizan para ampliar la información extraída de las bases de datos internas y considerar la situación actual y futura de la empresa. Por ejemplo, datos de gran interés son los datos estadísticos publicados por organizaciones nacionales e internacionales, consultoras y bancos, que muestran tendencias de mercado, hábitos de compras, entre otros.

Hoy en día existen numerosas y significativas fuentes externas de información: Sistemas de Información Geográfica (SIG o GIS, *Geographic Information Systems*), cuyas aplicaciones facilitan la adquisición, organización, almacenamiento y presentación de datos del territorio y, esencialmente, datos de posición geográfica específicos (coordenadas de latitud y longitud) de la organización, departamentos, sucursales, y, fundamentalmente, datos de geolocalización y sistemas GPS.

En la actualidad, otras fuentes de información vitales para las empresas son los datos procedentes de medios sociales (redes sociales generalistas, como Facebook, Twitter o Tuenti, o redes sociales profesionales como Linkedin, Vimeo, Xing, etcétera.).

Datos personales

Cada día es más frecuente el análisis de la información almacenada en hojas de cálculo, bases de datos locales, mensajes de textos, boletines de noticias, archivos y documentos adjuntos profesionales en correos electrónicos, etcétera. También, es habitual la integración de información operacional de la compañía con datos estructurados y no estructurados de empleados, clientes, proveedores (naturalmente, con el cumplimiento estricto de las leyes de protección de datos y privacidad vigente), que permitan su conversión en conocimiento para la toma de decisiones. En otras palabras, los datos personales formarán parte sustancial del sistema de gestión del conocimiento de la empresa, y el uso adecuado de herramientas colaborativas de flujo de trabajo condicionará la mejora en los procesos de tomas de decisiones (Capítulo 12).

4.7.4 INTEGRACIÓN DE DATOS

Dada la gran cantidad de fuentes de información distribuidas en diferentes lugares de una organización, en muchos casos, heterogénea, hace falta integrar toda esta información así como promover y facilitar su acceso. Es decir, se necesita extraer los datos de las diferentes fuentes, transformarlos (limpiar) para hacerlos eficaces y cargarlos en un *Data Warehouse* o, en su defecto, un *Data Mart*. Este proceso, de modo tradicional, se ha conocido como ETL (*Extraction, Transform, Load*), pero cada día más, se comienza a utilizar el término integración de datos debido, precisamente, al número creciente de modos en que los datos fuente pueden ser manipulados, tanto por la secuencia de operaciones como por la disparidad de fuentes sociales, sobre todo, porque en muchos casos no pueden ser tratados por herramientas ETL tradicionales.

La integración de datos se puede conseguir mediante técnicas diferentes que se agrupan, fundamentalmente, en dos categorías: extracción de datos mediante consulta con SQL (el lenguaje de consulta) en bases de datos o mediante software comercial especifico de integración de datos (ETL). También, es posible un tercer método, que es el desarrollo propio a medida del usuario, aunque cada vez es menos utilizado, ya que las herramientas ETL comerciales, o herramientas de aplicaciones web (apps) para escritorio o teléfonos inteligentes y tabletas, son cada día más abundantes.

El objetivo fundamental de la integración de datos es transformarlos haciéndolos más fáciles de utilizar. Por ejemplo, datos procedentes de diferentes fuentes se pueden integrar alrededor de una clave común, como puede ser la identidad digital de un usuario o cliente (nombre de usuario y contraseña).

Cada día, es más frecuente que el usuario acceda a Internet y, en consecuencia, a las aplicaciones corporativas, a través de diferentes fuentes: correo electrónico, la Web, contactos directos, call center, teléfono celular inteligente, mensajería instantánea (bien mediante los mensajes SMS tradicionales o con aplicaciones de mensajería instantánea (MI) como WhatsApp, Viber, Telegram, Line, Spotbros, Hangout, WeChat, etcétera) y tabletas. Imaginemos el caso de un banco: un cliente puede acceder directamente en la oficina de la empresa, mediante un cajero automático (ATM), por correo electrónico a través del sitio web del banco (con diferentes dispositivos), a través de las redes sociales del banco (Facebook, Twitter), etcétera. Por eso, es necesario integrar los datos de las diferentes fuentes del banco en un *Data Warehouse*.

Por otra parte, la integración de datos puede implicar otros tipos de transformaciones, por ejemplo, en lo relativo a formatos, género (hombre/mujer), fecha (día, mes, año), profesión, o en cuanto a duplicación de datos del cliente, originados por la diversidad de métodos de acceso y diferentes aplicaciones de una misma organización, etcétera. Estas integraciones de datos pueden exigir no sólo diferentes formas de extracción, sino de limpieza y carga de los datos.

Otra cuestión clave es la duración del tiempo de actualización. En muchos negocios, comienzan a requerirse que los datos se almacenen en los *Data Warehouse* cada muy poco tiempo, minutos o decenas de minutos. Imaginemos el caso de grandes almacenes (Walmart en Estados Unidos, El Palacio de Hierro en México, El Corte inglés en España, Media Markt y Carrefour en Europa, entre otros), que pueden necesitar almacenar los datos de compra o venta en el *Data Warehouse* de la empresa cada 10 ó 20 minutos, para poder estar disponibles para los análisis de datos que sean necesarios. Es decir, en los últimos años, han nacido herramientas de *Data Warehouse* en tiempo real, que han facilitado la integración de datos procedentes de múltiples fuentes mediante la actualización de datos en períodos prácticamente de tiempo real.

4.7.5 ALMACENAMIENTO DE DATOS

Existen diferentes arquitecturas para el almacenamiento de datos que ayudan en la toma de decisiones. La arquitectura más usual es un *Data Warehouse* central de empresa (*EDW, Enterprise Data Warehouse*). El *Data Warehouse* se configura como una plataforma central para cada organización sin uso de *Data Mart*. Este es el enfoque de la mayoría de las organizaciones, y en esta arquitectura los datos del *Data Warehouse* son accesibles por todos los usuarios.

La segunda arquitectura es la de *Data Mart*s independientes. Esta arquitectura se suele adoptar cuando el costo de la implantación de un único *Data Warehouse* no es asumible por la empresa o porque así lo consideran sus directivos o responsables de los sistemas de información. En este caso, los datos se almacenan en una o diversas aplicaciones, tales como finanzas, ventas,

operaciones o marketing. Esta arquitectura está centrada en las aplicaciones para el almacenamiento de datos, aunque los datos se pueden utilizar para otras aplicaciones de la organización, mediante los diferentes conectores (aplicaciones de middleware o intermediarias). En este caso, normalmente se requieren procesos de programación de la propia empresa o del proveedor del software. Los *Data Mart* independientes no suelen ser muy eficaces en las grandes organizaciones. Sin embargo, cumplen muy bien las necesidades de las pequeñas organizaciones o las de un departamento en una gran empresa. Uno de los inconvenientes de los *Data Mart* procede de las posibles inconsistencias de los datos y de la utilización de sistemas de fuentes diferentes (que pueden contener diferentes datos para los mismos elementos (la dirección del cliente o definición de las actividades con la propia organización que pueden no estar anotadas en su totalidad).

Existe una tercera arquitectura que algunos proveedores recomiendan y adoptan las organizaciones: *hub and spoke* (concentrar y hablar). Esta arquitectura almacena datos en un *Data Warehouse* central, mientras mantiene *Data Marts* que obtienen sus datos del depósito central de datos. Como los *Data Mart* adquieren sus datos del almacén central pueden tener más fiabilidad y consistencia, precisamente, por proceder de una fuente central. En el caso de los *Data Mart* dependientes, los datos se suelen almacenar con formatos apropiados al uso que se les va a dar y, de este modo, proporcionan respuestas más rápidas para consultas y ejecución de aplicaciones.

4.7.6 *MIDDLEWARE*

Las herramientas de middleware (intermediación o adaptación) facilitan el acceso al *Data Warehouse*. Dado que existen numerosas aplicaciones para que los usuarios puedan ver los resultados de los análisis o visualizaciones, se requiere un componente idóneo para conseguir que las diferentes herramientas de los usuarios puedan acceder sin problemas a los *Data Warehouse* y *Data Mart*.

4.7.7 USUARIOS

Una vez que los datos ya se han almacenado en el *Data Warehouse* o *Data Mart*, pueden acceder a ellos los diferentes usuarios del sistema de Inteligencia de Negocios (desarrolladores TI, trabajadores directamente relacionados con BI, analistas, directivos y gerentes, proveedores, clientes, etcétera).

Los beneficios de un sistema de almacenamiento de datos incluyen que:

- Los usuarios finales pueden acceder rápida y fácilmente mediante navegadores web, ya que estos datos están localizados en lugares concretos.

- Los usuarios finales pueden realizar un análisis exhaustivo de datos con algún método que anteriormente no fuera posible.
- Los usuarios finales obtienen una visión consolidada de datos organizados.

Estos beneficios pueden, además de mejorar el conocimiento de negocio, proporcionar ventajas competitivas y mejoras al servicio y satisfacción del cliente, facilitando la toma de decisiones. Los *Data Warehouse*, pese a sus muchas ventajas, también presentan inconvenientes que será preciso tener presentes. Según el proyecto específico, la construcción y mantenimiento pueden resultar muy caros. Otro inconveniente es la incorporación de datos desde sistemas informáticos obsoletos, que puede resultar difícil y costosa. Por último, suele producirse un hecho incontestable: las personas de un departamento pueden ser reacias a compartir datos con otros departamentos, por numerosas causas (esta situación se produce igual que sucede en algunos casos de gestión del conocimiento, donde las personas no desean compartir sus conocimientos con otros empleados).

4.8 METADATOS, CALIDAD Y GOBIERNO DE UN *DATA WAREHOUSE*

A fin de documentar el significado de los datos contenidos en un *Data Warehouse*, se recomienda establecer una infraestructura de información específica conocida como metadato. Un metadato es un dato que describe otro dato, es decir, son datos acerca de datos.

Tanto el personal especialista en TI, que opera y gestiona el *Data Warehouse*, como los usuarios que acceden a los datos, necesitan metadatos. El personal de TI necesita información relativa a las fuentes de datos, bases de datos, tablas, uso de datos, etcétera. Las necesidades de los usuarios incluyen definiciones de datos, herramientas disponibles de informes/consultas (*report/query*), distribución de informes e información de contactos para ayuda y seguridad.

Los metadatos indican para cada atributo de un *Data Warehouse* la fuente original de datos, su significado y las transformaciones a las que ellos se han sometido. La documentación proporcionada por los metadatos debe mantenerse constantemente actualizada, con el objetivo de reflejar cualquier modificación en la estructura del *Data Warehouse*. La documentación debe ser accesible directamente a los usuarios de almacenes de datos a través de un navegador web o de un tablero de control (*dashboard*), mediante los derechos de acceso que cada usuario establezca.

Vercellis (2009) considera que los metadatos deben realizar las siguientes tareas informativas:

- Una documentación de la estructura de los almacenes de datos: diseño, vistas lógicas, dimensiones, jerarquía, datos derivados y localización de todos los *Data Mart*.
- Una documentación de la genealogía de los datos, obtenida por el etiquetado de las fuentes de datos de las cuales se extrajeron, que describe cualquier transformación realizada en ellos.
- Un listado que mantenga las estadísticas del uso del *Data Warehouse*, que indique especialmente cuántos accesos a campos o a una vista lógica se han realizado.
- Una documentación del significado de los almacenes de datos con respecto al dominio de la aplicación, que proporcione la definición de los términos utilizados y que describa las propiedades de los datos, así como el propietario de los datos y las políticas de carga que se han utilizado.

4.8.1 CALIDAD DE LOS DATOS EN UN ALMACEN DE DATOS

La calidad de los datos en el *Data Warehouse* debe ser la adecuada para el cumplimiento de las necesidades del usuario. Si no sucede así, los datos no serán fiables y al final no se utilizarán. Muchas empresas consideran que los datos en los sistemas de fuentes de datos son pobres y han de ser mejorados antes que los datos se puedan almacenar en el *Data Warehouse*.

La necesidad de verificar, preservar y mejorar la calidad de los datos es una preocupación constante del responsable de diseño y actualización de un *Data Warehouse*. Los principales problemas que pueden comprometer la validez e integridad de los datos son: datos incorrectos, datos no actualizados y pérdida de datos.

Los factores importantes que pueden afectar a la calidad de los datos son:

- Precisión (*accuracy*). Los datos deben ser altamente correctos.
- Completitud o compleción. Los datos deben ser completos y se ha de procurar que no se produzcan pérdidas de valores.
- Consistencia. El formato y contenido de los datos debe ser consistente a través de las diferentes fuentes de datos, después de los correspondientes procedimientos de integración.
- Oportunos. Los datos deben actualizarse con frecuencia basados en los objetivos del análisis (actualizaciones regulares: diarias, semanales).
- No redundantes. Se debe evitar la repetición y redundancia de datos para no malgastar memoria y prevenir posibles inconsistencias.

- Significativos. Los datos deben ser significativos (relevantes) a las necesidades del sistema de análisis y toma de decisiones (Inteligencia de Negocio), con el objetivo de añadir valor real a la ejecución de todos los análisis posteriores.
- Accesibilidad. Los datos deben ser fácilmente accesibles por los analistas y las aplicaciones de apoyo a decisiones.

4.8.2 GOBIERNO DEL *DATA WAREHOUSE*

En las organizaciones, es necesario asegurarse que los diferentes sistemas de información que se nutrirán de los datos de los almacenes de datos, desde los ERP hasta los sistemas de BI, cumplen sus necesidades. Por estas razones, se deben poner en marcha buenas prácticas de gobierno de TI y, en particular, las relacionadas con los negocios fundamentales de la organización.

El gobierno del sistema de *Data Warehouse* debe ser una parte del gobierno de TI y requiere que las personas, comités y procesos se realicen en el momento oportuno y cumpliendo las reglas de gobiernos de las aplicaciones correspondientes. Las organizaciones suelen crear comités de personal técnico y de negocios que priorizan los proyectos, asignan recursos y aseguran que los negocios y los sistemas de información, en particular los almacenes de datos, estén alineados. Es necesario que estos comités y los equipos de trabajo que nombren supervisen los diferentes proyectos, para asegurar que se cumplen en su totalidad y de modo eficaz. También es preciso, y ya desde un punto de vista operacional, que los equipos de trabajo operacional, con la supervisión de los comités citados, ejecuten las tareas de creación de la definición de datos, identificación y resolución de los problemas de los datos. Todos los comités de empresa para el gobierno de las TI requieren la colaboración y las contribuciones especializadas del personal de TI y de los procesos de negocio.

4.9 HERRAMIENTAS ETL

La integración de los datos en un sistema de almacenamiento de datos, como ya se ha comentado, se realiza con tecnología de ETL (*Extraction, Transformation, Load*). ETL se refiere a las herramientas de software que se dedican a la ejecución automática de las tres tareas principales: extraer, transformar y cargar. El proceso de carga en un *Data Warehouse* requiere la realización de las tres fases citadas.

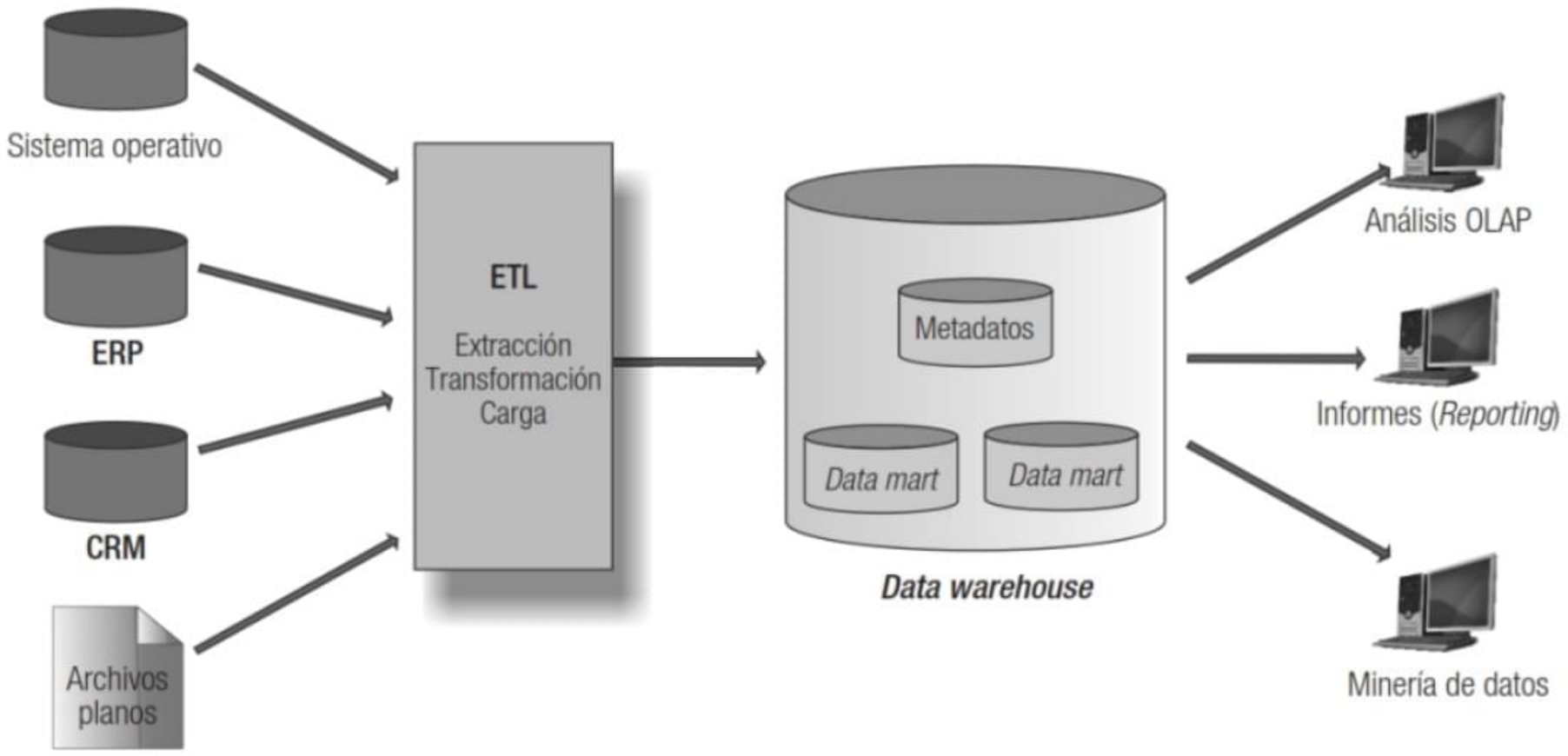


Figura 4.8. Herramientas ETL.

4.9.1 EXTRACCIÓN

Los datos se extraen de las fuentes internas y externas disponibles (lectura de datos de una o más bases de datos). Las fuentes de datos pueden constar de archivos extraídos de bases de datos OLTP, hojas de cálculo, bases de datos personales (Oracle, Microsoft, Access, etcétera) o archivos externos y archivos planos.

Es necesario considerar dos momentos en la extracción de datos: la extracción inicial, donde los datos disponibles relativos a períodos anteriores se introducen en los *Data Warehouse* vacíos, y las extracciones incrementales posteriores, que actualizan el *Data Warehouse* utilizando nuevos datos disponibles a lo largo del tiempo.

4.9.2 TRANSFORMACIÓN

Es la conversión de los datos extraídos en su formato anterior al formato que se requiere para que se puedan situar en un *Data Warehouse* o, simplemente, en otra base de datos. Esta fase se conoce como transformación, pero en realidad se realizan dos tareas: transformación y limpieza de datos.

El objetivo de esta fase (transformación y limpieza) es mejorar la calidad de los datos extraídos de las diferentes fuentes, mediante las correcciones de inconsistencias, imprecisiones y pérdida de valores. Las principales que se eliminan durante la fase de transformación y limpieza de datos son:

- Inconsistencias entre valores registrados con atributos diferentes que tienen el mismo significado.
- Duplicación de datos.

- Pérdida de datos.
- Existencia de valores inadmisibles.

4.9.3 CARGA

Una vez que los datos se han extraído, transformado y limpiado, se deben cargar en el *Data Warehouse* para hacerlos disponibles a los analistas y que puedan ser utilizados por las aplicaciones de apoyo a la decisión.

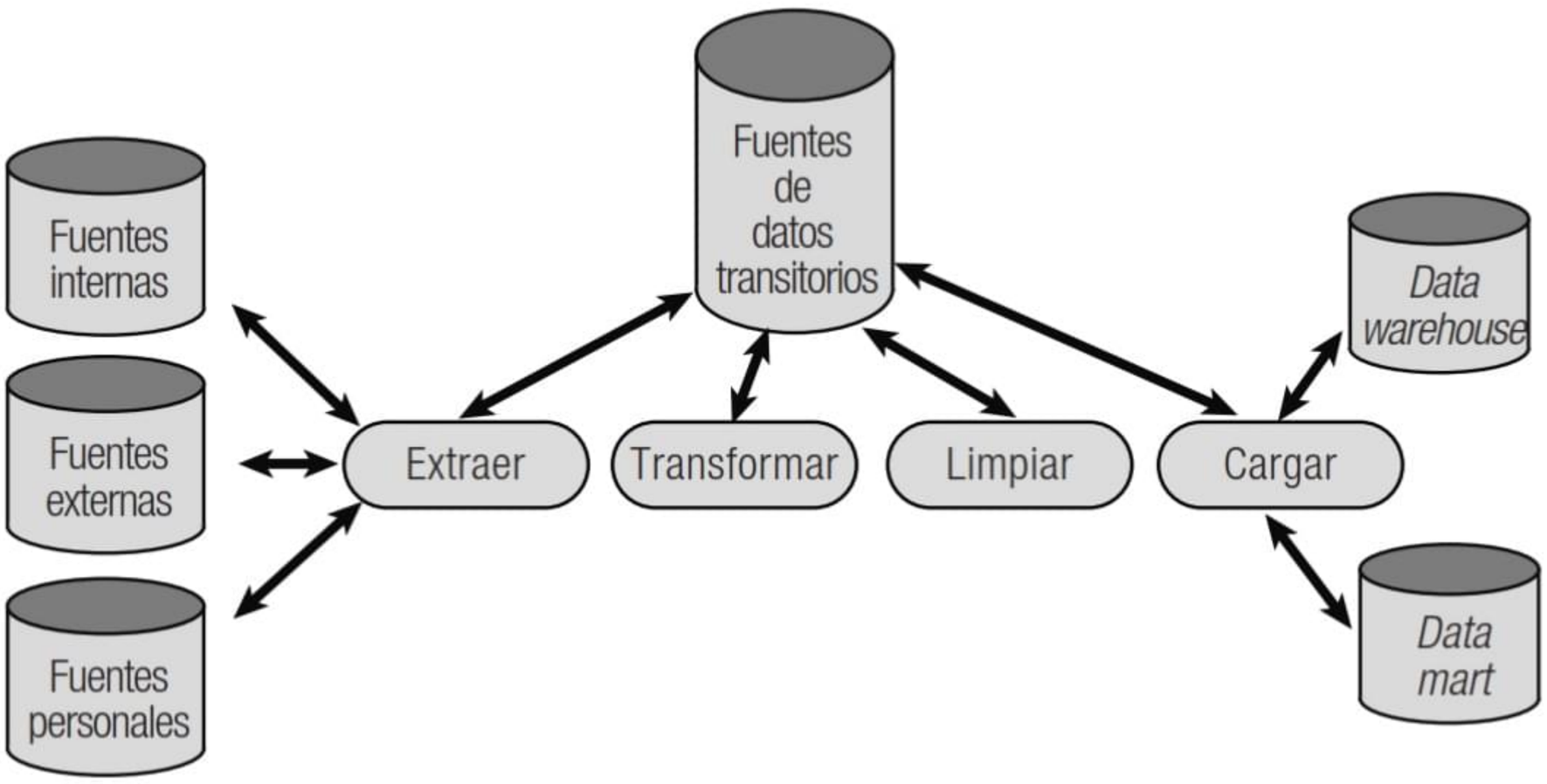


Figura 4.9. Fases de un sistema ETL.

4.10 DESAROLLO DE UN SISTEMA DE *DATA WAREHOUSE*

Un proyecto de *Data Warehouse* es una actividad muy importante dentro de cualquier organización y es más complicado que muchos de los restantes proyectos de computación, debido, fundamentalmente, a que afecta e influye en muchos departamentos e interfaces de entrada y salida. Además, son parte efectiva de los procesos de Inteligencia de Negocios de la compañía, desde los CRM de relación con los clientes hasta los sistemas de Analítica como OLAP o de análisis de datos, pasando por los sistemas de información básicos de la empresa, como los sistemas ERP.

Los beneficios de los sistemas de almacenes de datos para las organizaciones son numerosos y los clasificaremos en beneficios directos e indirectos (Turban, 2011):

- Los usuarios finales pueden realizar análisis amplios desde numerosas vistas.
- Es posible una visión consolidada de los datos corporativos.

Descargado en: eybooks.com

- La información es mucho mejor y oportuna.
- Se obtiene una mejora en el rendimiento o competencias del sistema.
- El acceso a los datos se simplifica.

Los beneficios indirectos se derivan del uso de los beneficios directos: mejora en el conocimiento del negocio, ventaja competitiva presente, mejora del servicio y de la atención al cliente, facilitación de la toma de decisiones y ayuda en las actualizaciones de los procesos de negocio. Por consiguiente, todos estos beneficios contribuyen a crear una ventaja competitiva con el uso del almacenamiento de datos (*Data Warehousing*).

Algunos autores añaden las siguientes características positivas de la utilización de un DW:

- Los usuarios finales pueden acceder a los datos que necesiten de un modo muy rápido y fácil, incluso vía navegadores web, ya que los datos estarán situados en servidores web.
- Los usuarios finales pueden realizar análisis amplios de los datos en formas que no eran posibles anteriormente con los sistemas tradicionales.

4.10.1 TIPOS DE DATA WAREHOUSE POR SU FUNCIONALIDAD

Vanden Hoven realizó una clasificación de los *Data Warehouse* que, pese a la antigüedad de su publicación, conserva en la actualidad toda su vigencia. Clasifica los almacenes de datos en tres categorías diferenciadas. La primera son los almacenes de datos que realizan funciones como localización, extracción, transformación, limpieza, transporte y carga de los datos en el Data Warehouse. La segunda categoría es como herramienta de gestión de datos: un motor de base de datos que almacena y gestiona los *Data Warehouse* así como los metadatos. La tercera categoría es una herramienta de acceso a datos que proporciona a los usuarios finales acceso para el análisis de los datos en el *Data Warehouse*. Esta última categoría incluye generadores de consultas, visualización, OLAP y Minería de Datos.

4.11 ENFOQUES DE DESARROLLO (MODELOS) DE UN SISTEMA DE *DATA WAREHOUSE*

Las organizaciones necesitan crear y desarrollar almacenes de datos que le ayuden de modo eficaz en sus procesos de toma de decisiones. Existen dos metodologías o enfoques para el desarrollo e implantación de un sistema de almacenamiento de datos, que, dependiendo del proveedor, se elegirá uno u otro. Son la metodología o desarrollo descendente (*top-down*) atribuida a Bill Inmon, conocido como padre del *Data Warehousing*, y la metodología o desarrollo

ascendente (*bottom-up*), atribuida a Ralph Kimball, otra de las grandes autoridades mundiales en almacenamiento de datos. Existen numerosos estudios comparativos de ambos métodos de desarrollo, por lo que haremos una breve descripción de cada uno.

4.11.1 MODELO INMON: MODELO EDW

El modelo Inmon soporta un enfoque de desarrollo descendente (*top-down*), que adapta las herramientas tradicionales de bases de datos relacionales a las necesidades de desarrollo de un *Data Warehouse* para la empresa o corporativo, conocido en inglés como EDW. Utiliza la metodología de desarrollo de las bases de datos establecidas en la empresa y herramientas como diagramas de entidad-relación, así como una adaptación del modelo de desarrollo en espiral. El enfoque EDW no impide la creación de *Data Mart* y suele ser ideal para el enfoque corporativo, ya que proporciona una visión consistente y completa de la empresa.

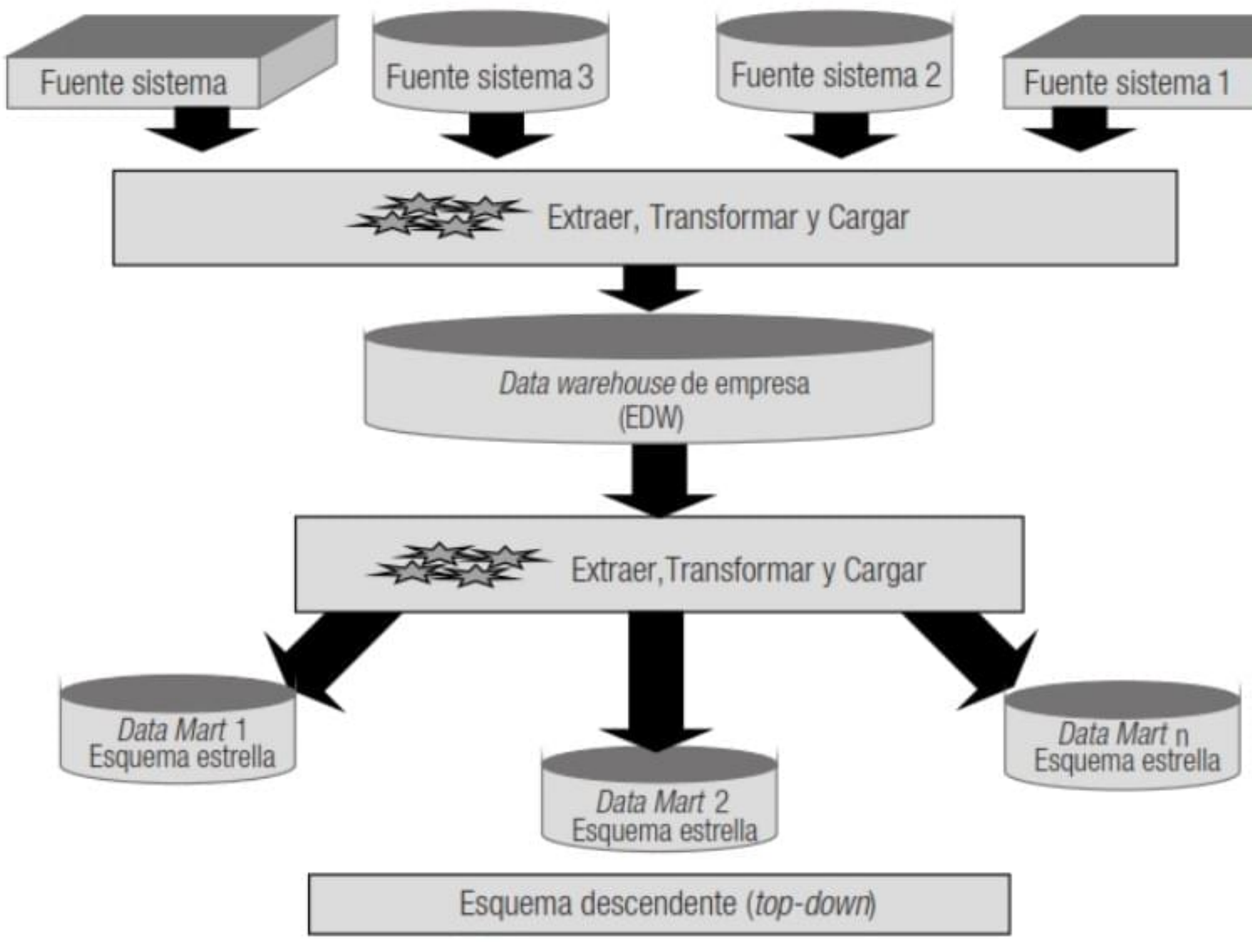


Figura 4.10. Modelo de *Data Warehouse* de Inmon.

4.11.2 MODELO KIMBALL: EL ENFOQUE DE *DATA MART*

El enfoque de Ralph Kimball es un modelo que propone un enfoque ascendente (*bottom-up*) que emplea el modelo dimensional conocido como enfoque de *Data Mart*. El enfoque de Kimball se centra en la expresión "*plan big, build small*" ("un gran plan, una construcción pequeña"). Recordemos que un *Data Mart* es un *Data Warehouse* orientado a un tema específico o a un departamento. En consecuencia, es una versión descendente de un *Data Warehouse*, que se centra en las peticiones y necesidades de un departamento específico como ventas, finanzas o contabilidad. Este modelo se aplica a modelado de datos

dimensionales que comienza con tablas y una metodología de desarrollo que implica un enfoque ascendente que, en el caso de un *Data Warehouse*, significa la construcción de un *Data Mart* en cada momento en que sea necesario.

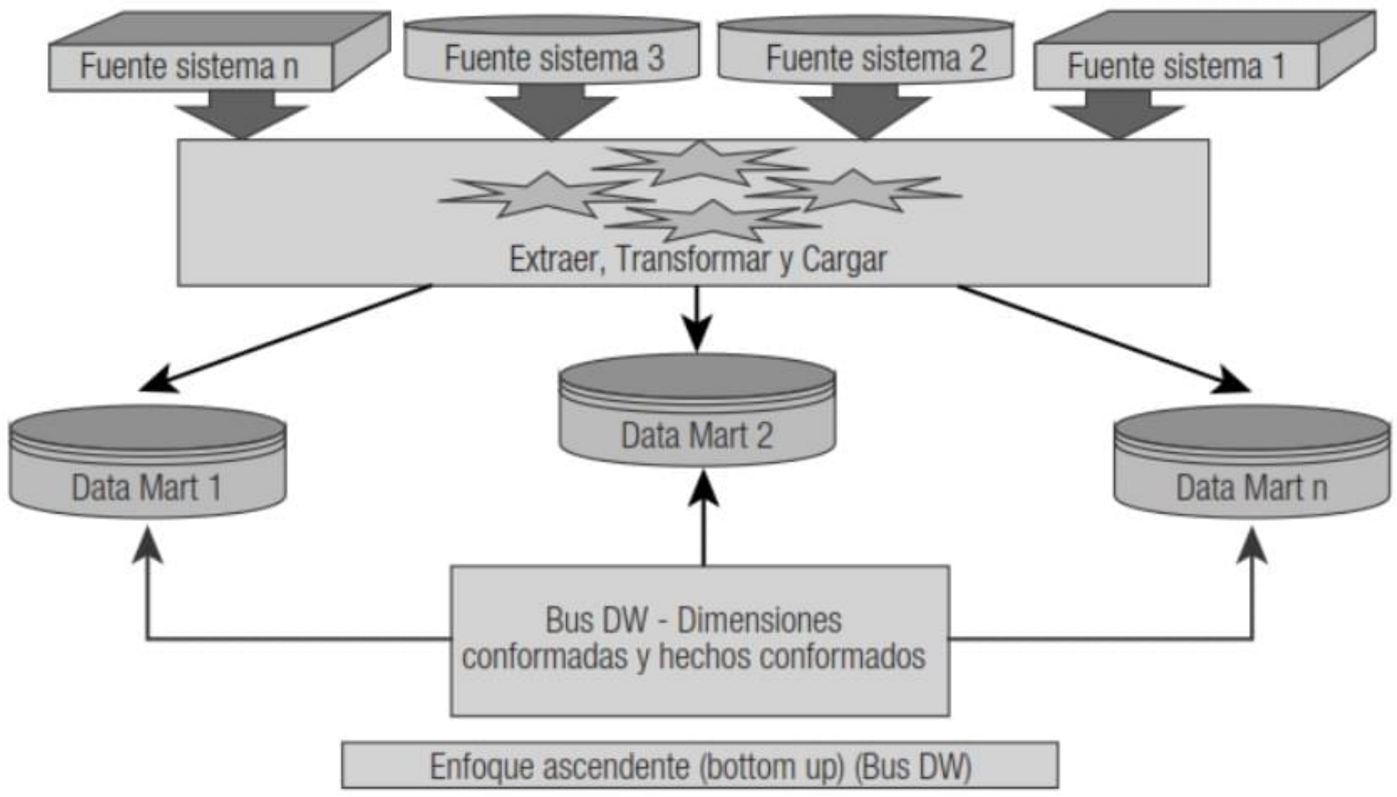


Figura 4.11. Modelo de *Data Warehouse* de Kimball.

4.11.3 IMPLANTACIÓN PRÁCTICA DE UN *DATA WAREHOUSE* EMPRESARIAL

No existe ninguna estrategia predeterminada para el desarrollo e implantación de un sistema de *Data Warehouse* empresarial. En muchos casos, vendrá impuesta por el proveedor y vendedor del software (los citados anteriormente); en otros, por el que la empresa haya seleccionado.

Una estrategia de almacenamiento de datos (*Data Warehousing*) puede evolucionar desde un sencillo *Data Mart* a un complejo *Data Warehouse* en respuesta a las demandas del usuario, los requerimientos de negocio de la empresa y la experiencia y madurez de la empresa en la gestión de recursos de datos. En muchas empresas, se utiliza la implantación previa de un *Data Mart* como experiencia previa en la construcción y gestión de datos, a la par que va presentando a los usuarios del negocio los beneficios de un mejor acceso a los datos a través del *Data Mart*. Según la compañía, el desarrollo de *Data Mart* independientes puede suponer grandes ventajas desde el punto de vista del tiempo, ahorro de costos y facilidad de acceso ante los posibles problemas de implantación de un sistema de almacenamiento global basado en un gran *Data Warehouse* a gran escala.

4.11.4 *DATA WAREHOUSE* 2.0 Y *DATA WAREHOUSE* EN TIEMPO REAL

Bill Inmon ha creado la definición más aceptada de *Data Warehouse*: "un *Data Warehouse* es un conjunto de datos integrados orientados a un tema, que varían con el tiempo y que no son transitorios, los cuales soportan el proceso de toma

de decisiones en la gestión". Esta definición se centraba en datos estructurados. Con el paso del tiempo, y como ya conoce el lector, el número de datos no estructurados que manejan las organizaciones y empresas ha crecido en grandes volúmenes y llega al 80 ó 90% del total que gestionan las compañías. Bill Inmon, atento a estas circunstancias, publicó un nuevo libro de *Data Warehouse*, en 2008, que denominó *Data Warehouse* 2.0 (Inmon 2008) donde se adelantó a la expansión de los grandes volúmenes de datos y, sobre todo, de datos no estructurados que se produjo al principio de esta década. En el libro, plantea el procesamiento de ambos tipos de datos: estructurados y no estructurados. Inmon define *Data Warehouse* 2.0 como la segunda generación de *Data Warehousing* con las características principales de "inclusión de datos no estructurados, metadatos en los negocios, procesamiento de exploración y estadística de datos, seguridad, granularidad y desempeño o rendimiento de los sistemas".

En paralelo, con el advenimiento de esta segunda generación de almacenamiento de datos, ha surgido una nueva tendencia denominada *Data Warehouse* en tiempo real, que sería el proceso de carga y de provisión de datos, vía los almacenes de datos, a medida que están disponibles. Estas nuevas herramientas pretenden responder a las transacciones más inmediatas que se producen en tiempo real o, al menos, a la mayor velocidad posible.

De esta forma, los *Data Warehouse* corporativos actuales ofrecen soluciones para manejar datos estructurados y no estructurados, así como el procesamiento de dichos datos en tiempo real, o al menos a la mayor velocidad posible, en lugar de las actualizaciones periódicas que se producen en los sistemas de almacenes de datos tradicionales con periodicidad de días o semanas.

4.12 OLAP (PROCESAMIENTO ANALÍTICO EN LÍNEA)

El análisis de datos en los sistemas de Inteligencia de Negocios se realiza fundamentalmente con herramientas de análisis multidimensional (también denominado OLAP), Minería de Datos y sistema de apoyo a la decisión. El creador de la idea OLAP fue Edgar Frank Codd, conocido precursor de las bases de datos relacionales. La propuesta inicial consistió en realizar una disposición de los datos en vectores para permitir un análisis rápido. Estos vectores son los llamados cubos OLAP, muy adecuados para el análisis instantáneo de grandes cantidades de datos.

Un ejemplo simple para ver la aplicación práctica de OLAP puede ser el siguiente: Una empresa comercializa cuatro productos diferentes, por ejemplo, aparatos de Tv, frigoríficos, lavadoras y calefactores, en las regiones del Norte, Sur, Este, Oeste y en la costa (por diferenciar específicamente el sector costero, con independencia de su situación geográfica). Las bases de datos tradicionales responden muy bien a consultas tales como: ¿cuántas ventas se han realizado el último trimestre del año en la costa? Sin embargo, si se desea conocer cuántas

lavadoras se vendieron en cada una de las regiones de venta y comparar sus resultados con sus previsiones de ventas, las bases de datos relacionales resuelven el problema de un modo más complejo que los sistemas OLAP que, en este caso concreto, son más adecuados.

OLAP (*On Line Analytical Procesing*) es una aplicación conocida también como análisis multidimensional (procesamiento analítico en línea). OLAP es una base de datos multidimensional donde el almacenamiento físico de los datos se realiza en un vector unidimensional. Los cubos OLAP se suelen comparar, a veces, con una hoja de cálculo ampliada. Soporta análisis de datos multidimensional y facilita a los usuarios su visión en diferentes formas, utilizando múltiples dimensiones. Cada aspecto o indicador de una información (producto, precio, coste, región, período) representa una dimensión diferente.

La razón de utilizar OLAP para consultas es la rapidez de respuesta en el caso de múltiples dimensiones. Las bases de datos relacionales son idóneas para registrar datos provenientes de las transacciones ordinarias (sistemas OLTP, procesamiento de transacciones en línea), ya que almacenan los datos en tablas discretas que han sido normalizadas. Esta estructura es idónea para los OLTP, como ya se ha comentado, pero para consultas complejas tipo multitabla se vuelve relativamente lenta; en este caso, el modelo más adecuado es una base de datos multidimensional o cubo de datos OLAP. Algunas características de los sistemas OLAP son:

- El análisis multidimensional permite a los usuarios visualizar los mismos datos desde diferentes sitios, utilizando múltiples dimensiones. Cada característica de la información de una aplicación de una empresa (productos, precios, artículos vendidos, año, región) se representa por una dimensión diferente.
- Repuestas rápidas a problemas ad hoc.
- Alto nivel de detalle en cada operación.

Los elementos de control son:

- Variables de decisión: representan una medición del negocio.
- Se basan en el concepto de cubo de datos OLAP.
- Las herramientas OLAP se caracterizan por:
- Ofrecer al usuario una visión multidimensional de los datos.
- No imponer restricciones sobre el número de dimensiones.
- Ofrecer simetría para las dimensiones.
- Permitir definir de forma flexible (sin limitaciones) sobre las dimensiones: restricciones, agregaciones y jerarquías entre ellas.
- Ofrecer operadores intuitivos de manipulación (*drill-down, roll-up, slice-and-dice, pivot,* etcétera).

- Ser transparentes al tipo de tecnologías que soporta el *Data Warehouse* ROLAP o MOLAP.
- Permitir consultas a la herramienta OLAP, seleccionando atributos en este esquema multidimensional si se conoce la estructura interna (esquema físico) del *Data Warehouse* caducado.
- Generar la correspondiente consulta y enviarla al gestor de consultas del sistema (mediante una sentencia SELECT).

4.12.1 CUBOS OLAP

Para funcionar, las aplicaciones OLAP utilizan un tipo de base de datos que posee la peculiaridad de ser multidimensional, denominada comúnmente cubo OLAP. Un cubo OLAP es una base de datos que posee varias dimensiones, que amplía las posibilidades que hasta el momento ofrecían las conocidas hojas de cálculo. Mediante la incorporación de estos vectores o cubos, se han ampliado las posibilidades de las bases de datos relacionales, permitiendo el procesamiento de importantes volúmenes de información que, de lo contrario, serían imposibles de realizar. Cada una de las dimensiones que posee la base de datos incorpora un campo determinado para un tipo de dato específico, que luego podrá ser comparado con la información contenida en el resto de las dimensiones para hacer posible la evaluación de la información realmente relevante para una compañía.

Una base de datos multidimensional puede contener varios cubos o vectores, que extenderán las posibilidades del sistema OLAP con el cual se trabaja, conocidos como hipercubos. En general, los sistemas OLAP suelen estar compuestos por tres dimensiones (figuras 4.11 a y b), pero pueden albergar más de tres dimensiones, mediante la utilización de estos cubos OLAP, donde cada una de las dimensiones o escalas del cubo corresponde básicamente a una jerarquía de datos.

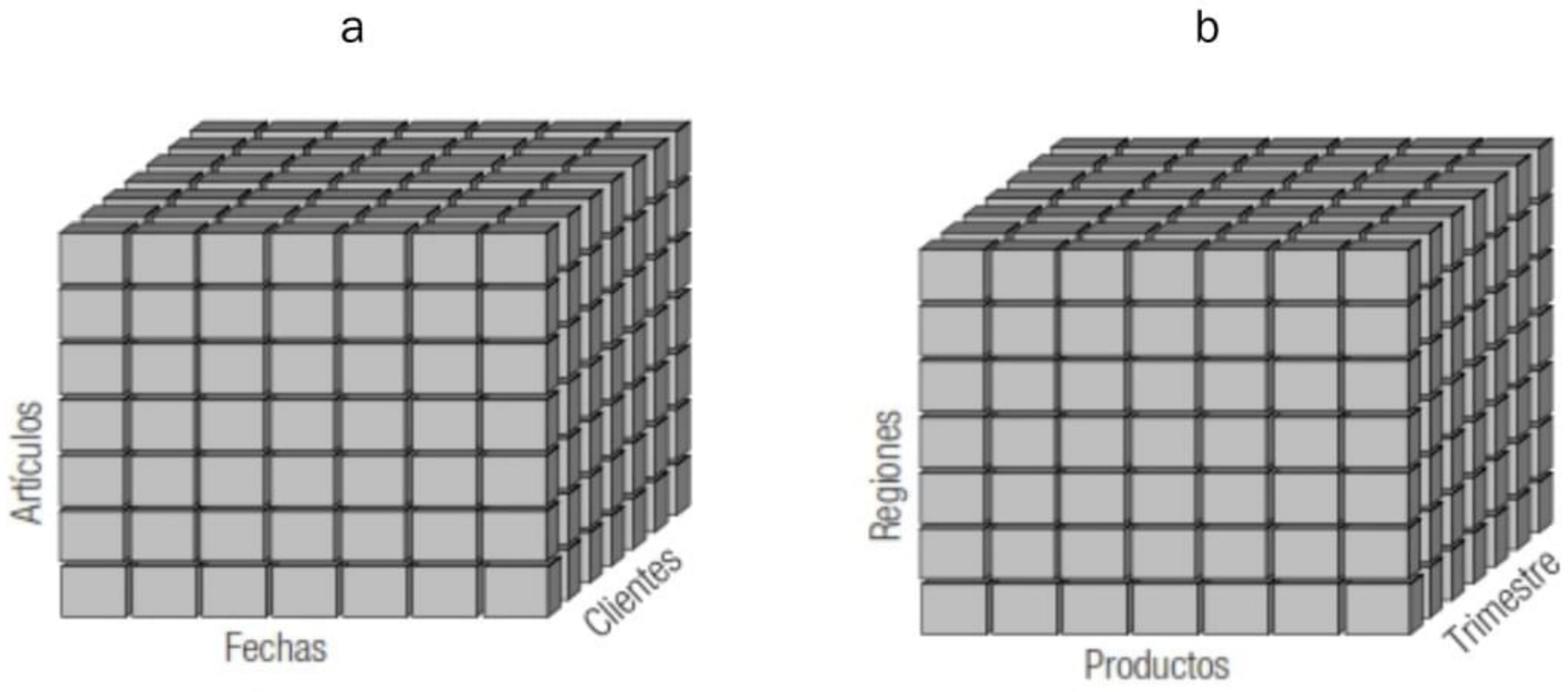


Figura 4.12. Cubos OLAP básicos de tres dimensiones, (a) y (b).

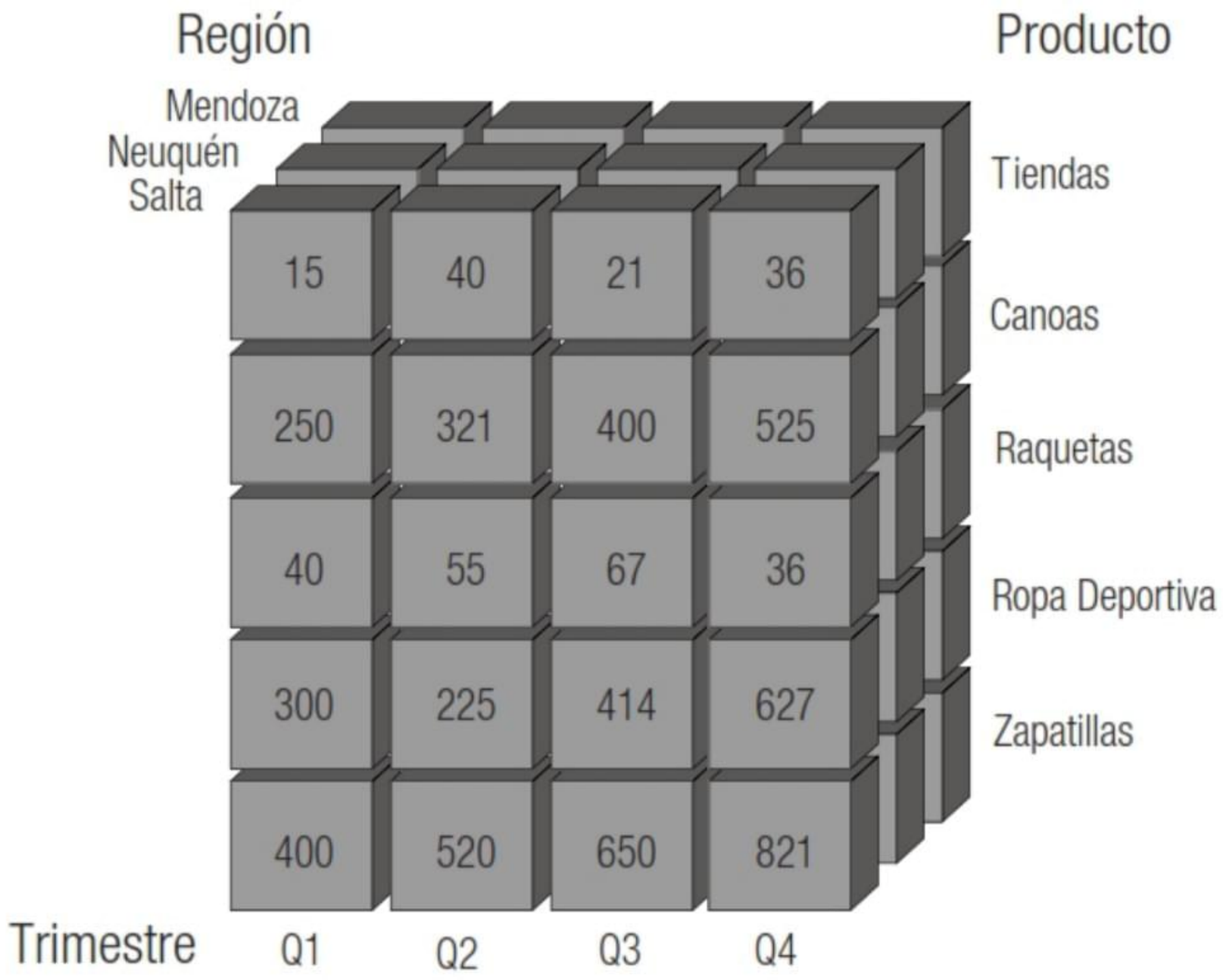


Figura 4.13. Cubo OLAP con las dimensiones generales: Trimestre, Región, Producto

Las herramientas OLAP se clasifican en tres grandes grupos básicos: MOLAP, ROLAP y HOLAP; también, han ido surgiendo otras herramientas OLAP, como el caso de DOLAP (sistemas OLAP de escritorio).

4.12.2 MOLAP

MOLAP, la arquitectura de OLAP, se implementa en una base de datos multidimensional (para ser visualizada en múltiples dimensiones), donde se almacenan los datos para proporcionar su análisis. Los datos se organizan en una estructura tipo cubo que el usuario puede rotar. Es muy adecuado para resúmenes e informes financieros.

4.12.3 ROLAP

Es un sistema OLAP relacional. Se implementa con una base de datos relacional para el manejo, acceso y obtención de datos, pero disponen de algunas facilidades para mejorar el rendimiento. El sistema accede directamente a los datos almacenados en un *Data Warehouse* para proporcionar los análisis

especificados. Puede crear vistas multidimensionales, pero no de estructura cubo de datos.

La base de datos relacional maneja los requerimientos de *Data Warehouse* y el motor de ROLAP proporciona la función analítica. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.

4.12.4 HOLAP (OLAP HÍBRIDA)

Almacena algunos datos en un motor relacional, y otros, en una base de datos multidimensional.

Trata de combinar las ventajas de MOLAP y ROLAP. Por ejemplo, cuando se necesita información tipo resúmenes, HOLAP potencia la tecnología de cubos para el desempeño más rápido. Cuando se necesita una información detallada, HOLAP se apoya en los datos relacionales, mediante técnicas *drill through*, que van desde los cubos hasta las capas de datos relacionales.

La tabla 4.1 muestra las características fundamentales de las cuatro categorías de OLAP.

Rendimiento	**MOLAP** Base datos multidimensional	Los datos que subyacen en los hipercubos son almacenados junto con las agregaciones en una estructura multidimensional.
Capacidad	**ROLAP** Base datos Relacional	Los datos que subyacen en los hipercubos son almacenados junto con las agregaciones en una estructura relacional.
	HOLAP Sistema híbrido	Los datos que subyacen en los hipercubos son almacenados en una estructura relacional y las agregaciones en una estructura multidimensional.
	DOLAP	Instalación MOLAP en un equipo cliente.

Tabla 4.1. Categorías de OLAP

4.12.5 PROVEEDORES DE OLAP

- OLAP MiCROSOFT (`http://office.microsoft.com/es-es/excel-help/informacion-general-sobre-el-pro-cesamiento-analitico-en-linea-olap-HP010177437.aspx`)
- Business Object (`businessobject.com`)
- QlikView de Qliktech (`qlikview.com/es`)

Descargado en: eybooks.com

- MicroStrategy OLAP
- Cognos OLAP
- IBM OLAP
- Information Builders
- Oracle
- SAS institute
- IBM SPSS

4.13 *DATA LAKE* (LAGOS DE DATOS): LOS NUEVOS DEPÓSITOS DE ALMACENAMIENTO DE DATOS

En la actualidad los repositorios de información de la empresa se dividen en dos grandes grupos:

1. Los almacenes de datos (*Data Warehouse* y *Data Mart*) soportados por bases de datos relacionales, que soportan datos estructurados organizados en filas y columnas (tablas).
2. Sistemas de Big Data que soportan grandes volúmenes de datos estructurados, no estructurados y semiestructurados, basados esencialmente en marcos de trabajo Hadoop y, cada vez con mayor frecuencia, Spark.

Un almacén de datos (*Data Warehouse*) es el primer paso natural para almacenar todo tipo de datos de un proyecto Big Data, pero su efectividad se reduce cuando el proyecto madura y los datos comienzan a crecer. Todos los datos son estructurados y procesados de una determinada forma, reduciendo su agilidad y coste al gestionar grandes volúmenes de datos.

Con el objeto de resolver los problemas típicos de un almacén de datos, cuando los datos son de gran volumen y no son estructurados, nacieron las infraestructuras de Big Data con el sistema Hadoop y las bases de datos NoSQL y "en memoria" como soluciones eficientes y fiables. Sin embargo, estos sistemas pensados para la recopilación de grandes volúmenes de datos y de diferentes formatos (estructurados, no estructurados y semiestructurados), así como su creación a grandes velocidades, exige procesos de ETL de extracción, transformación y carga de los datos en bruto (raw data) para su almacenamiento y posterior análisis con las herramientas adecuadas.

Sin embargo, las empresas, en sus estrategias organizativas, pueden decidir conservar todos los datos capturados que generan sus diferentes fuentes de información y sólo utilizarlos cuando se necesiten. Para tratar de resolver estos problemas, surgieron los *Data Lake* (Lagos de Datos). Un **Lago de Datos** resuelve

los problemas anteriores, ya que son sistemas de almacenamiento de datos sin procesamiento previo de los datos, pero que exigen soluciones técnicas muy sofisticadas y, normalmente, perfiles profesionales especializados que sean capaces de gestionar esos datos.

Origen del término *Data Lake*

El origen del término se le atribuye a James Dixon[1], CTO de Pentaho (uno de los proveedores más populares de software libre para herramientas de Big Data, fundamentalmente Hadoop), que en el blog oficial de la compañía, y presentando sus soluciones de Big Data, introdujo (en un artículo publicado en octubre de 2010) el término *Data Lake* para referirse a las mejoras introducidas por Pentaho en los sistemas de Big Data.

Dixon reflexionaba sobre el trabajo realizado para lanzar su primera distribución de Hadoop y lo que había aprendido de la interacción durante ese período con diferentes empresas usuarias de Hadoop. Analizando los factores comunes que había visto en dichos casos, extrajo la conclusión de que la gran mayoría de dichas empresas manejaban datos estructurados o semiestructurados provenientes típicamente de una única aplicación o sistema, en un volumen que hacía inviable técnica o económicamente el uso de un sistema de base de datos relacional para almacenarlos y que, aunque algunas de las preguntas que se querían hacer sobre los datos eran conocidas de antemano, muchas no lo eran e irían surgiendo en el futuro.

Partiendo de estas conclusiones, Dixon propuso una visión de cómo deberían almacenarse este tipo de datos, sin tener que pasar por un preprocesamiento que extrajera los atributos que, en principio, se pensase que iban a ser utilizados en las consultas y sin tener que agregarlos necesariamente, para así no perder capacidad de ser usados a posteriori según otras necesidades. De ahí surge la analogía del "lago de datos", donde los datos (el agua del lago, según la analogía) provienen de un origen y "llenan el lago" sin haber sido aún limpiados, procesados o empaquetados a priori. De esa manera, los datos se almacenarán en crudo y en su totalidad, y serán las necesidades de los diferentes grupos de usuarios las que permitirán identificar las versiones filtradas o procesadas de dichos datos que harán falta para resolver dichas necesidades (bien sea en forma de consultas preprocesadas por ser las más frecuentes, o como entrada para un *Data Warehouse* con una estructura de datos determinada para agregarla a otras fuentes)

Definición de *Data Lake*

Un *Data Lake* (Lago de Datos) es un repositorio de almacenamiento que contiene una gran cantidad de datos en bruto en su formato original, incluyendo datos estructurados, semiestructurados y no estructurados, que se guardan sin ningún procesamiento (raw data). Los datos se guardan en bruto y sin ningún tipo de esquema; su estructura y los requisitos de los datos no se definen hasta que

éstos se necesitan. En resumen, en un *Data Lake* se recogen todos los datos (en forma directa o formato nativo) pero no se alteran, limpian o manipulan; su valor se mantiene en bruto (nativo) y no se transforma previamente hasta su análisis y explotación. Los datos se vierten por las organizaciones y se recuperan cuando es necesario y, en ese instante, se procede a ordenarlos y dar una estructura que permite análisis posteriores.

Un *Data Lake* tiene una arquitectura plana para almacenamiento de datos. Los datos (estructurados y no estructurados) se mantienen en el formato en que fueron ingeridos hasta que sean solicitados para su consulta. Para el procesamiento de los datos, a cada dato se le asigna una etiqueta o identificador único y se organizan como un conjunto de etiquetas de metadatos extendidas; con este sistema de almacenamiento cuando se invoca un dato o una consulta, se asocia a su identificador y, en consecuencia, se puede realizar el proceso en el lago de datos seleccionando un conjunto de datos más pequeños asociados a sus etiquetas y se realiza el proceso correspondiente. El lago de datos se asocia, normalmente, con el almacenamiento de objetos orientado a Hadoop como se comenta a continuación.

Un Lago de Datos es un depósito de datos masivo y de fácil acceso para almacenar Big Data. Hadoop es la tecnología más utilizada para crear Lagos de Datos. En esencia, un *Data Lake* es un tipo de almacenamiento en el que la información almacenada tiene una estructura variable (diferentes tipos de datos, texto, imágenes, mensajes, audios, ubicaciones físicas, etcétera), es masiva, de fácil y rápido acceso y resiliencia, sin atender a una lógica de negocio específica. Los Lagos de Datos almacenan los datos en su formato más básico (en bruto), se actualizan añadiendo más información pero nunca se modifica la información ya existente. De este modo, los *Data Lake* permiten almacenar los datos en bruto y estar disponibles en todo momento y, casi, en tiempo real en su formato original. El uso de fuentes muy variadas permite realizar análisis complejos y modelos predictivos.

Un *Data Lake* no es un *Data Warehouse* 4.0 ni tampoco un sustituto. Ambos sistemas conviven y cada uno está diseñado para un determinado objetivo, por lo que los directivos de TI deben ver cuál es la mejor herramienta para su trabajo. En cualquier forma, existen consultoras y proveedores de soluciones de software que definen a los *Data Lake* como *Data Warehouse* mejorados o modernos.

En realidad, un *Data Lake* es un almacén de datos en el que se guardan todos los datos que necesita la organización sin ningún tratamiento ni estructura; son almacenes de datos en crudo (archivos, logs, tablas de cualquier sistema ERP, imágenes de Facebook, tuits o cookies de los navegadores de usuarios). En esencia, y dado que permite el almacenamiento de datos no estructurados, un Lago de Datos es un repositorio de datos Hadoop. Desde el punto de vista técnico, un Lago de Datos tiene una arquitectura plana de almacenamiento; a cada elemento dato de un lago se le asigna un identificador único y se le etiqueta con un conjunto de etiquetas de metadatos ampliados. Cuando se realiza una

consulta al Lago de Datos en busca de datos, se analiza el conjunto de datos para responder a dicha consulta.

En resumen, *Data Lake* ha sido definido como la futura generación de sistemas de almacenamiento de datos. El *Data Lake* no sustituye ni elimina los almacenes de datos (*Data Warehouse*), sino que amplía las capacidades de la organización en lo relativo a la gestión de datos voluminosos y variables.

Los *Data Lake* se pueden desarrollar con múltiples tecnologías, tales como Hadoop, NoSQL, bases de datos relacionales, etcétera.

En un *Data Lake*, al disponer de todos los datos en bruto (raw data), el acceso a los datos originales es más directo y reduce los pasos intermedios necesarios (ETL) para su procesamiento (evitando la necesidad de cargar modelos previos), con una estructura de datos no definida hasta que los datos son necesarios.

4.13.1 PROPIEDADES, VENTAJAS E INCONVENIENTES DE LOS *DATA LAKE*

Capacidades del Lago de Datos

1. Capturar y almacenar datos primarios a escala para reducir costes, sin procesamiento (en bruto).
2. Soporte para almacenar todos los tipos de datos existentes en el mismo depósito y sin necesidad de procesarlos.
3. Soporte para todo tipo de perfiles de usuarios, tanto para modelos empresariales profesionales como para modelos de investigación.
4. Mayor facilidad para cambiar y actualizar el sistema de datos utilizado.
5. Definir la estructura de los datos en el momento en que se utilicen.

Ventajas del Lago de Datos

- El Lago de Datos permite que los usuarios comerciales **tengan acceso inmediato a todo el almacén de datos.**
- Los datos situados en el lago **no se limitan** a los datos relacionales o transaccionales.
- Con un Lago de Datos, el usuario **nunca necesita desplazar los datos.**
- El Lago de Datos **otorga facultades a los usuarios comerciales** y los libera de las ataduras que supone la dominación de TI.
- El Lago de Datos acelera la entrega, permitiendo que las unidades de negocio alimenten las aplicaciones rápidamente.

Desventajas del Lago de Datos

- Área desconocida del procesamiento de datos.

- Gobernanza de datos.
- Gestionar el caos.
- Problemas de privacidad.
- Complejidad de los datos heredados.
- Gestión del ciclo de vida de los metadatos.
- Islas de datos aislados.

4.13.2 PROVEEDORES DE SOLUCIONES DE *DATA LAKE*

Los *Data Lake* se construyen normalmente en Hadoop y algunos proveedores actualizan las versiones de los almacenes de datos cuando sólo se necesitan datos estructurados. Por estas razones, las distribuciones comerciales de Hadoop y las proveedoras de soluciones de software de Big Data son las que ofrecen mejores soluciones de arquitecturas de *Data Lake*.

Las proveedoras de Big Data Hortonworks y MapR ofrecen excelentes y económicas versiones de arquitectura de Lagos de Datos. Empresas de software propietario que ofrecen soluciones de almacenamiento de *Data Lake* son: Informática con el producto Intelligent *Data Lake*, Microsoft con Azure *Data Lake*, EMC, Amazon con sus productos AWS y EC2.

4.14 *DATA WAREHOUSE* VERSUS *DATA LAKE*

Andy Patrizio[2],un prestigioso periodista de la revista Datamation, publicó a mediados de 2017 un excelente artículo muy ilustrativo sobre las diferencias entre *Data Lake* y *Data Warehouse*:

¿Qué es un *Data Lake*? Un Lago de Datos es un repositorio de almacenamiento que contiene una gran cantidad de datos en bruto, en su formato nativo, y se almacenan sin procesar hasta que se necesiten. Un Lago de Datos utiliza una arquitectura plana para almacenar datos, que son totalmente no estructurados y se dejan en el formato original en el que fueron recolectados, aunque a cada elemento del lago se le asigna un único identificador y se etiqueta con un conjunto amplio de etiquetas de metadatos. De este modo, cuando se ejecuta una consulta, se puede realizar frente a un pequeño conjunto de datos con las etiquetas específicas en lugar del procesamiento de todos los datos del lago.

¿Qué es un *Data Warehouse*? Es el término opuesto a Lago de Datos. Es un repositorio de datos estructurados, integrados, de múltiples fuentes, organizados para la recreación de informes analíticos. Normalmente, utilizan múltiples bases de datos para diferentes tipos de almacenamiento de datos, tales como

ingestión, *staging* y transformación, y procesamiento, tales como procesamiento analítico en línea (OLAP) o procesamiento de transacciones en línea.

Ambos términos se utilizan para almacenamiento de datos, pero adoptan diferentes enfoques. El Almacén de Datos se adapta a una estructura bien definida, mientras que el Lago de Datos es más flexible. El Lago de Datos contiene datos estructurados, no estructurados y semiestructurados y, por consiguiente, no tiene requerimientos de estructuras de datos rígidas. En ocasiones, se considera que un *Data Warehouse* es como un subconjunto de un ecosistema de grandes datos. Ambos pueden coexistir y funcionar juntos de modo eficiente y ambos aportan conocimiento y profundidad de ideas a los negocios. Se consideran estas dos estructuras de datos como los complementos ideales de Big Data.

Por lo anterior, hay que considerar que un *Data Lake* no es un *Data Warehouse*. Como el término está de moda, muchas veces los empresarios quieren estar a la vanguardia y pretenden usar esta tecnología cuando no es lo que requieren y comprometen el actuar de sus áreas de TI. Otra fuente de referencia que señala las principales diferencias entre ambos términos, es la reseña de Tamara Dull (2016)[3]en un artículo publicado en el reputado y mencionado portal KDnuggets:

Datos. Un *Data Warehouse* sólo almacena datos que han sido modelados/estructurados, mientras que un *Data Lake* no tiene formato de datos y almacena todo tipo de datos (estructurados, no estructurados y semiestructurados).

Procesamiento. Antes de que se puedan cargar datos en un *Data Warehouse*, primero hay que darles formato y estructura, es decir, hay que modelarlos. Esta operación se llama esquema de escritura. Un *Data Lake* contiene los datos en bruto, tal y como son; posteriormente, cuando se requiera usar los datos, es cuando se les da formato y estructura. Esta operación se llama esquema de lectura. En este sentido, se trata de dos enfoques muy diferentes.

Almacenamiento. Una de las características principales de las tecnologías de datos grandes como Hadoop es que el costo de almacenamiento de datos es relativamente bajo en comparación con el *Data Warehouse*. Hay dos razones principales para esto: en primer lugar, Hadoop es un software de código abierto, por lo que el soporte de licencias y la comunidad son gratuitos. Y, en segundo lugar, Hadoop está diseñado para ser instalado en hardware de bajo costo.

Agilidad. Un *Data Warehouse* es un repositorio altamente estructurado por definición. No es técnicamente difícil cambiar la estructura, pero puede ser que implique modificar todos o gran parte de los procesos de negocios que están vinculados a la misma. Un *Data Lake*, por el contrario, carece de la estructura de un *Data Warehouse*, lo que ofrece a los desarrolladores y científicos de datos la posibilidad de configurar y reconfigurar sus modelos, consultas y aplicaciones sobre la marcha con facilidad.

Seguridad. Las tecnologías de *Data Warehouse* han existido durante décadas, mientras que las tecnologías de Big Data (por ejemplo, las bases de datos de un *Data Lake*) son relativamente nuevas. Por lo tanto, la capacidad de asegurar los datos en un *Data Warehouse* es mucho más madura que la seguridad de datos en un *Data Lake*, aunque la seguridad en Big Data cada día es mayor.

Usuarios. En el estado de desarrollo en que se encuentran los Lagos de Datos, estos sistemas de almacenamiento están más pensados para expertos en desarrollo y análisis de grandes volúmenes de datos (analistas, ingenieros y científicos de datos).

DATA WAREHOUSE (Almacén de Datos)	Vs.	*DATA LAKE* (Lago de Datos)
Procesado, estructurado	DATOS	estructurados/semiestructurados/ no estructurados, (en bruto)
Esquema de escritura	PROCESAMIENTO	esquema de lectura
Caro para grandes volúmenes de datos	ALMACENAMIENTO	diseñado para bajo costo
Configuración fija, menos ágil	AGILIDAD	Sistema en experimentación (no maduro)
Sistema experimentado (maduro)	SEGURIDAD	muy ágil, se configura y reconfigura a medida que se necesite
Profesionales de negocios	USUARIOS	Preferentemente analistas y científicos de datos

Tabla 4.2. Características de un *Data Warehouse* y un *Data Lake*[4].

Las principales diferencias entre un *Data Warehouse* y un *Data Lake* son:

- Estructura de los datos recolectados. Un *Data Warehouse* almacena los datos que han sido estructurados, mientras que un *Data Lake* no utiliza ninguna estructura. En el almacén tradicional los datos son sometidos a procesos ETL para asegurar que adquieren el formato idóneo y están estructurados de la forma adecuada antes de quedar guardados. En el *Data Lake* este paso no es necesario. Se gana tiempo (al reducirse considerablemente el tiempo de carga) y se optimiza el aprovechamiento de recursos ya que, en función de las necesidades del usuario final, se estructurarán de una forma u otra.
- Procesamiento. Los datos se procesan antes de ser cargados en un *Data Warehouse* para darle algún tipo de modelo. Los datos en un *Data Lake* son

datos en bruto y así se conservan hasta que se necesiten y sean procesados.

- Modo de almacenamiento. Los requisitos de procesamiento previo y el rendimiento analítico también son diferentes El *Data Warehouse* se basa en la tecnología de base de datos relacional, que sólo puede almacenar datos coherentes y estructurados, mientras que el *Data Lake* se basa en tecnologías que permiten almacenar datos sin procesar y luego aplicar incrementalmente la estructura, según los requisitos analíticos. Los *Data Warehouse* tienden a ser grandes bases de datos, mientras que los Lagos de Datos corren sobre Hadoop, que está diseñado para bajo coste y scaleout hardware.
- Reconfiguración. Un Almacén de datos es un repositorio muy estructurado, de modo que no responde bien a los cambios; se pueden realizar pero se tarda mucho tiempo. Un Lago de Datos carece de cualquier tipo de estructura, de modo que se puede configurar y reconfigurar a medida que se necesiten los cambios.

4.15 PROVEEDORES DE SOLUCIONES DE *DATA WAREHOUSE*

Existen un gran número de proveedores de soluciones de *Data Warehouse*: los proveedores tradicionales de bases de datos relacionales que han diversificado sus soluciones de software y proveedores especializados en estos sistemas de almacenamiento de datos. De igual forma que con otras soluciones de Inteligencia de Negocios, existen proveedores de software propietario y proveedores de código abierto y/o gratuitos.

CASO DE ESTUDIO DE PROVEEDORES DE *DATA WAREHOUSE*

Existe un gran número de proveedores de soluciones de *Data Warehouse* muy cualificados y con ofertas muy variadas, que normalmente se ajustarán a los requerimientos de la empresa. Los propios portales web de los proveedores y sitios especializados (recursos del capítulo) le proporcionarán las funcionalidades y características de las soluciones, y grandes consultoras ofrecen estudios pormenorizados sobre clasificación de las herramientas comerciales de los diversos proveedores. Este es el caso del estudio de la consultora Gartner, Cuadrante mágico de proveedores de *Data Warehouse*, que todos los años publica, con un estudio comparativo de los mejores vendedores de soluciones y sus características fundamentales, y que es referencia obligada para profesionales de TI y directivos de empresa.

Normalmente, los proveedores de DW suelen ofrecer otras herramientas de Inteligencia de Negocios y sistemas de información, por lo que, en muchos casos, le ofrecerán aplicaciones para integración de solución. Una lista de los

proveedores más reconocidos, a expensas de luego analizar el Cuadrante Mágico de Gartner 2017 de DW, es:

- Computer Associates (www.ca.com).
- Embarcadero (www.embarcadero.com).
- Business Object (www.businessobject.com).Es en la actualidad propiedad de SAP.
- HP (hp.com).
- Hyperion Solutions (www.hyperion.com).
- IBM (www.ibm.com).
- informatica (www.informatica.com).
- Microsoft (www.microsoft.com).
- Oracle (www.oracle.com).
- SAS Institute (www.sas.com).
- Siemens (www.siemens.com).
- Sybase (www.sybase.com).
- Teradata (www.teradata.com).

Cuadrante Mágico de Gartner de soluciones de *Data Warehousing* y *Analytics*: *Management Solutions for Analytics*

Uno de los numerosos informes que publica Gartner es relativo a soluciones de almacenamiento de datos y se llama Cuadrante Mágico de Gartner de *Data Management Solutions for Analytics*" **DSMA**" (Soluciones Analíticas de Gestión de Datos o Soluciones de Gestión de Datos para Analítica). A partir del año 2017introdujo este nuevo título, ya que en los años anteriores se llamaba *Magic Quadrant for Data Warehouse and Database Management Systems.* Con este cambio al igual que ya hemos visto con otros informes, Gartner quería resaltar la importancia de la analítica de datos y de negocios *(Analytics*).

El informe refleja los casos de usos que reflejan la diversidad de datos:

- *Data Warehouse* tradicional.
- *Data Warehouse* en tiempo real.
- *Data Warehouse* independiente del contexto.
- *Data Warehouse* lógico (LDW, Logical Data Warehouse).

Gartner evalúa en el nuevo informe soluciones de *Big Data* y de *Analítica de Big Data* y por ello considera todos los posibles casos de uso de los almacenes de datos: tradicional, en tiempo real, independiente del contexto y lógico.

El informe de 2018 evaluó 22 proveedores de soluciones de DW con un enfoque en su visión empresarial, así como la capacidad de ejecución de las compañías que incluyen optimizaciones específicas para dar soporte al procesamiento analítico incluyendo:

- Soporte para el procesamiento relacional.
- Soporte para el procesamiento no relacional (incluyendo gráficos).
- Aprendizaje automático (*Machine Learning*) e Inteligencia Artificial.
- Lenguajes de programación enfocados en analítica de datos como R y Python.

Los datos de las soluciones de software que se evalúan, no solo se almacenan en estructuras relacionales (bases de datos relacionales), sino que pueden utilizarse múltiples modelos como, por ejemplo, relacional, XML, JSON, textos, gráficos y geoespacial (Véase capítulos 5 y 6).



Figura 4.14 Cuadrante Mágico de Gartner 2018 de *Data Management Solutions for Analytics*
Fuente: Gartner (Febrero de 2018). `www.gartner.com` y `https://gtnr.it/2Nstd44`

Los proveedores líderes de Gartner en 2018 son: Teradata, Oracle, AWS (Amazon *Web Services*), Microsoft, IBM y SAP. Como proveedores "*visionarios*" destacan MarkLogic y Google. Entre los proveedores de soluciones "*jugadores de nicho*" sobresalen numerosas soluciones de *data warehouses* y bases de datos NoSQL, tales como: Pivotal, Cloudera, MapR Technologies, Hortonwork, Neo4Jy Actian. Los proveedores "*aspirantes*" a líderes son: Snowflake y MemSQL.

Otros cuadrantes mágicos de Gartner de gestión de datos

Gartner publica también otros informes anuales (Cuadrantes Mágicos) relacionados con la gestión y análisis de datos, además del ya comentado DMSA y son:

Cuadrante Mágico de Gartner para Herramientas de Infraestructura de Datos

El Gartner 2018 *Magic Quadrant for Data Integration Tools* (julio, 2018), requiere las siguientes funcionalidades exigidas por la consultora a las empresas proveedoras de herramientas de infraestructuras de datos son:

- Adquisición de datos para Inteligencia de Negocios, Analítica y *Data Warehousing*.
- Gestión de datos maestros (*Master Data Management*).
- Consistencia de datos entre aplicaciones comerciales.
- Compartición de datos entre empresas.
- Gestión de datos en Lagos de Datos (*Data Lakes*).
- Migración de datos.



Source: Gartner (July 2018)

Figura 4.15. ***Magic Quadrant for Data Integration Tools*** **2018**
Fuente: Gartner, julio 2018.

Las empresas seleccionadas por Gartner en el informe de 2018 son: *Líderes*, IBM, SAP, SAS, Oracle, Informática y Talend; *Aspirantes*, Microsoft, Attunity y Denodo; *Nicho*: TIBCO Software. Hitachi Vanta, Adeptia, Syncsort y Actian; *Visionarios*, Information Builder.

Cuadrante Mágico de Gartner para Sistemas de Gestión de Bases de Datos Operacionales

Este informe de Gartner de gestión de datos está más directamente relacionado con proveedores de soluciones de bases de datos y *big data* (ver capítulos 5 y 6) pero por su relación con la gestión y administración de datos los incluiremos en esta sección. Gartner incluye productos de bases de datos relacionales y no relacionales adaptados a un amplio rango de aplicaciones transaccionales a nivel de empresa, incluyendo productos de sistemas de información de empresa como ERP y CRM e incluso soluciones para gestión de los datos del Internet de las Cosas. De igual forma Gartner, trata de examinar las empresas que ofrecen herramientas modernas y arquitecturas abierta para un procesamiento de datos

más eficiente. Por ello también busca proveedores de soluciones que ofrezcan bases de datos tradicionales de inteligencia de negocio con los modernos lagos de datos (*Data Lakes*) como es el caso de las infraestructuras de *Apache Spark* para procesamiento de datos en tiempo real.

Los requerimientos más destacados, exigidos por Gartner a los proveedores de soluciones de bases de datos para poder figurar en el cuadrante mágico de bases de datos operacionales son:

- Alto rendimiento y seguridad.
- Inteligencia de negocios móvil y en la nube, integral en cualquier dispositivo.
- Análisis avanzado en la base de datos tanto "*en la empresa, in-situ*" como en la nube.

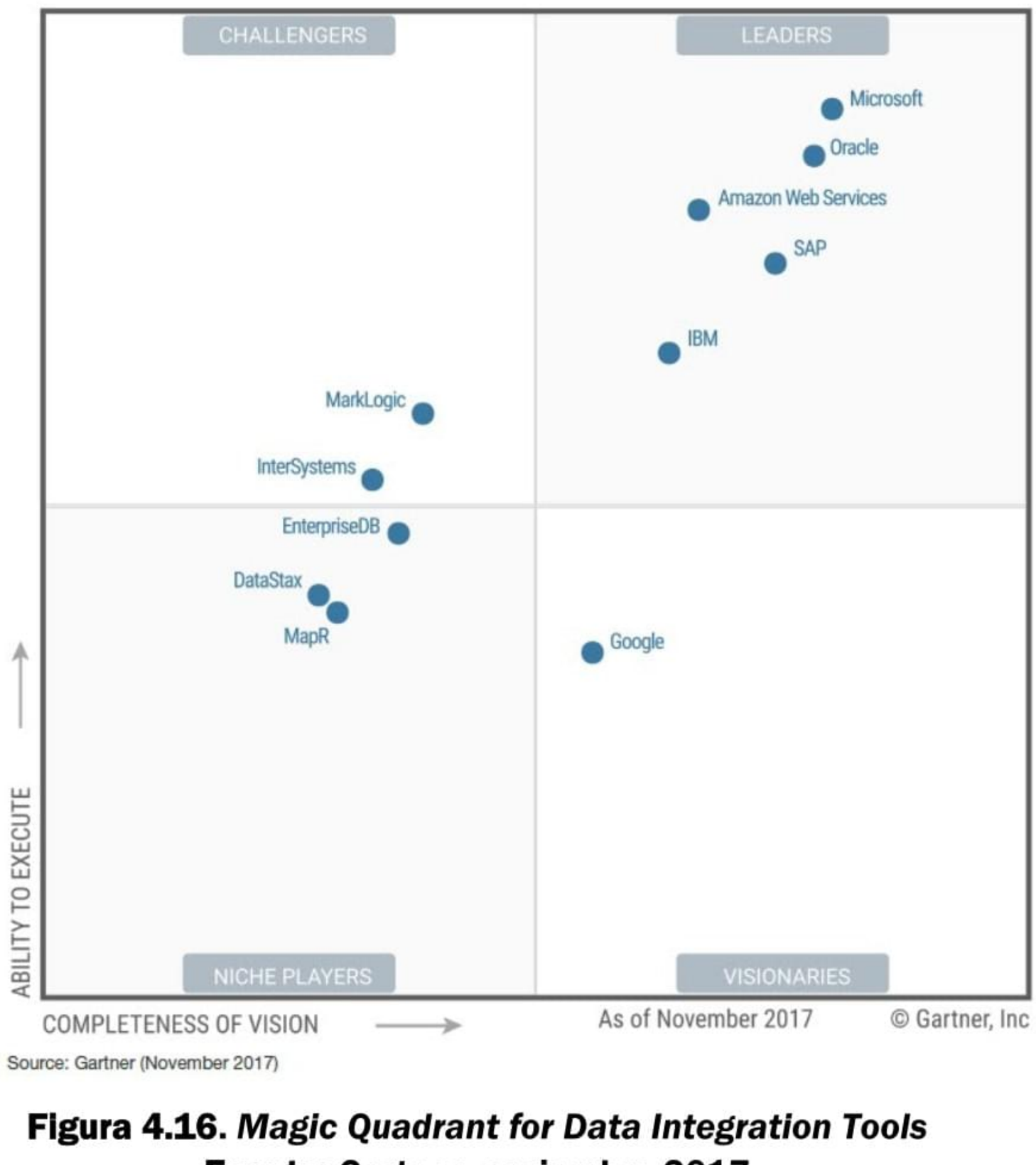


Figura 4.16. ***Magic Quadrant for Data Integration Tools***
Fuente: Gartner, noviembre 2017

Las empresas seleccionadas por Gartner en el informe de 2018 son: *Líderes*, AWS de Amazon, Microsoft, Oracle, SAP, IBM; *Aspirantes*, MarkLogic y InterSystems; *Nicho*: EnterpriseDB, MapR y DataStax; *Visionarios*, Google.

4.16 RESUMEN

- Los **datos** son uno de los activos estratégicos más importantes de las organizaciones. Son elementos discretos, sin valor por sí solos, porque su valor reside en una gestión adecuada para convertirse en **información** y, luego, en **conocimiento**.
- Los datos se caracterizan por su **formato** (estructura de datos), su **contenido** (tipos de datos) y **procedencia de origen** (de dónde vienen). El formato donde se almacenan los datos es de dos tipos: **estructurado** y **no estructurado**.
- Los principales problemas que pueden comprometer la validez e integridad de los datos son: **datos incorrectos**, datos no actualizados y **pérdida de datos**.
- Un *Data Warehouse* agrega datos de bases de datos transaccionales para que puedan ser analizados y puedan servir, por ejemplo, para identificar y examinar tendencias de los negocios y ayudar en la planificación y toma de decisiones. Los *Data Warehouses* y los *Data Marts* son el soporte fundamental de las aplicaciones de Inteligencia de Negocios (Business Intelligence).
- El gobierno del sistema de *Data Warehouse* debe ser una parte del gobierno de TI, y requiere que las personas, comités y procesos se realicen en el momento oportuno y cumpliendo las reglas de gobiernos de las aplicaciones correspondientes.
- Una estrategia para la implementación del gobierno de datos es la gestión de datos maestro (*Master Data Management*), que abarca todos los procesos y aplicaciones de una organización. Proporciona a las empresas la capacidad de almacenar, mantener, intercambiar y sincronizar los datos maestros, de modo que sean consistentes, precisos (correctos) y oportunos.
- Existen dos tipos de bases de datos: centralizadas y distribuidas.

NOTAS

[1]*https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/*

[2]*Andy Patrizio. Data Lake vs Data Warehouse. Datamation, junio 2017. http://www.datamation.com/big-data/data-lake-vs-data-warehouse.html*

[3]*Tamara Dull es directora de Emerging Technologies, SAS Best Practices en SAS Institute. http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html*

CAPÍTULO 5

BIG DATA: ARQUITECTURA, ECOSISTEMA HADOOP Y *OPEN DATA)*

CONTENIDO

5.1 Introducción
5.2 Definición de Big Data
5.3 Tipos de datos
5.4 Características de Big Data
5.5 Breve reseña histórica de Big Data
5.6 Fuentes de datos
5.7 Datificación
5.8 Datos en organizaciones y empresas
5.9 Arquitectura de Big Data
5.10 Ecosistema Hadoop
5.11 Herramientas más utilizadas de Hadoop en Big Data
5.12 Open Data: el movimiento de los datos abiertos
5.13 Iniciativas e índices internacionales de Open Data
5.14 RESUMEN

OBJETIVOS
• Examinar y conocer la definición y características principales de Big Data.
• Identificar y analizar los diferentes tipos de datos manejados en los sistemas de Inteligencia de Negocios.
• Localizar e identificar las diferentes fuentes de datos.
• Introducción a la tendencia de *datificación*.
• Describir la arquitectura de Big Data.
• Identificar el ecosistema Hadoop (soporte de Big Data) y sus componentes principales.
• Conocer e identificar las herramientas más utilizadas en los sistemas de Big Data.
• Examinar y analizar el concepto de Open Data y su impacto en los sistemas de Inteligencia de Negocios.

Disponible online

CAPÍTULO 6
BASES DE DATOS NOSQL Y "EN MEMORIA"

CONTENIDO

6.1 Introducción
6.2 Tipos de bases de datos
6.3 Sistemas de bases de datos MPP
6.4 ¿Qué es NoSQL?
6.5 Bases de datos NoSQL
6.6 Modelos de bases de datos NoSQL
6.7 Breve historia de NoSQL
6.8 Computación "en memoria" (*in memory*)
6.9 Bases de datos "en memoria"
6.10 Proveedores de soluciones de software de bases de datos
6.11 RESUMEN

OBJETIVOS
• Conocer y revisar los fundamentos y características de las bases de datos NoSQL.
• Conocer los diferentes tipos de bases de datos NoSQL.
• Conocer una breve historia de NoSQL y cómo se ha llegado a ellas.
• Entender el concepto de computación "en memoria", y por extensión, las bases de datos "en memoria" (*in memory*).
• Conocer y comprender el concepto de Hadoop como marco de trabajo y ecosistema de programación y de la gestión de grandes volúmenes de datos.
• Identificar y diferenciar las plataformas más populares de Hadoop.

Disponible online

CAPÍTULO 7

VISUALIZACIÓN DE DATOS: INFORMES Y CONSULTAS, CUADROS DE MANDO (*DASHBOARDS*) Y CUADRO DE MANDO INTEGRAL (CMI)

CONTENIDO

7.1 Introducción
7.2 Conceptos generales de visualización de datos
7.3 Gráficos
7.4 Tipos de gráficos
7.5 Mapas
7.6 Infografías
7.7 Informes (*reporting*) y consultas (*query*)
7.8 Cuadros de mando (*dashboards*)
7.9 Narrativa de Datos (*Data Storytelling*)
7.10 Cuadro de Mando Integral (CMI) o *Balanced Scorecard*
7.11 Herramientas de visualización de datos
7.12 RESUMEN

OBJETIVOS
• Definir y conocer la visualización de datos: conceptos, gráficos y herramientas.
• Conocer cuáles son los diferentes tipos de gráficos utilizados en soluciones de Inteligencia de Negocios, Analítica de Datos y de Negocios.
• Conocer y comprender las diferencias de tipos de datos especiales: mapas, cartogramas, histogramas, infografías.
• Conocer las operaciones básicas de presentación de resultados: informes o reportes (*reporting*), consultas (*query*), cuadros de mando (*dashboards*).
• Conocer las herramientas de visualización más populares en organizaciones y empresas.
• Aprender y conocer el uso de los cuadros de mando (*dashboards*) para la ayuda en la toma de decisiones.
• Introducción a la disciplina de Analítica Visual de Datos (*Visual Analytics*).
• Conocer y aprender a utilizar el Cuadro de Mando Integral (*Balanced Scorecard*), así como las diferencias esenciales con los cuadros de mando o tableros de control (*dashboards*).
• Conocer cuál es la misión del Cuadro de Mando Integral: características y componentes.
• Conocer los conceptos modernos de Narración de Datos (*Data Storytelling*).

7.1 INTRODUCCIÓN

Una vez que se ha realizado el análisis de datos, se disponen de resultados que se han de comunicar y presentar a los diferentes niveles de empleados de la organización —y, en su caso, clientes, *partners*—, de modo que aprendan de ellos y les ayuden a tomar las decisiones y conclusiones más adecuadas para el negocio. La información que se ha de presentar debe ser clara, sencilla y lo más intuitiva posible.

El análisis de datos, en general, y de *Big Data*, en particular, sólo es útil si garantiza que las personas a quienes va destinado obtienen la información adecuada, en el formato adecuado y el momento en que se necesita, de modo que puedan tomar las decisiones correctas de la forma más eficiente posible y siempre de modo ubicuo. Una vez analizados los datos, se ha de tener en cuenta quién necesita los resultados, con el objeto de tomar mejores decisiones estratégicas y adaptar la visualización de datos a las necesidades de cada rol profesional.

La visualización de datos puede darse en numerosas formas: **gráficos y tablas** (cuadros), **informes tradicionales**, **infografías** (con mapas, textos, datos, comportamientos, emociones), **nubes de palabras, mapas mentales**, etcétera. Otros medios a disposición de las empresas y de gran impacto corporativo son los cuadros de mando o tableros de control (*dashboards*) y, desde el punto de vista

de gestión empresarial, los Cuadros de Mando Integral (*Balanced Scorecards*), CMI, que son herramientas de gran eficiencia en las empresas.

Las corporaciones disponen de un gran número de herramientas de visualización, tanto de *software* propietario como de código abierto (*open source*) y gratuitas.

7.2 CONCEPTOS GENERALES DE VISUALIZACIÓN DE DATOS

En la era de los grandes datos, las compañías necesitan información más relevante, más útil, de mejor calidad y disponer de ella en el momento adecuado. La presentación de datos en formato visual, mediante la visualización de datos y herramientas de Analítica Visual, ayudan a los usuarios a ver patrones y relaciones en grandes cantidades de datos, lo que sería difícil de discernir si los datos se presentaran en formato tradicional de listas de texto o números. Las personas suelen verse más ágiles cuando ellas pueden acceder a filtrar información que se presenta visualmente y, de este modo, pueden aumentar sus capacidades de percepción y creación de nuevas ideas sobre un determinado asunto.

El banco español BBVA —líder en tecnologías y transformación digital en la banca— define visualización de datos como la disciplina que permite transformar las relaciones numéricas de los datos en impactos visuales y la comprensión de la información. Otra definición muy sobresaliente es la dada por el fabricante de soluciones de *software* de visualización de datos, SAS (uno de los proveedores más acreditados en herramientas de Analítica Visual), que define visualización de datos como la exploración visual e interactiva y la correspondiente representación gráfica de datos de cualquier dimensión, naturaleza (estructurados y no estructurados) y origen. La visualización de datos permite a las personas ver cosas que antes no eran evidentes; se pueden identificar tendencias rápida y fácilmente. Transmiten información de una manera universal y simplifican la tarea de compartir ideas con otras personas. Es una manera de obtener información rápida mediante exploración visual, por medio de informes consistentes y de intercambio flexible de dicha información.

La visualización de datos trata de explorar y analizar los datos de una forma visual y rápida; permite compartir información con rapidez y tomar decisiones adecuadas en el momento adecuado. Otra ventaja que ofrece la visualización de datos es mantener bajo control a las métricas y, sobre todo, a los indicadores clave de rendimiento, KPI empresariales.

Analítica Visual de *Big Data*

El análisis de los grandes volúmenes de datos requiere la identificación de la información y la necesidad de que los resultados de los informes se han de presentar a las personas responsables, de un modo significativo, y que los

puedan entender con facilidad. Es muy importante que la información adecuada se presente de modo comprensible y se comunique a los usuarios correspondientes en el formato adecuado y cuando la necesiten, con el objeto de que puedan tomar las decisiones correctas en el momento oportuno; cada vez con mayor frecuencia, se requiere que esta información esté disponible en tiempo real y, también, que se pueda acceder a ella con cualquier dispositivo y en modo **omnicanal**.

Tradicionalmente, los informes para comunicar los resultados utilizan cuadros, diagramas y tablas, junto a los formatos numéricos puros (sin gráficos ni cuadros). Cada día con mayor frecuencia, se presentan los resultados mediante conversaciones, charlas y tertulias, respaldados por presentaciones numéricas o gráficas, en una conversación oral.

La proliferación de aplicaciones de *Big Data* en la Inteligencia de Negocios ha facilitado la creación de una alta gama de herramientas de visualización, de modo que permiten presentar los resultados de los análisis de los datos con un aspecto más atractivo y comprensible. Una de las tendencias de mayor impacto en las herramientas de visualización son las conocidas como **Narraciones de Datos** (*Data Storytelling*), que se están convirtiendo en una de las más empleadas para la presentación de resultados y son compatibles con las infografías, mapas mentales o nubes de palabras.

Las herramientas de visualización de datos se construyen mediante soluciones de *software* propietario o de código abierto, que se adquieren a los proveedores comerciales de *software*, bien como programas independientes o integrados con otras herramientas de Inteligencia de negocios, a modo de soluciones integradas (*suites*).

Tradicionalmente, los informes han utilizado varios tipos de gráficos y cuadros o tablas para ayudar a la visualización de resultados. En los próximos apartados, veremos los más comunes.

En los últimos años se ha consolidado una tendencia: la integración de datos en formatos sencillos (de una página, cartón) dentro de una misma presentación y que mezcla todo tipo de datos. Hoy en día, estas soluciones muy eficaces se pueden obtener mediante herramientas de visualización conocidas como cuadros de mando o tableros de control (*dashboards)* y que vienen incluidas en numerosas soluciones de *software* de numerosos proveedores.

7.3 GRÁFICOS

En el proceso de realización de un informe, uno de los aspectos más importante que se debe considerar es la selección del tipo de gráfico, tabla o diagrama, que en general denominaremos gráficos. La definición de gráfico según la RAE (Real Academia Española) tiene dos acepciones relacionadas con la visualización de datos:

4(m). Representación de datos numéricos por medio de una o varias líneas que hacen visible la relación que esos datos guardan entre sí.

5(f). Representación por medio de líneas/gráfica.

El INE (Instituto Nacional de Estadística de España) denomina gráfico estadístico a una representación visual de una serie de datos estadísticos y considera que es una herramienta muy eficaz ya que:

- Capta la atención del lector.
- Presenta la información de forma sencilla, clara y precisa.
- No induce a error.
- Facilita la comparación de datos y destaca las tendencias y las diferencias.
- Ilustra el mensaje, tema o trama del texto al que acompaña.

En resumen, el gráfico es una herramienta muy eficaz para presentar la información de manera, clara, precisa y sencilla, facilitando la comparación de datos y resaltando diferencias y tendencias.

El INE considera que existen diferentes tipos de datos y, en función de ellos, se pueden representar a través de los gráficos. Los tipos de datos utilizados en los gráficos se clasifican en dos grandes grupos:

- **Cualitativos.** Se refieren a cualidades o modalidades que no pueden expresarse numéricamente. Pueden ser:
  - **Ordinales** (siguen un orden o secuencia): por ejemplo, abecedario, meses del año, etcétera.
  - **Categóricos** (no siguen ningún orden): por ejemplo, el estado civil de las personas).
- **Cuantitativos.** Se refieren a cantidades o valores numéricos. Pueden ser:
  - **Discretos**. Valores enteros (0, 1, 2, 3): por ejemplo, número de hijos, número de alumnos, etcétera.
  - **Continuos**. Toman cualquier valor dentro de un intervalo.

Existen numerosos tipos de gráficos (diagrama es un término sinónimo muy empleado) y uno de los aspectos más importantes y, a veces más complejos, es la selección del tipo de gráfico que se ha de utilizar en un informe o reporte, o en un cuadro de mando. En el apartado 7.4 del sitio web del libro se describirán los más populares utilizados en los diagramas de visualización, así como otros diferentes gráficos que han alcanzado gran notoriedad en los cuadros de mando o paneles de control que se describen más adelante.

Además, y como recursos muy prácticos, recomendamos al lector la descarga de los documentos, en formato libre, del INE y de Tableau:

Guías prácticas de tipos de datos
INE. Portal Explica del Instituto Nacional de Estadística (España).
`www.ine.es/explica/explica_pasos_tipos_graficos.html`
Tableau. ¿Qué gráfico es el más adecuado?
`https://www.tableau.com/es-es/learn/whitepapers/which-chart-or-graph-is-right-for-you`

7.4 TIPOS DE GRÁFICOS

Disponible online.

7.5 MAPAS

Cartograma

Es un mapa en el que se presentan datos estadísticos por regiones, colocando un número o coloreando las diferentes zonas, en función del dato que representa.



Figura 7.1. Cartograma

Mapas de calor

Es una representación gráfica de los datos, donde los valores individuales contenidos en una matriz se representan como colores. Los mapas de calor son una excelente manera de comparar datos en dos categorías mediante color. El efecto es ver rápidamente donde la intersección de las categorías es más fuerte y dónde es más débil.

Se usan para ver la visualización de relación entre dos factores, por ejemplo, análisis de segmentación de mercado objetivo, adopción de productos en regiones, ejemplos de ventas por agente individual, etcétera.

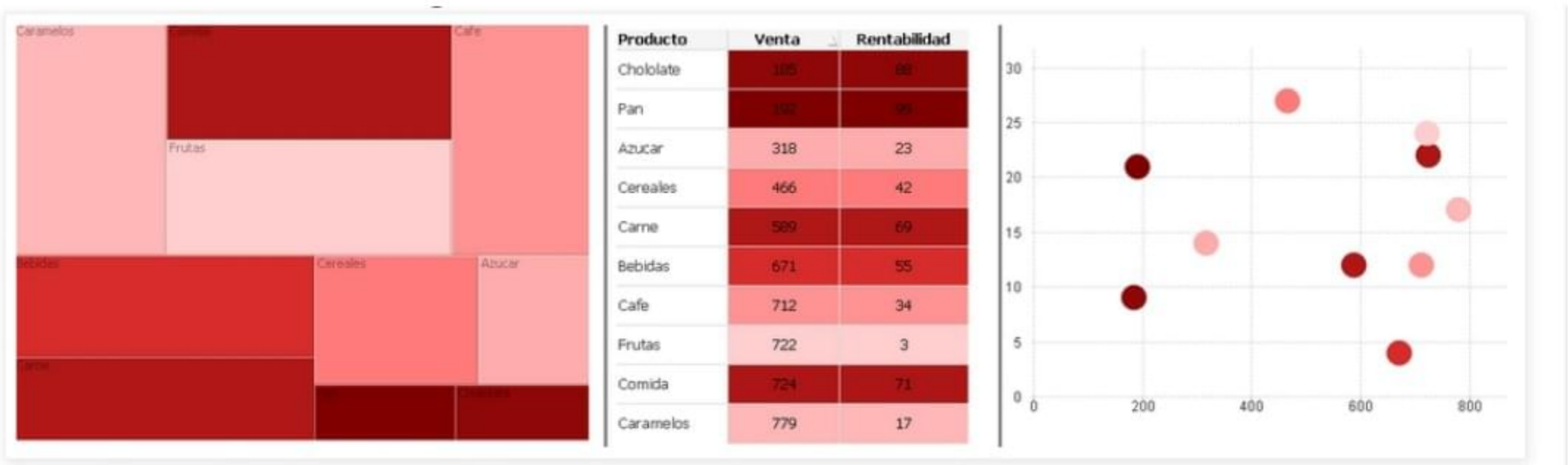


Figura 7.2 Mapa de calor.

Fuente: http://qlikviewapuntes.blogspot.com/2014/02/mapa-de-calor-usando-asistente.html

Diagrama de Sankey

Es un diagrama de flujos antiguo que se ha revitalizado gracias a las nuevas herramientas que facilitan su realización (en lugar de hacerlo de modo manual). Los anchos de las flechas representan el volumen.

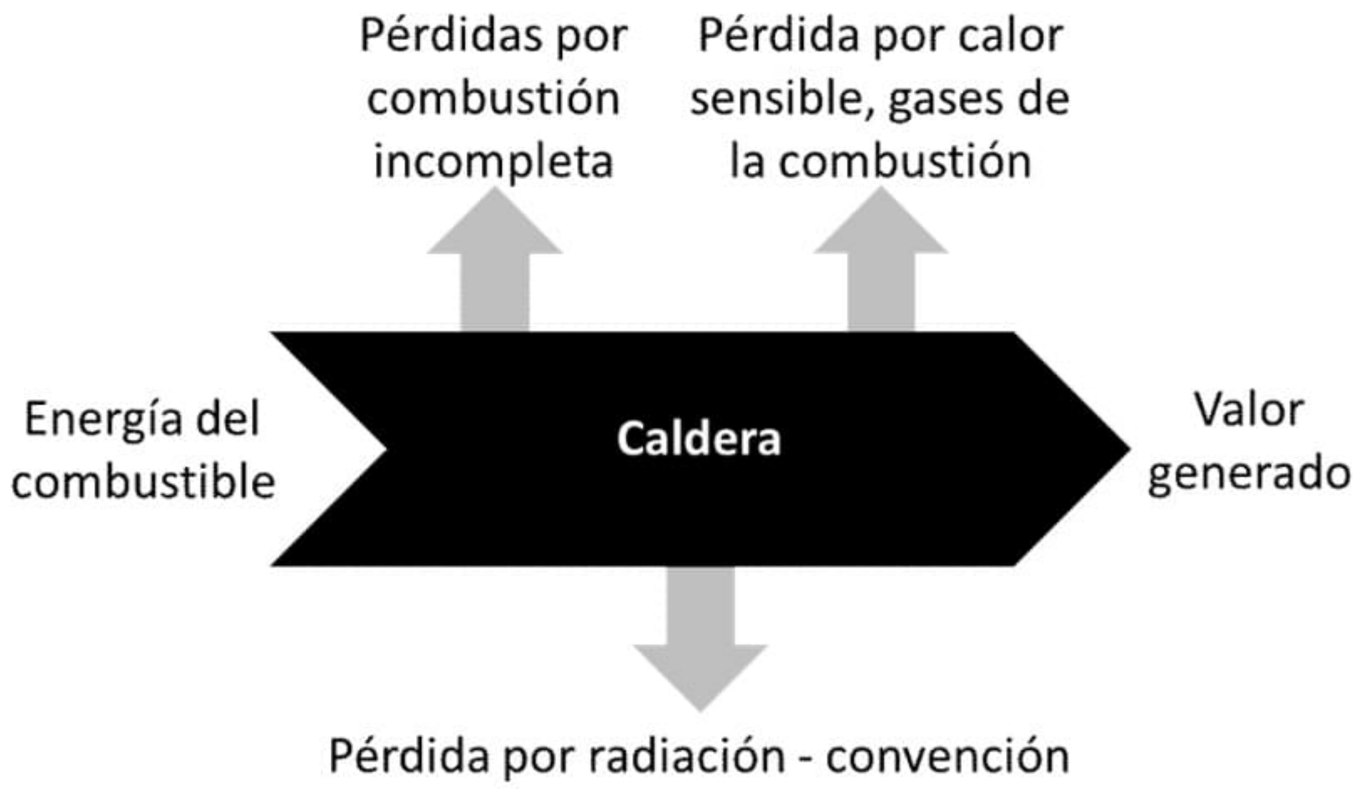


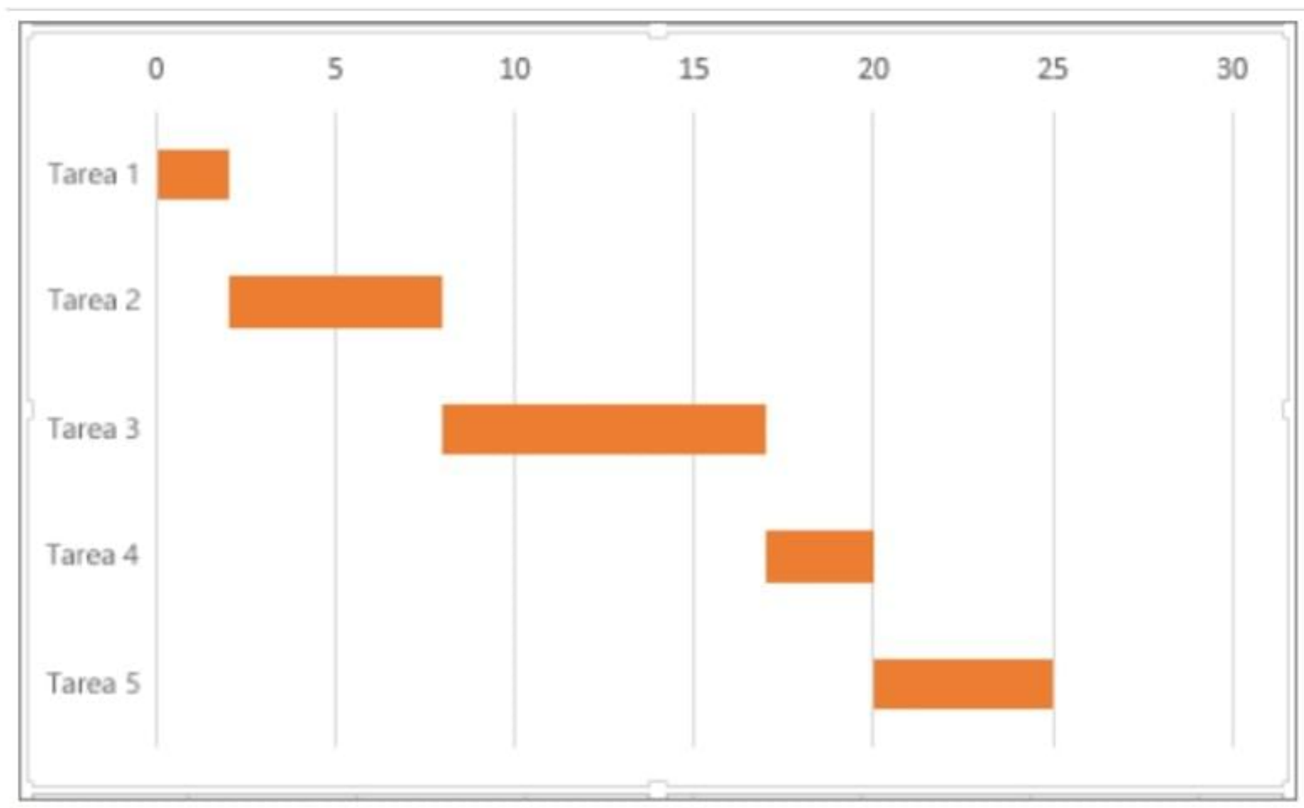
Figura 7.3 Diagrama de Sankey de una caldera

Fuente: https://www.researchgate.net/figure/Figura-7-Diagrama-de-sankey-de-una-caldera-1_fig6_316859930

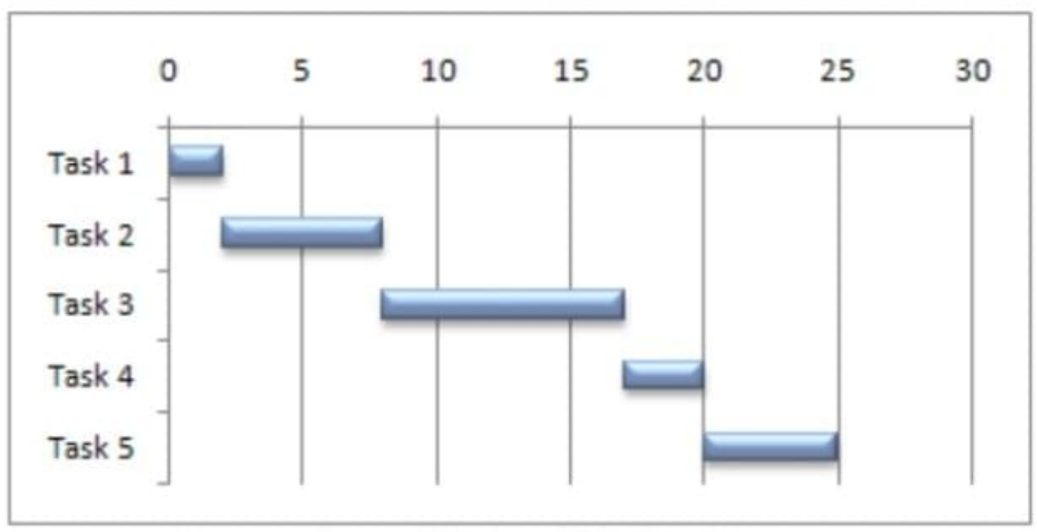
Diagrama de Gantt

Un gráfico o diagrama de Gantt es una herramienta muy eficaz para la ilustración de fechas de inicio y terminación de los elementos de un proyecto. El cumplimiento de los pazos es primordial para el éxito de un proyecto y requiere conocer las necesidades que se han de cumplir y cuando se han de alcanzar. Los diagramas de Gantt se han asociado tradicionalmente con la administración de proyectos, pero pueden ser utilizados para otras aplicaciones donde las personas o máquinas cambian sus actuaciones con el tiempo; por ejemplo, en la planificación de recursos ver los tiempos que las personas alcanzan hitos u objetivo explícitos, niveles de certificación y como se distribuyen en el tiempo.

Los gráficos de Gantt se pueden utilizar: 1. Visualización de del programa de un proyecto; 2. Visualización de otros temas en uso con el tiempo.



Ejemplo de simulación de diagrama de Gantt en Excel 2016 y 2013



Ejemplo de simulación de diagrama de Gantt en Excel 2010 y 2007

Figura 7.4 Diagrama de Gantt
Fuente: Microsoft. https://support.office.com/es-es/article/presentar-datos-en-un-diagrama-de-gantt-en-excel-f8910ab4-ceda-4521-8207-f0fb34d9e2b6

Mapa de árbol

Es un método para la visualización de datos jerárquicos mediante el uso de rectángulos anidados. Utilizan una serie de rectángulos, dentro de otros rectángulos, para mostrar los datos jerárquicos como una proporción del todo. Los datos se observan relacionados como en un árbol; a cada rama se le asigna un rectángulo que representa cuántos datos comprende. Cada rectángulo se subdivide, a su vez, en rectángulos más pequeños o subramas, nuevamente basándose en su proporción con el todo.

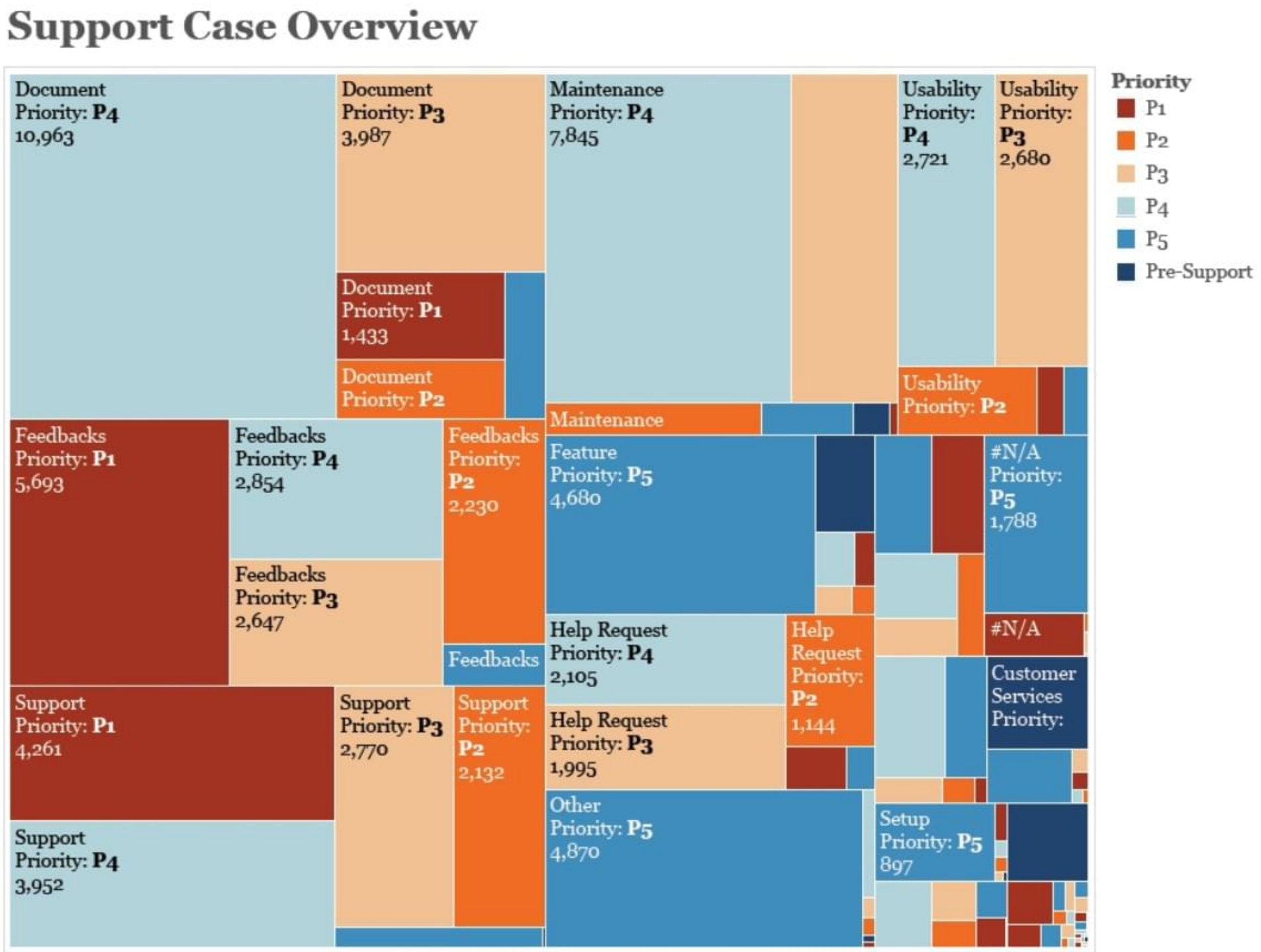


Figura 7.5 Mapa de árbol. Fuente: Maila Hardin *et al. ¿Qué tabla o gráfico es el adecuado para usted?* Tableau (pág. 23)

Diagrama de caja (*Box Plot*)

Es un tipo de gráfico que utiliza los cuartiles para representar un conjunto de datos. Permite observar de un vistazo las distribuciones de los datos y sus principales características: centralidad, dispersión, simetría y tamaño de las cajas.

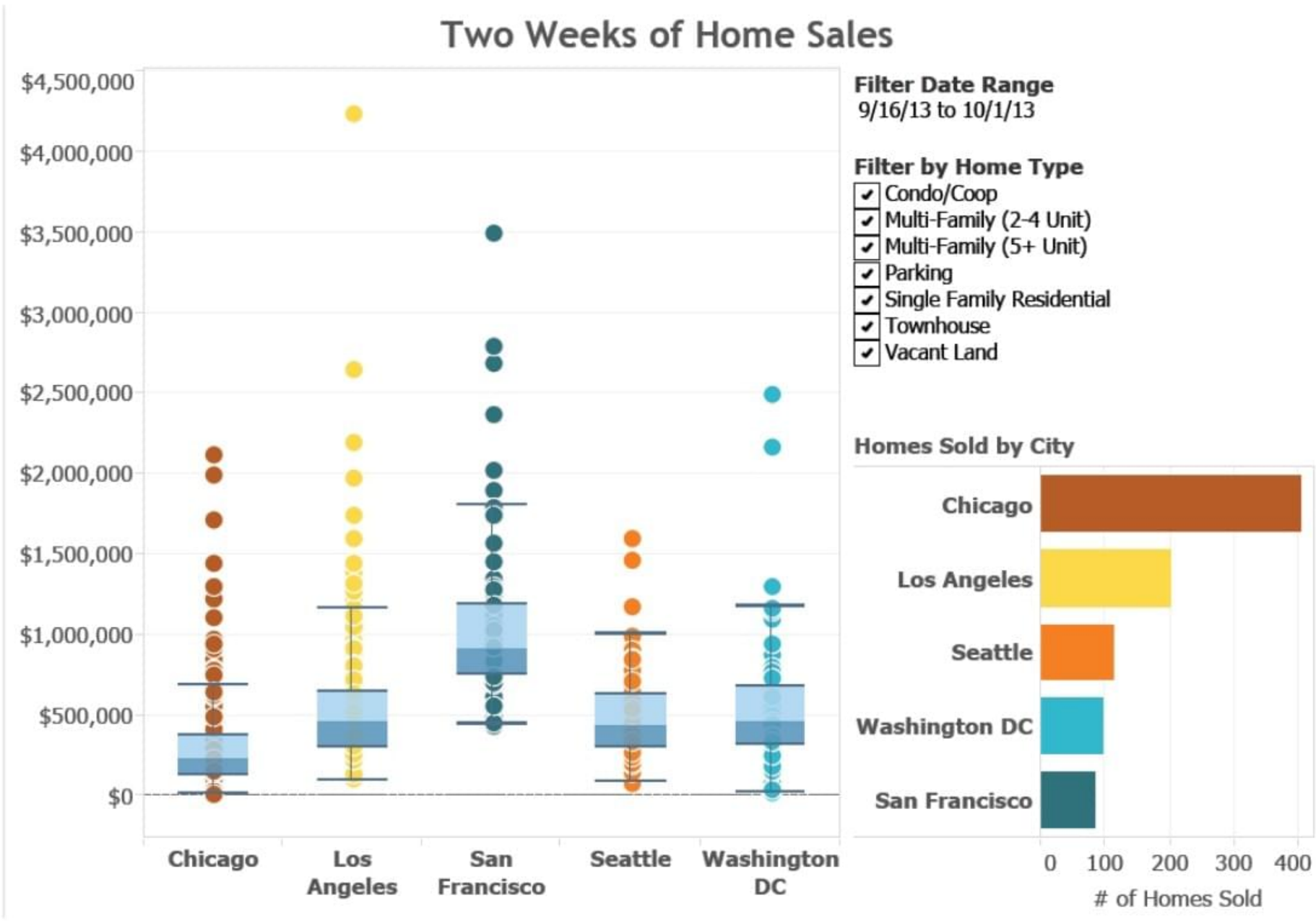


Figura 7.6 Diagrama de cajas

Árbol de nodos

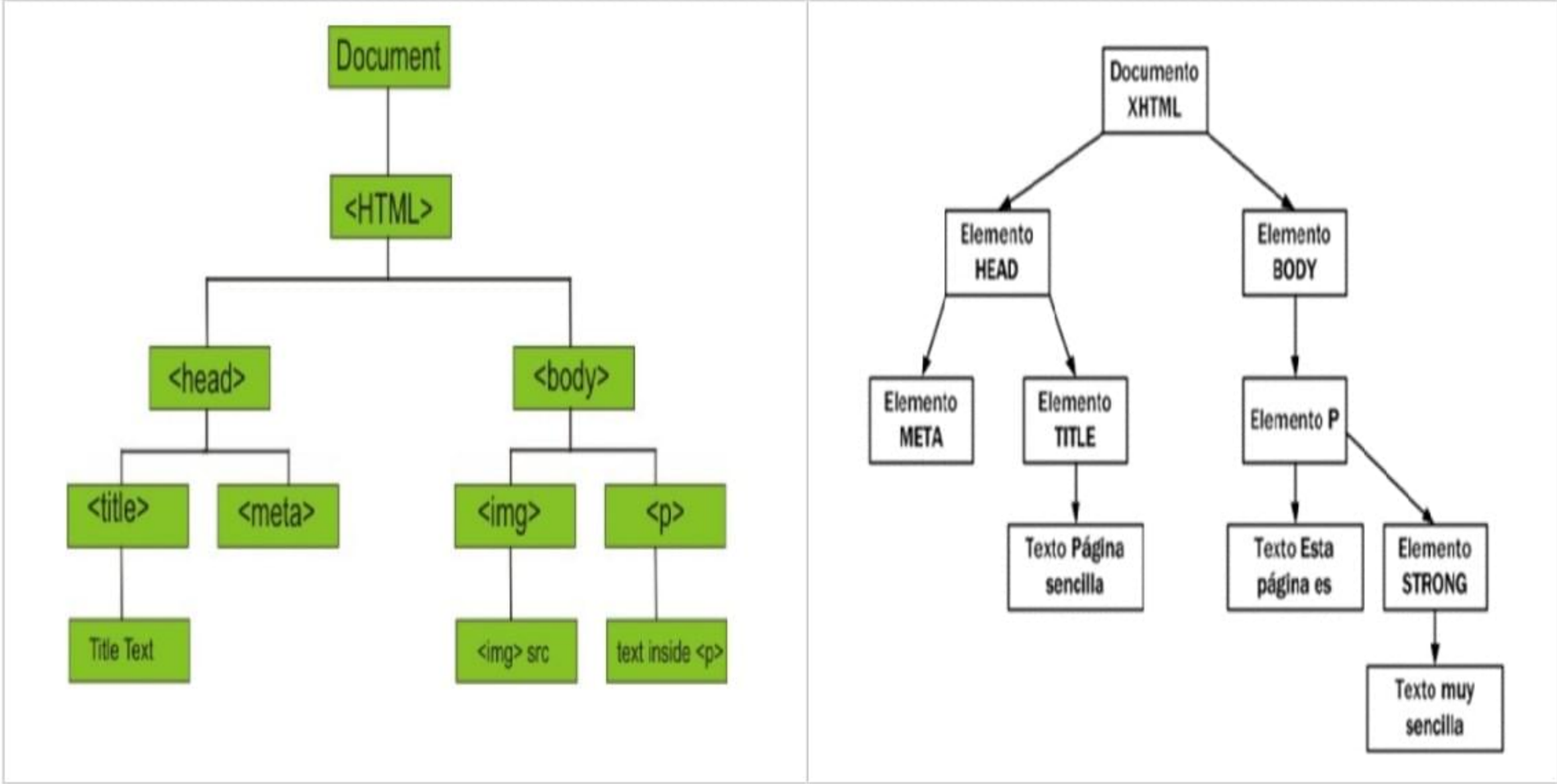


Figura 7.7 Árbol de nodos. Fuente: Árbol de nodos generados automáticamente por DOM (Document Model Objetct) a partir del código HTML de la página:
`http://librosweb.es/libro/javascript/capitulo_5/arbol_de_nodos.html`

Es un diagrama muy utilizado para visualizar relaciones. Las visualizaciones de grandes de datos lo han actualizado, ya que son una evolución de los

organigramas y árboles familiares. La necesidad de visualizaciones de grandes volúmenes de datos ha convertido a estos gráficos en herramientas imprescindibles en numerosas aplicaciones de negocios.

7.6 INFOGRAFÍAS

Las infografías son una de las herramientas visuales que más han crecido desde la implantación de las tecnologías de *Big Data*, especialmente el análisis de datos y las tecnologías habilitadoras y facilitadoras de la Industria 4.0 ya descritas.

Las oportunidades y beneficios que ha traído el análisis de los grandes volúmenes de datos han obligado a utilizar nuevos modos de comunicar e informar los resultados y, por consiguiente, nuevas herramientas de visualización adaptadas a estos grandes datos.

Una infografía es una mezcla de información y gráficos, como señala su nombre, y una representación visual de una página (o página desplazable a lo largo de una página web que puede ocupar una o varias páginas, según la escala de la infografía), que trata de visualizar los datos, la información y el conocimiento necesario para ayudar en la toma de decisiones, de una manera rápida, clara y sencilla.

Las infografías se suelen utilizar con mucha profusión en periódicos, revistas y otros medios de comunicación. Sin embargo, a medida que los *Big Data* y sus herramientas siguen introduciéndose en organizaciones y empresas, los ejecutivos y directivos necesitan un sistema de visualización de datos coherente, simple e intuitivo, que les permita extraer los mensajes clave que se adapten a sus prioridades y le ayuden a su necesaria toma de decisiones.

Una infografía debe ser una solución innovadora que sintetice los mensajes clave y describa la historia de los datos en una página física (o desplazable). Bernard Marr (2016)[1] considera que una buena infografía debe tener tres partes muy claras:

- **Atractiva visualmente**: uso de color, gráficos e iconos.
- **Contenido útil**: uso de marcos temporales, estadísticas y referencias.
- **Transmitir conocimientos**: uso de hechos y deducciones.

Marr señala la necesidad de combinar las características anteriores para contar una historia, ya que la infografía no consiste solamente en encontrar la manera más accesible de presentar datos, sino que se trata de encontrar la mejor manera de iniciar visualmente conversaciones e interacciones, que puedan mejorar el proceso de toma de decisiones y el rendimiento. La mejor forma de conseguirlo, insiste Marr, es poner el centro de atención en el mensaje y utilizar lo

que exprese dicho mensaje a partir de una combinación de palabras, números y colores, de modo que "entre por los ojos" y llame la atención del lector.

La avalancha de datos que viven, día a día, organizaciones y empresas ha potenciado la realización de buenas infografías que muestren de un modo visual e instantáneo los informes que necesitan los ejecutivos y, en su caso, los enlaces web necesarios (o documentos fuente impresos) para aquellos que necesiten o deseen conocer más detalles.



Figura 7.8. ¿Qué pasa en un minuto en Internet?

Fuente: https://www.digitalinformationworld.com/2018/05/infographic-internet-minute-2018.html

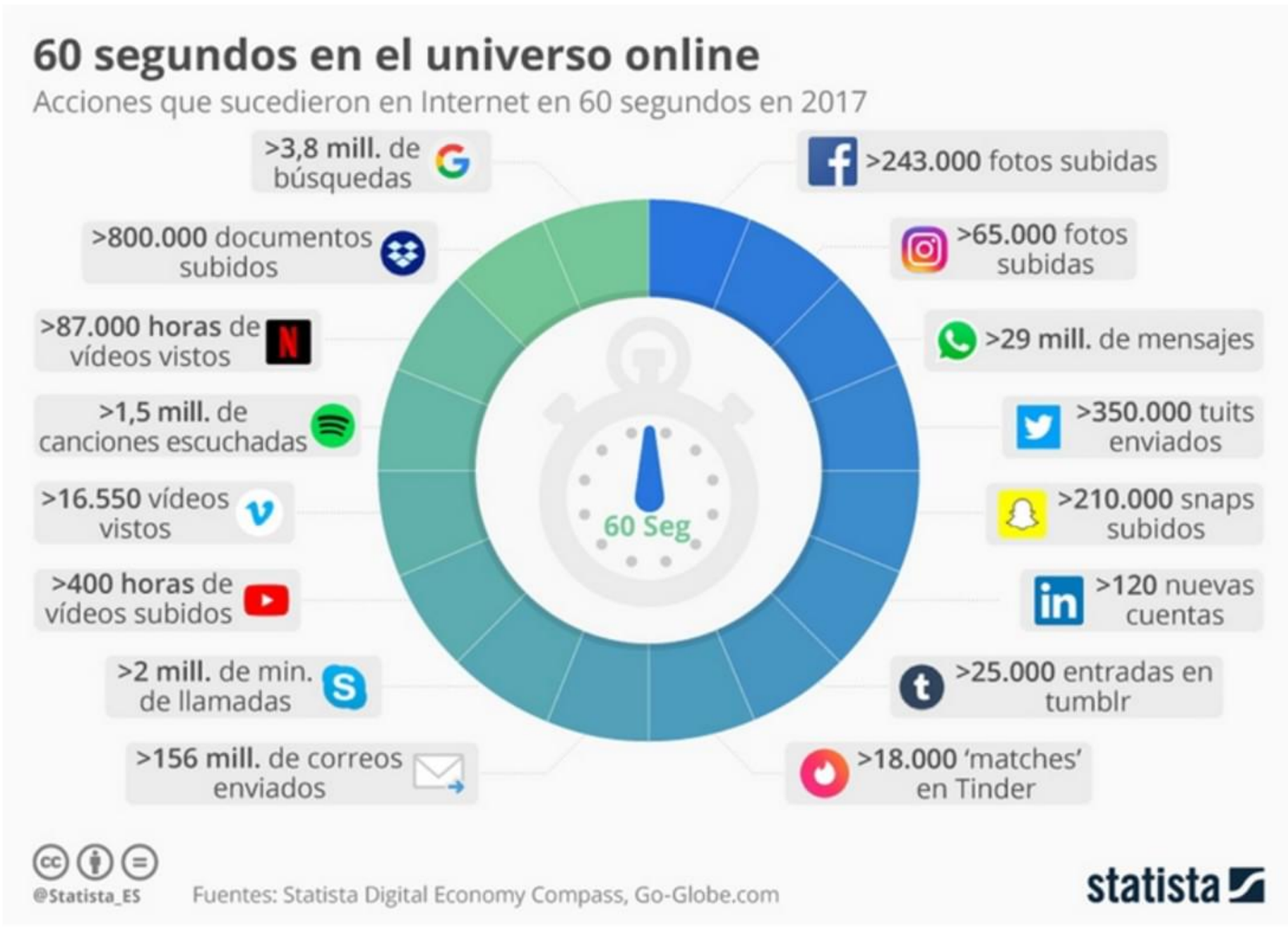


Figura 7.9. Infografía "60 segundos en el universo online"

Fuente: https://es.statista.com/grafico/13159/que-sucede-en-un-minuto-en-internet/



Figura 7.10. Infografía sencilla de ofertas de vuelos de Iberia

7.7 INFORMES (*REPORTING*) Y CONSULTAS (*QUERY*)

La etapa de visualización de resultados en una solución de Inteligencia de Negocios requiere la generación y distribución de informes que permitan tomar decisiones en los tres niveles organizativos de una empresa: operativo, táctico y estratégico.

La Analítica de Negocios incluye informes o reportes (*reporting*) estáticos y dinámicos, todo tipo de consultas, descubrimiento de información, vistas multidimensionales, desgloses detallados, etcétera.

7.7.1 INFORMES (REPORTES)

¿Qué es un informe? Un **informe** o **reporte** (*report*) es un documento mediante el cual se presentan los resultados de uno o varios procesos de negocio. Normalmente, contiene texto acompañado de elementos, tales como gráficos, tablas o diagramas, que facilitan la comprensión de la información presentada. En la actualidad, a los gráficos, mapas y cartogramas se han comenzado a añadir, con gran frecuencia, infografías generalistas o específicas de otras fuentes.

Un **reporte (informe)** ha de ser ágil, inmediato, flexible, fiable, seguro, personalizado, con la información empresarial actualizada. Las soluciones de *reporting* requieren la posibilidad de generación de informes para el usuario final, tanto sean especialistas como simples usuarios que no tengan amplios conocimientos técnicos o de programación. Estas herramientas de *software* específico deben contemplar e integrar soluciones de ERP, CRM, Recursos Humanos, etcétera, y permitir la creación y presentación de informes, reduciendo el tiempo de generación de reportes y facilitando también consultas de información. En la mayoría de las herramientas se deben poder generar reportes y también consultas con gran rapidez, y con mayor intensidad se deben entregar los resultados de manera visual, de modo que puedan ser interpretados y analizados con eficacia.

Los informes están destinados a los usuarios de negocio, en sus diferentes niveles profesionales, que necesitan conocer la información contrastada y consolidada como ayuda a la toma de decisiones.

La arquitectura de un sistema de Inteligencia de Negocios requiere una herramienta de generación de informes. Las fuentes de origen de los informes son diferentes y proceden de sistemas OLAP, almacenes de datos (*Data Warehouses* y *Data Marts*), ODS, ERP, CRM, SCM, redes sociales, bases de datos NoSQL, infraestructuras o marcos de trabajo (*frameworks*) como Hadoop, Spark, etcétera. La última tendencia en informes es incorporar las mayores capacidades de visualización junto con la posibilidad de incorporar documentos en PDF, PPT, dispositivos móviles, informes de la Web, etcétera.

7.7.2 HERRAMIENTAS DE REPORTE (INFORMES)

Las herramientas de informes o *reporting* son las soluciones de *software* que permiten diseñar, distribuir, planificar y administrar o gestionar informes en el contexto de una organización, departamento o área de la compañía. Una herramienta de informes debe permitir la distribución de los informes con gran rapidez, en el momento adecuado y cada vez, con mayor frecuencia, de modo *omnicanal* (cualquier canal de comunicación de los empleados por la compañía: correo electrónico, mensajería, redes sociales, etcétera) y con cualquier dispositivo (teléfono inteligente, tableta, PC, y, cada vez con mayor frecuencia, SmartTV, sobre todo en departamentos de *marketing* y de comunicación).

Tipos de informes

Existen diferentes tipos de informes en función de la interacción ofrecida al usuario y de su dependencia del departamento de informática, de recursos humanos, contaduría, etcétera:

- Rutinarios o predefinidos de producción (se suelen generar de manera automática y se distribuyen periódicamente a los suscriptores mediante correo electrónico o servicios de mensajería).
- Estáticos (tienen un formato definido e inamovible).
- Dinámicos (son formatos ágiles en el tiempo, contenido y forma).
- Paramétricos o parametrizados (presentan parámetros de entrada y son ideales para consultas múltiples y diferentes).
- *Ad Hoc* (creados para el usuario final y a medida de sus necesidades).
- Cuadros de mando (*dashboards* y *scorecards*).

Los informes pueden ser capaces, si el usuario lo requiere, de desglosar vistas detalladas de datos, capacidad para modelar y crear predicciones. Los informes se pueden generar directamente de datos operacionales (por ejemplo, sistemas CRM, ERP y SCM, puntos de venta, etcétera), o bien de almacenes de datos (*Data Warehouses* y *Data Marts*), y, cada día más, de bases de datos no relacionales (NoSQL) o sistemas de archivos de *Big Data*, como Hadoop o Spark.

Los informes tienen como complemento las consultas y búsquedas *ad hoc*.

7.7.3 CONSULTAS

Las consultas pueden ser ordinarias o *ad-hoc* (bajo demanda). Las consultas *ad hoc* permiten a los usuarios consultar información en tiempo real de la computadora, que no están disponibles en los informes periódicos. Las respuestas a esas consultas se necesitan para agilizar la toma de decisiones. El

sistema ha de ser lo suficientemente inteligente para entender lo que desea el usuario; en estos casos, se utilizarán programas escritos en lenguajes SQL y también pueden tener un enfoque dirigido a aplicaciones basadas en la Web.

Las consultas basadas en la Web, para navegar o buscar información propia o externa, permiten a los usuarios acceder, navegar y explorar datos relacionales, con el objeto de ayudar en las mejores tomas de decisiones en tiempo real. Estos métodos de consulta son muy utilizados, sobre todo en los departamentos de *marketing*, a fin de medir el éxito de una campaña, teniendo presente métricas de analítica web que veremos posteriormente.

Existen numerosas herramientas para facilitar las actividades de consultas. Una muy conocida, aunque de *software* propietario, es Cognos de IBM.

En la actualidad, la mayoría de las respuestas a consultas se pueden resolver con herramientas de visualización, aunque las herramientas de computación cognitiva, como IBM Watson, junto con los asistentes virtuales *bots* y *chatbots* (véase capítulo 13), se utilizarán, cada día más, con mayor profusión para responder a consultas especializadas sobre el negocio.

7.7.4 INFORMES GIS (SISTEMAS DE INFORMACIÓN GEOGRÁFICA)

Un sistema de información geográfica, SIG, (GIS en inglés) es un tipo especial de herramienta visual que si bien su uso más frecuente es tal como lo indican sus siglas, en los sistemas de información geográfica, cada día es más popular su uso para la realización de informes en las organizaciones y empresas.

Un sistema GIS es una herramienta muy especial que ayuda a los tomadores de decisiones a visualizar problemas que requieren conocimiento acerca de la distribución geográfica de las personas u otros recursos. El *software* de GIS conecta datos de posición a puntos, líneas y áreas en un mapa. Algunos sistemas GIS tiene capacidad de modelado para cambiar los datos y revisar automáticamente los escenarios de negocio, en función de su geolocalización o la posición de sus clientes o competidores.

Los GIS pueden utilizarse también para ayudar a los gobiernos estatales y locales a calcular tiempos de respuestas ante desastres naturales y otras emergencias; pueden ayudar a los bancos identificando las mejores posiciones geográficas para instalar una nueva oficina, un cajero automático, etcétera.

CASO DE ESTUDIO: DATATEAM CONSULTING

La consultora Datateam Consulting de la Ciudad de México, especializada en reportes y tableros de control, y proveedora de soluciones de *software* de Tableau y Alteryx Software, ofrece en su sitio web una serie de consejos profesionales que consideramos muy prácticos y eficaces para la generación de reportes y

consultas. El *dossier* de la consultora Datateam[2], con cuestiones (preguntas) de interés y respuestas profesionales a los reportes y consultas, es el siguiente:

1. ¿Qué se necesita para la generación de reportes y consultas?

Se necesita tener acceso a sus datos (fuentes de información), conocer cuáles reportes y consultas se desean realizar e instalar el *software* para su dispositivo.

2. ¿Sobre qué fuentes de datos puedo generar los reportes?

Acceso a archivos de texto, Access, Excel, Oracle, SQL Server, MySQL, Hadoop.

3. ¿Qué necesito saber para generar reportes?

Necesita conocer sus datos y sus fuentes. Saber qué reportes y consultas desea generar y estar preparado para visualizar datos. Excel permitirá generar reportes muy complejos.

4. ¿Qué conocimientos necesito para usar Tableau?

Definición de los reportes por generar y de las fuentes de información; acceso a los datos; instalación y licencia del *software* de Tableau; generación de las consultas y reportes siguiendo las instrucciones.

5. ¿Puedo compartir los reportes y consultas que se generan?

Sí, se necesita el *software* Tableau Software Server; se generan los reportes y se publican los informes en el Tableau Software Server.

7.8 CUADROS DE MANDO (*DASHBOARD*)

Los **cuadros de mando** o **tableros de control**[3] digitales (***dashboard*** o ***scorecard*** en inglés), similares a los cuadros de mando de un automóvil, se están convirtiendo en una herramienta de gran popularidad para ayudar a la gestión de la toma decisiones de directivos y profesionales. No debemos confundir el término en español con la metodología de Cuadro de Mando Integral (CMI) o *Balanced Scorecard*, que explicaremos en el apartado 7.10.

7.8.1 ¿QUÉ ES UN CUADRO DE MANDO?

Un cuadro de mando (*dashboard*) es una herramienta de Inteligencia de Negocios que visualiza en una única pantalla (de escritorio, tableta o teléfono inteligente) gráficos y diagramas o tablas de métricas o de indicadores clave de rendimiento (KPI) para ayudar a la gestión de una empresa. De modo más simple, un cuadro

de mando es una representación gráfica de los principales indicadores KPI de una empresa, que ayudan al empleado a identificar las áreas que necesitan mayor atención y que afectan a las consecuciones de la estrategia de la empresa. Un cuadro de mando, en esencia, es un panel de datos en el que las empresas visualizan la información más importante con el objetivo final de conseguir su mayor optimización.

Los cuadros de mando sirven para analizar los datos y detectar los posibles problemas de un negocio, ayudando al empleado a definir las líneas de actuación correctas para conseguir los objetivos establecidos. Al utilizar cuadros de mando, los directivos son capaces de examinar en detalle el nivel del producto y tomar las medidas proactivas o retroactivas con el fin de preservar la calidad y eficacia del mismo. De igual modo, los cuadros de mando ayudan, también, a reducir la cantidad de tiempo requerida por los planificadores y directivos de proyectos para su diseño, construcción y aprobación.

Los cuadros de mando permiten el análisis visual de la información de la empresa (*Visual Analytics*) y deben informar de la evolución correcta de los indicadores o métricas fundamentales de una organización o empresa, de sus departamentos o de sus áreas de negocio.

En síntesis, un cuadro de mando debe ayudar a conseguir los objetivos de la Inteligencia y de la Analítica de Negocios: transformar los datos en información, la información en conocimiento, y el conocimiento debe facilitar la tarea de la toma de decisiones a los empleados, según su nivel jerárquico dentro de la organización.

7.8.2 CARACTERÍSTICAS DE UN CUADRO DE MANDO

La información de un cuadro de mando para ayudar en la toma de decisiones debe constar de diferentes elementos: métricas y KPI, listas, gráficos, tablas, diagramas, diagramas de Venn, alertas visuales y automáticas, mapas, infografías, etcétera, que se han de combinar de manera precisa y uniforme; es importante que cada cuadro de mando disponga de un menú de navegación y en su caso, si fuera de interés, una infografía. Esta información se ha de basar, fundamentalmente, en los indicadores clave de rendimiento (KPI) del negocio y/o de la empresa, así como en las tendencias de negocio.

La creación de un *dashboard* deberá tener presente, en consecuencia, las siguientes características:

- **Indicadores clave de rendimiento KPI adecuados.** Se ha de realizar una búsqueda exhaustiva de los indicadores (palabras clave) que añadan valor al negocio.
- **Presentación visual.** Los cuadros, gráficos, tablas y restante información han de ser ilustrativos, claros y bien estructurados.

- **Datos comprensibles y accionables.** Los datos deben estar contextualizados para su comparación e interpretación, de modo que permitan establecer valoraciones útiles.
- **Personalizado.** Debe ser acorde a los objetivos específicos de cada empresa. Un cuadro de mando no es estándar para todas las empresas y estrategias funcionales. Se ha de elaborar una presentación a medida, en función de los objetivos.

De forma genérica, un *dashboard* engloba a varias herramientas que muestran información relevante para la empresa a través de una serie de indicadores de rendimiento, también denominados **KPI** (*Key Performance Indicators*).

Los cuadros de mando ofrecen una visualización intuitiva de los indicadores empresariales para aquellos perfiles de usuarios responsables de monitorizar, controlar y gestionar los procesos de negocio, y pueden ser personalizados para cada área de competencia (finanzas, ventas, *marketing*, RRHH, TI, etcétera).

Los **cuadros de mando** son herramientas de **administración del rendimiento empresarial (EPM)** que se presentan ante los usuarios como una visualización de los indicadores empresariales. Permiten monitorizar, controlar y gestionar los procesos de una organización a través de códigos semafóricos, que establecen alertas para disponer de una visión completa del rendimiento de la compañía. Los cuadros de mando de gestión, con sus capacidades de alerta, proporcionan una gran ventaja y muchos beneficios. Los indicadores de cumplimiento, evaluación, eficiencia y eficacia contenidos en ellos ofrecen una visión completa de la organización y su rendimiento, permitiendo comprobar, por ejemplo, si la actividad diaria está alineada con la estrategia corporativa o interpretar lo que está ocurriendo y saber si debemos tomar medidas de mejora. Según su función, se pueden clasificar en:

- **Cuadro de mando operacional.** Se utiliza para monitorizar procesos, eventos y actividades específicas a medida que se producen: ayuda en la ejecución de procesos.
- **Cuadro de mando táctico.** Mide y analiza el desempeño o rendimiento de proyectos específicos, procesos y actividades. Muestra información que ayuda a controlar procesos.
- **Cuadro de mando estratégico (*scorecards*).** Normalmente, se despliega de un modo descendente para revisar el progreso que conduce a conseguir los objetivos del negocio. Interviene en la gestión del proceso para la consecución final de objetivos.

Presentación de un cuadro de mando

El cuadro de mando ha de cumplir con la misión de monitorear el rendimiento de la organización y su progreso, por estas razones es necesario que puedan acceder a las métricas en tiempo real con el objeto de poder informar a los usuarios de aquello que están haciendo.



Figura 7.11. Presentación de un cuadro de mando (*dashboard*)

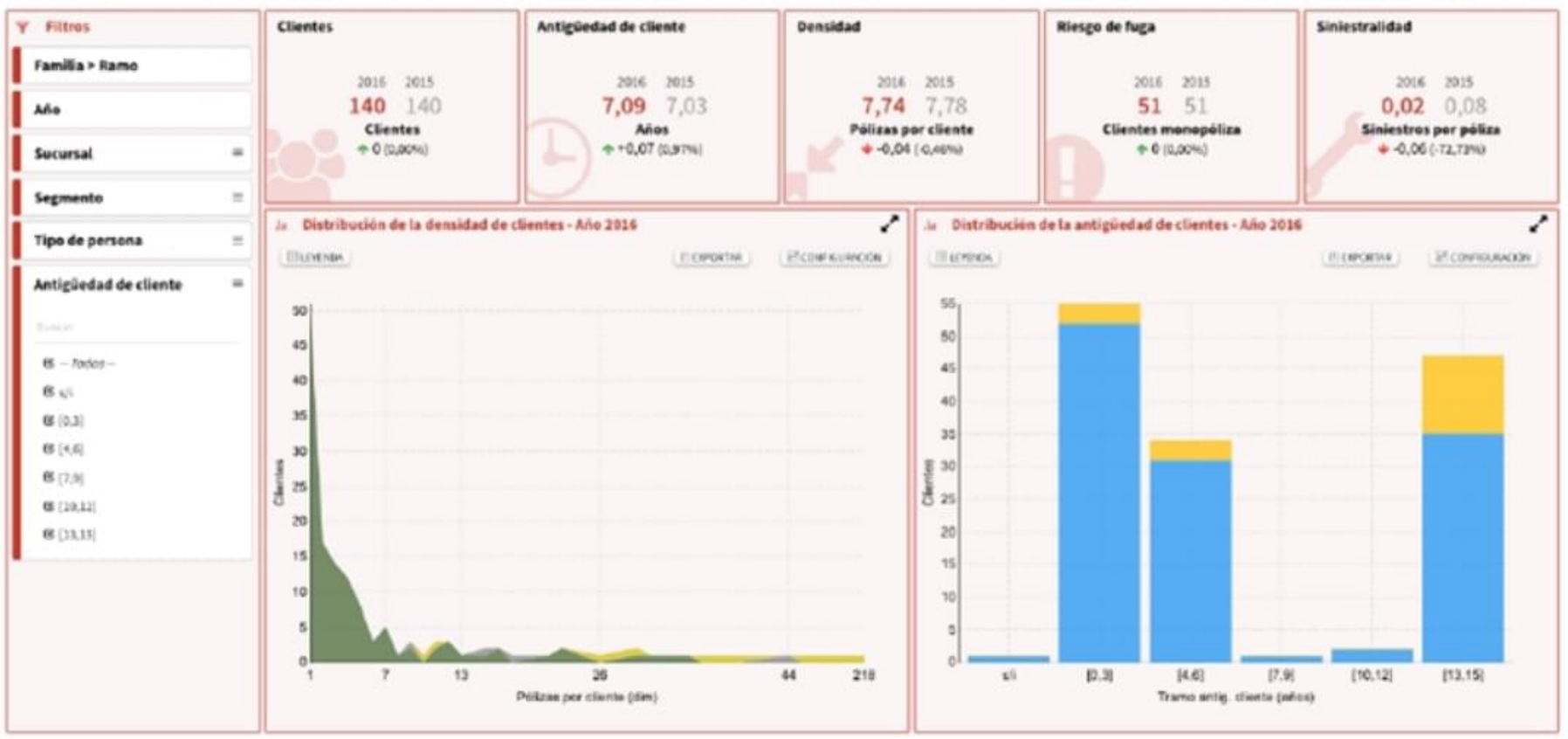


Figura 7.12. Cuadro de mando empresarial. Fuente: Red.es/Iniciativa Aporta. *Visualización de datos*, noviembre 2016

http://datos.gob.es/es/documentacion/visualizacion-de-datos-definicion-tecnologia-y-herramientas.
http://datos.gob.es/sites/default/files/doc/file/informe_herramientas_visualizacion.pdf

Componentes de un cuadro de mando

La mayoría de los cuadros de mando se basan en una serie de gráficos, tablas, medidores u otros indicadores visuales que el usuario ha seleccionado para supervisar; la personalización es importante a la hora de maximizar lo esencial. Los indicadores pueden ser de importancia operativa o estratégica. Así, pues, los componentes de un cuadro de mando pueden ser:

- Tablas (estática, dinámica o herramienta de análisis como OLAP)
- Gráficos
- Mapas
- Métricas (fundamentalmente KPI)
- Listas (de métricas y KPI)
- Menús de navegación
- Alertas (informan del cambio de estado de los datos e indicadores; es convenientes que sean visuales y automáticas al estilo de las alertas de las apps de los teléfonos inteligentes).

Creación de un cuadro de mando

Una vez definidos los componentes y la presentación a utilizar en un cuadro de mando, el proceso siguiente es su creación. Existen numerosas herramientas (ver apartado 6.8) tanto en el mercado del software propietario como en el de software o código abierto.

La hoja de cálculo Excel de Microsoft ha sido la herramienta tradicional y todavía sigue siendo la herramienta imprescindible en organizaciones y empresas. El diseño de cuadros de mando[4] utilizando tablas dinámicas con Excel es una decisión muy acertada.

Curto (2016)[5] propone un proceso iterativo que combina varios pasos:

1. Identificar la necesidad de negocio y los potenciales usuarios del cuadro de mando.
2. Elegir los datos que se van a mostrar en el cuadro de mando.
3. Elegir el formato de presentación conjuntamente.
4. Combinar datos y presentación
5. Planificar la interactividad del usuario.
6. Implementación del cuadro de mando

¿Cómo se utilizan los cuadros de mando en tiempo real?

Cada día, sensores y dispositivos generan cantidades ingentes de datos: dispositivos GPS en vehículos, objetos y personas, sensores de monitorización del medio ambiente, *feeds* de video en directo, sensores de velocidad en autovías,

feeds de redes sociales, etcétera. Esto significa que tenemos más fuentes de datos valiosos, conocidos como **datos en tiempo real**. Sólo recientemente ha surgido la tecnología necesaria para incorporar estos datos en tiempo real a aplicaciones de SIG.

Las funciones de SIG en tiempo real de la plataforma ArcGIS han transformado el modo de utilizar la información en una situación específica. Los cuadros de mando en tiempo real proporcionan vistas útiles de las operaciones diarias de las organizaciones, con lo cual, quienes toman las decisiones y las partes interesadas, disponen de la información más reciente, necesaria para impulsar las ideas y estrategias actuales y futuras. Los *dashboards* responden a preguntas como: ¿Qué sucede ahora mismo? ¿Dónde sucede? ¿A quién afecta? ¿Qué activos están disponibles? ¿Dónde está mi gente?

Algunas aplicaciones de los cuadros de mando en tiempo real

- Los gobiernos locales utilizan la información en tiempo real para gestionar operaciones, tales como el seguimiento y la monitorización de las máquinas quitanieves y los camiones de la basura.
- Las empresas de suministro monitorizan servicios públicos para los consumidores, tales como el agua, las aguas residuales y la electricidad.
- Los departamentos de transporte realizan el seguimiento de autobuses y trenes y monitorizan el tráfico, las condiciones de las carreteras y los incidentes.
- Las autoridades aeroportuarias y las agencias de la aviación realizan el seguimiento y la monitorización del tráfico aéreo en todo el mundo.
- Las empresas de petróleo y gas realizan la monitorización de camiones cisterna y equipos sobre el terreno.
- Las agencias de seguridad pública monitorizan los crímenes mientras suceden, además de las llamadas entrantes en los números de emergencias.
- Las agencias federales, como la Federal Emergency Management Agency (FEMA), US Geological Survey (USGS), National Oceanographic and Atmospheric Administration (NOAA) y la Environmental Protection Agency (EPA), recopilan cantidades ingentes de información sobre el medio ambiente. Monitorizan el tiempo, la calidad del aire y el agua, las inundaciones, los terremotos y los incendios forestales.
- Las empresas utilizan los *feeds* de las redes sociales (como Twitter) en tiempo real para recopilar comentarios y monitorizar la reacción social ante temas específicos de interés.
- Las personas utilizan *smartphones*, *smartwatches*, sensores inteligentes, identificadores por radiofrecuencia (RFID), pulseras de *fitness*, etcétera, para capturar y visualizar información, sobre todo acerca de lo que hacen (Internet de las cosas).

- Las agencias de gestión de emergencias monitorizan la seguridad pública en grandes eventos, tales como maratones o los Juegos Olímpicos.

El concepto de tiempo real

Los datos en tiempo real están tan actualizados como la fuente de datos que los actualiza, tanto si esos datos se actualizan cada segundo como si lo hacen cada minuto, hora o día. El tiempo real para una organización podría no serlo para otra, según el tipo de situaciones que se monitoricen.

El tiempo real es un concepto que, por lo general, se refiere al conocimiento de eventos a la misma velocidad que se desarrollan o en el mismo momento en que lo hacen (sin un retardo significativo). Generalmente, esto se confunde con la frecuencia o el intervalo entre eventos, lo cual es, básicamente, con qué frecuencia se actualizan los eventos. El intervalo de actualización, o la frecuencia se relaciona con el término **resolución temporal**, que puede variar de una aplicación a otra.

Por ejemplo, la mayoría de los sistemas de monitorización de aeronaves proporcionan una o dos actualizaciones cada segundo, mientras que la actualización de las condiciones meteorológicas puede hacerse una vez cada hora. Para monitorizar sus redes, los suministradores de energía utilizan sistemas, también llamados SCADA (Supervisión, Control y Adquisición de Datos), que muestrean datos de tensión, flujo, presión y otros parámetros de dispositivos analógicos a frecuencias muy elevadas (por ejemplo, 50 Hz). Todo esto puede requerir gran cantidad de recursos, como el ancho de banda de la red, la memoria del sistema y el volumen de almacenamiento.



Figura 7.13. Tablero de control (cuadro de mando)

CASO DE ESTUDIO: Google Data Studio 360

A finales del año 2016, y ya en pleno funcionamiento en 2018, Google presentó una herramienta de *dashboard* integrada en su herramienta de analítica web, Google Analytics 360. Tiene una versión gratuita (*freemium*) que tiene todas las

características de las versiones *premium* y de empresa, excepto el límite del número de informes (5) que se pueden crear y el tipo de asistencia técnica (correo electrónico o asistencia telefónica).

Google Data Studio 360 —según el propio Google— nació con el objetivo de permitir a cualquier persona, de una forma muy sencilla, la posibilidad de crear cuadros de mando con un potente diseño visual, mediante la conexión de la herramienta con una multitud de fuentes de información.



Figura 7.14 Cuadro de mando de Google Data Studio

Las principales aportaciones que ofrece Google Data Studio[6] son:

- **Ahorro de tiempo** y recursos en la creación de nuestros informes (*report*), lo que nos permitirá centrarnos más aún en el análisis de los datos.
- Poder combinar y ofrecer en **un solo cuadro de mando** (*dashboard*) todas las herramientas del día a día de nuestros clientes.
- **Automatizar los informes** para poder aportar conocimiento (ideas) de una manera más rápida.
- Facilitar a los responsables de *marketing* de las empresas el **reporte a otros departamentos,** o a escalafones superiores, de un *dashboard ad hoc* para cada caso.



Figura 7.15 Tablero de control de Youtube

Dashboard de Google Analytics

Google define el potencial de la herramienta de visualización de datos en tres características:

- **Conectar.** Para poder crear un cuadro de mando se necesitan datos que proporcionen información a Google Data Studio 360 y, para ello, permite conectar con múltiples fuentes de información para generar nuestros cuadros de mandos (todos los datos procedentes de todas las herramientas de Google, además otras como MySQL y Cloud SQL).
- **Visualizar.** Representación de datos muy visual y atractiva, que permite tener una visión 360 de nuestro negocio.
- **Compartir/Colaborar.** Nos permite la posibilidad de compartir informes y colaborar con nuestros compañeros/clientes de forma muy sencilla.



Figura 7.16 Cuadro de mando (*Dashboard*) Google Analytics

Las opciones de Google Data Studio 360 son:

- **Informes** (creación y generación de informes).
- **Fuentes de datos** (extracción de datos de numerosas fuentes de datos propias de Google y externas).

7.9 NARRATIVA DE DATOS (*DATA STORYTELLING*)

La visualización de datos permite recibir de forma gráfica los datos en bruto (fríos) que, pese a los numerosos formatos y técnicas empleadas, a veces, tienen dificultad para transmitir el "conocimiento" o "emociones" (*insights*) implícitos en dichos datos. La narración de datos es una nueva técnica que está emergiendo y que permite transformar los datos en un discurso o narración visual. La narrativa de datos combina datos y ayuda a explicar lo que está sucediendo en ellos y las razones de su importancia en casos determinados. Es una técnica que ha ido naciendo a la vez que se expandía el concepto de periodismo de datos, y se ha ido trasladando a la visualización de datos para profesionales, esencialmente desde la creciente implantación del análisis de *Big Data*.

Data Storytelling (narración de datos) se ha convertido en un método de contar una historia (narrativa o narración) de datos apoyados en gráficos, mapas, infografías, etcétera, de modo que la típica narración de historias se ha ido adaptando a la narración de datos. De este modo, la narración de datos ayuda a contar historias que permiten explicar mejor las cifras, los resultados o los hechos, facilitando la creación de una historia que permite explicar mejor los datos para ayudar en la toma de decisiones.

La capacidad de contar historias con datos es una tendencia muy importante para ayudar a las técnicas de visualización de datos para profesionales, ante el creciente volumen de datos y la necesidad de tomar decisiones basadas en todo lo que ellos representan. Una visualización efectiva de datos es una aportación considerable a la hora de comunicar las conclusiones y resultados del análisis de datos. Es una nueva técnica de visualización, donde se pasa de la tendencia tradicional de mostrar o presentar datos a la nueva tendencia de contar historias sobre esos datos: se pasa de la narración de historias a la narración de datos.

Definición de narración de datos

De un modo sencillo, la narración de datos es una visualización de datos muy eficaz, que permite la creación de gráficos de datos visualmente muy atractivos. Es una narración estructurada con el objetivo de comunicar información sobre los datos e implica una combinación de tres elementos clave: datos, visualización y narrativa (Dykes, 2016).

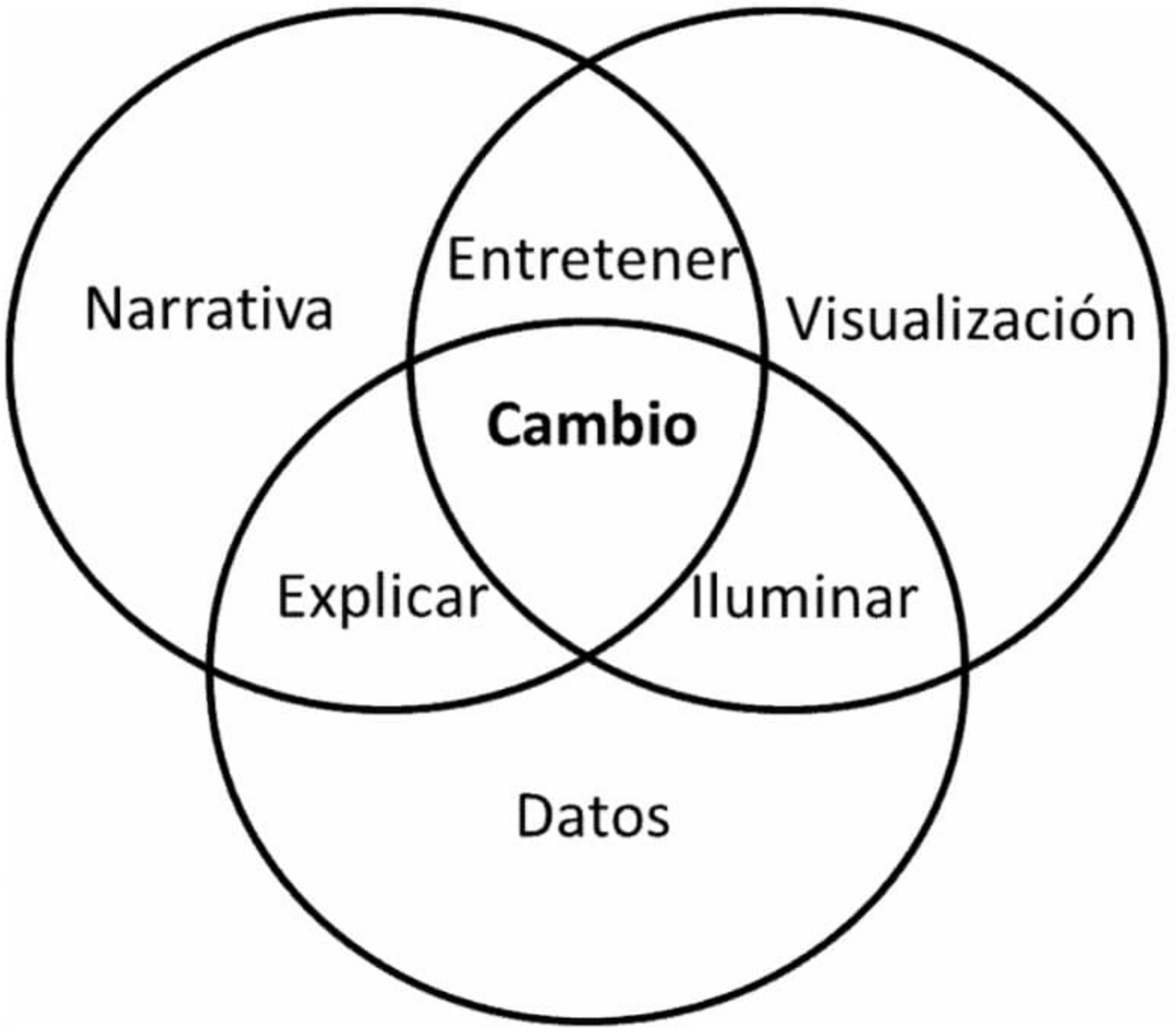


Figura 7.17 Narración de datos

The Data Warehousing Institute[7] —una de las fuentes de información más fiables en análisis de datos— define *Data Storytelling* como: "la práctica de construir una narrativa alrededor de un conjunto de datos y sus visualizaciones correspondientes para ayudar a transmitir el significado de esos datos de un modo potente y convincente". El portal *TechTarget*[8] (una referencia universal en temas de tecnologías de la información en general, e Inteligencia de Negocios y Analítica de Datos en particular) define *Data Storytelling* como: "el proceso de traducción del análisis de datos en términos simples con el objetivo de influir en una decisión o acción de negocios".

Por último, Ton Davenport[9] una de las autoridades más reconocidas en analítica y en *Big Data*, define la narrativa de datos como "la forma de simplificar y dar sentido a un mundo complejo; suministra contexto, la perspectiva, la interpretación en otras propiedades. Todas las cosas que hacen significativo un análisis de datos además de más relevante e interesante".

Knaflic, una prestigiosa analista de datos y especialista en *Data Storytelling,* en su obra de referencia (Knaflic, 2017)[10] considera que la narrativa de datos "es la capacidad de contar historias con datos y su importancia aumenta cada día en el mundo, en el que cada vez hay más datos y más necesidad de tomar decisiones basadas en ellos". Knaflic destaca en su obra cómo mejorar las habilidades de una persona mediante la creación de una historia con datos atractiva, informativa y convincente.

Herramientas de *Data Storytelling*

Los proveedores de herramientas de visualización de mayor impacto en el mercado, tales como SAS, Qlik y Tableau, ofrecen soluciones de narrativas de datos, bien específicas o bien integradas dentro de sus herramientas de visualización.

Un caso muy implantado es *Qlik Sense*, una herramienta de descubrimiento de datos y soporte a la toma de decisiones mediante historia de datos. La narración de datos en Qlik Sense permite combinar informes, presentaciones y técnicas de análisis exploratorio para crear y compartir una experiencia de colaboración. En la página web de Qlik Sense, se resalta que esta herramienta ofrece un mecanismo que toma capturas de imágenes estáticas de los datos (informes) e inserta esas capturas en historias estructuradas compuestas de una o más diapositivas. Qlik Sense señala la importancia de la narración de historias de manera interactiva y atractiva durante la presentación de la historia de datos ante la audiencia, según los criterios de la herramienta de Qlik. Una historia de datos, según Qlik, es una combinación efectiva de datos, narrativa y visualización que trata de explicar, iluminar y comprometerse con el significado de los datos.

Tableau, uno de los líderes mundiales en visualización de datos, dispone de herramientas específicas para el tratamiento de historias y narraciones de

datos[11]. *Tableau Desktop* es la herramienta profesional que puede probarse durante un periodo de 14 días. Además de esta herramienta profesional, Tableau ofrece una versión gratuita, *Tableau Public*, idónea para prácticas de visualización de datos, y que se puede utilizar para prácticas educativas y de aprendizaje. Lógicamente, si desea tener de manera profesional todas sus propiedades a nivel de empresa, deberá pensar en la solución de *software* propietario (*Tableau Desktop*).



Figura 7.18. Narración de datos

Fuente: aulaPlaneta (Área de Educación del Grupo Planeta)
http://www.aulaplaneta.com/2015/11/07/infografias/storytelling-con-las-tic/

7.10 CUADRO DE MANDO INTEGRAL (CMI) O *BALANCED SCORECARD*

Disponible online.

7.11 HERRAMIENTAS DE VISUALIZACIÓN DE DATOS

Existen numerosas herramientas de visualización de datos proporcionadas por numerosos fabricantes. Dada la popularidad que han alcanzado para la ayuda en la toma de decisiones, casi todos los grandes fabricantes de *software* tienen soluciones independientes o integradas de visualización de datos.

El gran reto del análisis de datos es conseguir ser capaces de recolectar, relacionar, procesar, analizar y extraer el verdadero valor, de forma que la información se pueda presentar de manera clara, concisa y comprensible. En esencia, las herramientas han de permitir explorar y analizar visualmente los datos.

Herramientas *Big Data* para visualizar datos

Además de saber cómo almacenar datos, procesarlos y analizarlos, ser un **experto en *Big Data*** conlleva saber comunicar la información que esos datos, tras su clasificación y estudio, nos han proporcionado. Para ello, es fundamental **"representar" los datos** en un contexto familiar y efectivo, que facilite la labor de interpretarlos, visualizándolos de manera sencilla y asequible.

En el mercado hay **herramientas de visualización de datos** asequibles, tanto para desarrolladores o diseñadores como para personal menos técnico. La mayoría cuenta con versiones de pago y versiones gratuitas y ofrece gráficos optimizados para su uso en redes sociales. Entre las más populares se encuentran Tableau, Weave, Datawrappper, Gephi, Infogram, Many Eyes, Piktochart, NodeXL, Chartblocks, d3, Thinglink, Axiis, QuickView y Google Fusion Tables.

El banco español BBVA —uno de los más innovadores— ofrece un gran número de portales web donde publica de modo gratuito una gran cantidad de información solvente sobre tecnologías de la información. Una de las publicaciones más reconocidas sobre visualización de datos es un *ebook*[12], relativo a herramientas de visualización, donde explica los conceptos fundamentales y recomienda una serie de herramientas, entre las que destacan:

- Google Fusion Tables
- CartoDB ahora Carto
- Tableau Public (herramienta gratuita, aunque existe una versión de pago; sólo exige un registro previo: `https://www.tableau.com/es-es/products/desktop/download`)
- ICharts . Versión gratuita (Basic) y dos de pago Platinum y Enterprise.
- Smart Data Report. Análisis de datos gratuitos y generación de informes

OTROS proveedores

- SAS Visual Analytics
- QlikView
- Gephi
- R (con su proyecto RStudio)
- Infogr.am
- ESRI (soluciones basadas en sistemas de información geográfica, GIS)
- Pentaho
- Jaspersoft
- ManyEyes
- Quadrigam
- Nodebox
- D3.js (D3)
- Cyte.com
- Caspio.com
- Zoho.com
- Siense.com
- Rapahael

7.12 RESUMEN

La visualización de datos es una de las etapas clave en los sistemas de Inteligencia de Negocios, cuyo objetivo principal es mostrar en forma visual los resultados de las etapas de análisis de datos. Uno de los líderes mundiales en analítica, SAS, define la visualización de datos como: "la exploración visual e interactiva y la correspondiente representación gráfica de datos de cualquier dimensión, naturaleza (estructurados y no estructurados) y origen".

- El crecimiento exponencial de datos y las técnicas de visualización de datos han dado origen al nacimiento de una nueva técnica de análisis: Analítica Visual.
- Las técnicas de visualización de datos son numerosas: gráficos, tablas, mapas, infografías, mapas mentales, nubes de etiquetas, narrativas o historias de datos (*Data Storytelling*), etcétera.

- Herramientas clave en la visualización de datos: reportes o informes y consultas.
- Los cuadros de mando o tableros de control (*dashboards*) son una de las herramientas más populares y utilizadas en la ayuda de toma de decisiones para profesionales y directivos.
- Un cuadro de mando es una herramienta que visualiza, en una única pantalla, gráficos y diagramas, tablas, métricas, indicadores KPI, etcétera, para ayudar en la gestión de una empresa.
- La narrativa de datos (*Data Storytelling*) es una de las técnicas de visualización de datos más innovadoras. Es un método de contar una narración o historia de datos apoyada en gráficos, mapas, infografías, etcétera.
- El Cuadro de Mando Integral (*Balanced Scorecard*) es un sistema de gestión que permite traducir los objetivos estratégicos de una organización en resultados. Es una metodología de planificación estratégica basada en indicadores (métricas y KPI), creada por los profesores de la Universidad de Hardvard, Robert S. Kaplan y Daniel P. Norton, que relaciona factores medibles de procesos con la consecución de objetivos estratégicos.
- Las perspectivas (puntos de vista) son las dimensiones clave de la empresa y permiten relacionar los objetivos de todas las áreas o departamentos. Las cuatro perspectivas del CMI son: finanzas, clientes, aprendizaje y crecimiento, procesos internos.

NOTAS

[1] *Bernard Marr. Big Data. La utilización del big data, el análisis y los parámetros SMART para tomar mejores decisiones y aumentar el rendimiento. TEELL Editorial, 2016.*

[2] `https://datateam.com.mx/reportes`

[3] *No existe unanimidad —en la bibliografía de Inteligencia de Negocios (en español) y de los proveedores de soluciones— en la traducción del término dashboard, por esta razón, como sucede con otros términos de origen inglés, se suele utilizar en la biblioteca técnica el nombre original de dashboard. Sin embargo, existe un mayor uso del término cuadro de mando, pero también en algunos países latinoamericanos se utiliza tablero de control o cuadro de control.*

MUÑIZ, Luis (2017) Diseñar cuadros de mando con Excel utilizando tablas dinámicas, Barcelona, Editorial Profit.

[5] *CURTO, Josep (2016). Introducción al Business Intelligence. Barcelona: Editorial UOC.*

[6] https://loogic.com/data-studio-el-dashboard-del-2017/

https://www.google.com/analytics/data-studio/

[7] *TDWI Glossary. What is data storytelling?https://tdwi.org/portals/what-is-data-storytelling-definition.aspx*

[8] *Margarert Rouse.* http://searchcio.techtarget.com/definition/data-storytelling

[9] *Tom Davenport. Why Data Storytelling Is So Important. 22 de enero, 2015.*

http://www.tomdavenport.com/blogs-articles/deloitte-university-press/;

https://dupress.deloitte.com/dup-us-en/topics/analytics/data-driven-storytelling.html

[10] *Cole Nussbaumer Knaflic (2017). Storytelling con datos. Visualización de datos para profesionales. Madrid: Anaya multimedia / Willey*

[11] *Mackinlay, Jock et al. Relato de historias con datos. Tableau.*
https://www.tableau.com/sites/default/files/media/Whitepapers/whitepaper_datastorytelling_es_0.pdf

[12] *BBVA Open4U. ebook "Herramientas de visualización".*
https://www.bbva.com/es/wp-content/uploads/2015/08/bbva-open4u-ebook-herramientas-visualizacion-datos.pdf

PARTE III
ANALÍTICA DE NEGOCIOS Y ANALÍTICA DE DATOS

CAPÍTULO 8
MINERÍA DE DATOS

CONTENIDO

8.1 Introducción
8.2 Minería de Datos: conceptos, definiciones y aplicaciones
8.3 Aplicaciones de la Minería de Datos
8.4 Proceso de descubrimiento del conocimiento: KDD
8.5 Proceso de Minería de Datos: metodología CRISP-DM
8.6 Proceso de Minería de Datos: metodología SEMMA
8.7 Modelos, algoritmos y técnicas de Minería de Datos
8.8 Relaciones de la Minería de Datos con otras disciplinas: de Big Data a Data Science
8.9 Herramientas de software de Minería de Datos
8.10 RESUMEN

OBJETIVOS
• Conocer y comprender los conceptos fundamentales de Minería de Datos.
• Aprender las técnicas fundamentales de la Minería de Datos.
• Conocer aplicaciones prácticas de Minería de Datos.
• Aprender el proceso de Minería de Datos y metodologías más utilizadas.
• Conocer los diferentes tipos de Minería de Datos.
• Conocer y comprender las distintas etapas o fases que componen el proceso de Minería de Datos.
• Conocer las técnicas y los métodos más empleados en Minería de Datos.
• Conocer las herramientas de software más populares empleadas en Minería de Datos.

8.1 INTRODUCCIÓN

Minería de Datos es un conjunto de técnicas que se utilizan para optimizar el desarrollo de la Inteligencia de Negocios a partir de los datos que una organización recolecta, organiza y almacena, para, posteriormente, realizar el análisis de datos y contribuir del modo más eficiente posible en la toma de decisiones.

Existen grandes volúmenes de datos almacenados en las bases de datos, *Data Warehouses* y otros tipos de almacenes de datos como las bases de datos NoSQL, "en memoria" y MPP (bases de datos de procesamiento paralelo masivo) y los modernos repositorios de datos conocidos como Lagos de Datos (*Data Lakes*). Esta ingente cantidad de datos son cada día más ricos y rentables para las organizaciones, pero requieren un análisis eficiente, no siempre fácil, para su conversión en conocimiento para la toma de decisiones.

La Minería de Datos busca descubrir el conocimiento de los datos recopilados y almacenados, de manera oportuna y en una forma escalable (para pasar de modo eficiente de gigabytes – terabytes – petabytes - exabytes y, en un futuro no lejano, a zettabytes).

Se desarrollan los conceptos fundamentales de Minería de Datos y las técnicas para el descubrimiento de patrones de datos de interés a partir de datos de aplicaciones diversas. Analizaremos los fundamentos de la Minería de Datos, sus aplicaciones y campos de implantación, así como las técnicas más utilizadas en el desarrollo de herramientas de Minería de Datos eficaces y eficientes. Asimismo, realizaremos un examen práctico de las herramientas de *software* de Minería de Datos más empleadas, tanto de *software* propietario y comercial, como de *software* de código abierto y gratuito.

Las técnicas de Minería de Datos están evolucionando y se están integrando con las técnicas de aprendizaje automático dentro de la Inteligencia Artificial y la **Ciencia de Datos** (*Data Science*). Se hará una introducción a la Ciencia de Datos

como una disciplina que tiene como espina dorsal la Minería de Datos y que se estudiará con mayor profundidad en el capítulo 14.

8.2 MINERÍA DE DATOS: CONCEPTOS, DEFINICIONES Y APLICACIONES

El término Minería de Datos (*Data Mining)* se utiliza para describir el descubrimiento o "minado" (*mining*) del conocimiento a partir de grandes cantidades de datos. La Minería de Datos nació como una analogía de la minería tradicional (minería de oro ode cobre). Por esta razón, como se trataba de minar, buscar conocimiento significativo en ingentes volúmenes de datos, también se ha la denominado a veces como "minería de conocimiento" (*knowledge mining*), "descubrimiento del conocimiento" (*knowledge discovery*) o, incluso, "extracción del conocimiento". Asimismo, en muchas ocasiones, dado que la Minería de Datos recurrirá a realizar en la práctica análisis de datos a través de patrones previamente seleccionados, también se la ha denominado, "análisis de datos", "análisis de patrones de datos" y "arqueología de datos" (Turban 2011 y Sharda 2014).

En este sentido, deseamos destacar un término sinónimo y muy utilizado en los orígenes de la Minería de Datos: **KDD** (*Knowledge Discovery from Data*) o **Descubrimiento de Conocimiento en Bases de Datos**; sin embargo, como veremos posteriormente, el proceso KDD es un término más global y la Minería de Datos es una etapa dentro del proceso global KDD.

Desde un punto de vista técnico, Minería de Datos es el proceso que utiliza técnicas matemáticas, estadísticas y de Inteligencia Artificial para extraer datos, identificar información útil y conocimientos posteriores en forma de patrones. Estos patrones pueden estar en diferentes formatos, tales como reglas de negocio, afinidades, correlaciones, asociación, tendencias o "modelos de predicciones". Una definición muy aceptada de Minería de Datos es: "el proceso no trivial de identificación de patrones de datos almacenados en bases de datos estructuradas y que tiene las propiedades de no triviales, válidos, nuevos (noveles), posteriormente útiles y en última instancia comprensibles, donde los datos se organizan en registros estructurados por variables categóricas, ordinales y continuas" —Fayyad *et al* (1996), citado por Turban (2011)—. El proceso de Minería de Datos implica una serie de etapas iterativas. Más adelante veremos que existen diferentes modelos de proceso de Minería de Datos.

La Minería de Datos es una disciplina en la que confluyen muchas otras disciplinas clásicas en el mundo científico y de negocios: estadística, Inteligencia Artificial, aprendizaje automático o de máquina (*machine learning*), sistemas de información, algoritmos y bases de datos y, recientemente, *Data Science* (Ciencia de Datos) o análisis de grandes volúmenes de datos o *Big Data.*

Data Mining (**Minería de Datos**): Proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten and Frank, 2016).
Knowledge Discovery in Databases — **KDD (Descubrimiento de Conocimiento en Bases de Datos):** Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos (Fayyad et al. 1996).

8.3. APLICACIONES DE LA MINERÍA DE DATOS

La Minería de Datos es una de las herramientas más utilizadas en la actualidad en un gran número de áreas, que van desde aplicaciones empresariales, de negocios, industriales hasta la investigación científica en numerosos campos como la medicina, la bioquímica, meteorología, astronomía, genérica, bioinformática, física, etcétera. Un área muy destacada de aplicaciones de Minería de Datos es la Web y los textos que han dado lugar al nacimiento de minerías de datos específicas, conocidas como Minería Web y Minería de Textos. Es muy difícil, por no decir imposible, enumerar todas las aplicaciones en las que la Minería de Datos juega un papel fundamental. En esta sección tratamos de describir aplicaciones sobresalientes en diversos y variados campos.

Medicina

La Minería de Datos en medicina es una de las aplicaciones más prácticas, debido a que complementa la investigación médica en análisis clínicos y en el trascendental campo de los diagnósticos, entre otras especialidades:

- Identificación de patrones novedosos para mejorar la supervivencia de pacientes con cáncer.
- Predicción de tasas de éxito en trasplantes de órganos a pacientes para desarrollar políticas de donantes/receptores en el tratamiento clínico.
- Genómica. Identificación de los diferentes genes del cromosoma humano.
- Selección de embriones en reproducción artificial.
- Genética. Estudio del ADN
- Descubrimiento de las relaciones entre síntomas y enfermedades, así como entre enfermedades y tratamientos con éxito.
- Identificación de terapias para diferentes enfermedades.
- Estudio de factores de riesgo en diferentes patologías.

- Segmentación de pacientes por grupos afines.
- Gestión hospitalaria y clínica para planificación temporal de habitaciones, quirófanos, salas de consulta, etcétera.

Industria

En el sector industrial son numerosas las aplicaciones y, también, en numerosos campos.

- Fabricación y producción. Predecir fallos de máquinas antes de que ocurran a través del uso de datos de sensores.
- Procesos industriales. Automatización del control de procesos: optimización del rendimiento de forma adaptativa; implementación de programas de mantenimiento predictivo.
- Control de calidad. Identificación de posibles causas de fallos; detección y localización precoz de defectos industriales; detección precisa de productos defectuosos; descubrimiento de patrones noveles para identificar y mejorar la calidad de los productos.

Banca

La Minería de Datos ayuda al sector bancario en numerosas aplicaciones:

- Detección de patrones de uso fraudulento de tarjetas y transacciones de banca en línea (*online*).
- Automatización de los procesos de concesión de préstamos para predecir, con la mayor precisión posible, los morosos más probables.
- Estudio de concesión de tarjetas de crédito.
- Determinación del gasto en tarjetas por segmentación de grupos.
- Identificación de reglas de comportamiento del mercado de valores a partir de los registros históricos de dichos mercados.
- Predicciones de hábitos y patrones de compra en grandes almacenes y en mercados en línea.
- Detección de segmentos de clientes predispuestos a la compra de determinados artículos, bien en el lanzamiento o cuando ya están en el mercado.
- Identificación de clientes fieles y también de fuga de clientes.

Seguros

La industria de seguros también utiliza técnicas de Minería de Datos en numerosas aplicaciones:

- Predicción de clientes propensos a suscribir nuevas pólizas con características especiales.
- Identificación y prevención de pagos de reclamaciones incorrectas y actividades y comportamientos fraudulentos.
- Identificación de grupos de clientes y patrones de riesgo.
- Previsión de gastos médicos de sectores de clientes específicos.
- Identificación de fidelidad de clientes.
- Identificación de fugas de clientes.

Hardware y *software* de computadoras

Un equipo de computación con un *hardware* adecuado y un *software* eficiente aumentará el desempeño del proceso de buscar y analizar información; la Minería de Datos puede ser muy provechosa en el caso de poseer datos sobre sus procesos productivos, datos de seguimiento de clientes, datos externos del mercado, datos sobre la actividad de competidores, etcétera. Aplicaciones concretas pueden ser:

- Predicción de fallos de unidades de disco, memorias *pendrive*, etcétera, antes de que ocurran realmente.
- Identificación y filtrado de contenido web no deseado y mensajes de correo electrónico.
- Detección y prevención de puentes (*bridges*) de seguridad de redes de computadoras.
- Identificación de productos de *software* potencialmente inseguros.

Turismo

En la industria del turismo existe una gran variedad de aplicaciones para hoteles, líneas aéreas, *resorts*, viajes, alquiler de automóviles, trenes, etcétera:

- Predicciones de ventas de diferentes servicios (reserva de asientos en diferentes clases, reserva de habitaciones en hoteles/*resorts*, reserva de autos en compañías de alquiler, etcétera).
- Identificación de los clientes más rentables para proporcionarles mejores servicios (por ejemplo, las tarjetas de fidelización "millas" de los clientes "viajeros frecuentes", a los que se ofrecen beneficios como "prioridad en salas VIP", *upgrade* (subida) de categoría, ofertas especiales en función de la tarjeta de fidelización, etcétera).
- Predicción de ocupación en aviones, trenes, etcétera, dependiendo de rutas viajeras, épocas del año, entre otras.

Gestión de relaciones con los clientes (CRM)

La gestión y administración de relaciones con los clientes es una de las actividades de las empresas donde más aplicaciones tiene la Minería de Datos. El objetivo del CRM es crear relaciones personales con los clientes para facilitar los servicios que se han de proporcionar. CRM tiene dos objetivos principales:

- Retención de clientes mediante la satisfacción del cliente.
- Desarrollo de relaciones con los clientes mediante el conocimiento de los mismos.

Las técnicas de Minería de Datos tienen una gran aplicación, sobre todo en la segmentación de clientes, campañas de *marketing* directo, campañas de *marketing* de contenidos, etcétera.

Así, aplicaciones globales de Minería de Datos clásicas son:

- Segmentación de clientes.
- Campañas de *marketing* directo.
- Análisis de secuencia y la bolsa de la compra.

También existen aplicaciones más específicas:

- Identificar compradores/demandantes de nuevos servicios/productos.
- Comprender las razones para mejorar la retención de clientes.
- Descubrir relaciones en el tiempo entre productos y servicios para maximizar ventas y valores del cliente.
- Identificar los clientes más rentables y sus preferencias para fortalecer las relaciones con ellos y maximizar las ventas.

Deportes

La Minería de Datos también tiene, curiosamente, numerosas y prácticas aplicaciones en el mundo del deporte. En casi todos los deportes se encuentran aplicaciones de minería. Una de las más conocidas en baloncesto, donde es clásica la aplicación de la NBA (National Basketball Association) de Estados Unidos, que desarrolló una aplicación de Minería de Datos para PC (Advanced Scout), que permite descubrir patrones de comportamiento de interés para el desarrollo de los partidos de la NBA.

Otro caso muy conocido es el equipo AC de Milán que —desde hace bastantes años— utiliza un sistema inteligente para prevenir lesiones. El club posee aplicaciones de redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta, de modo que pueda ayudar a seleccionar el fichaje de un posible jugador o a alertar al médico del equipo de una posible lesión. El sistema de Minería de Datos fue creado por la compañía Computer

Associates International y se alimenta de datos de cada jugador, relacionados con su rendimiento, alimentación, respuesta a estímulos externos que se obtienen y analizan cada quince días. El sistema dispone de más de 800 casos registrados, que permiten predecir alguna posible lesión.

8.4 PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO: KDD

La extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases*, KDD), que se refiere al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información [1]. No es un proceso automático, es un proceso iterativo que, exhaustivamente, explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad, y que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos. Las etapas del proceso KDD tradicional de Fayad *et al*[1]de vigencia total se dividen en 5 fases (figura 8.1) y son:

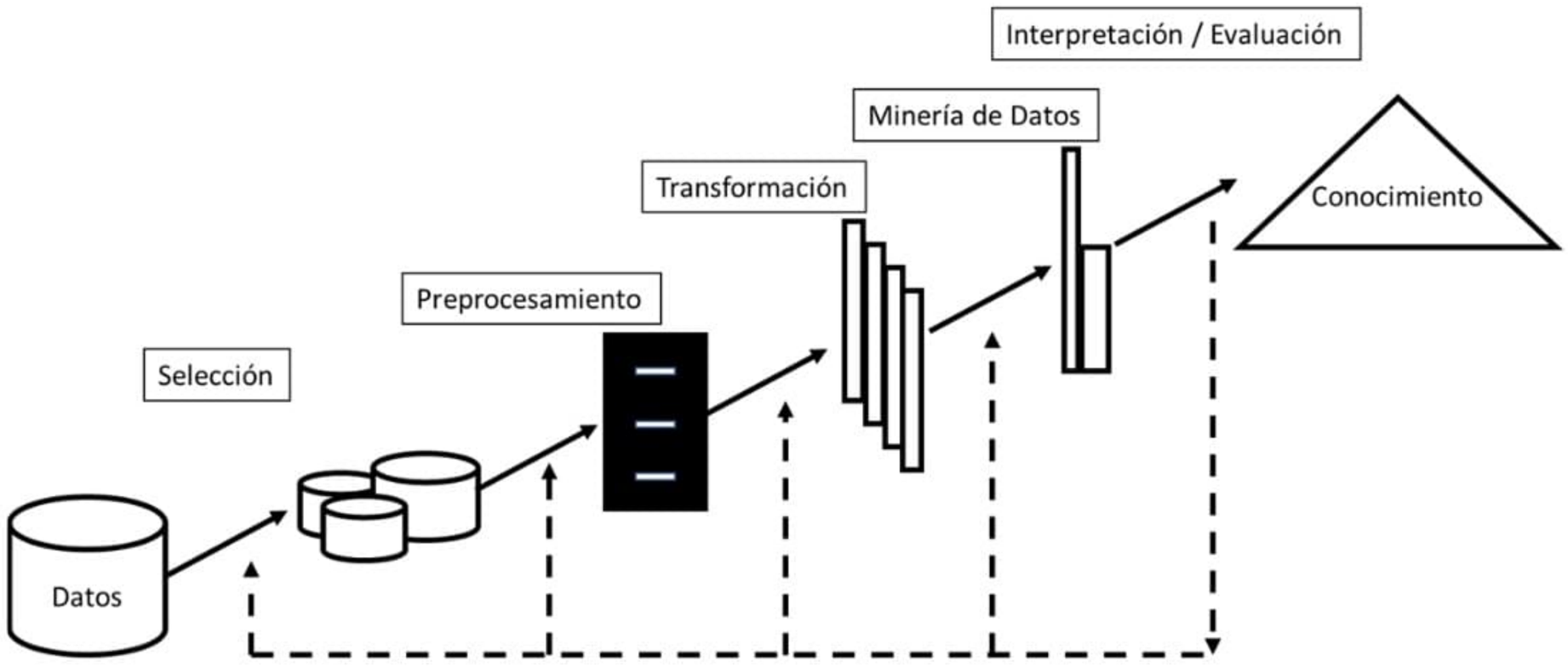


Figura 8.1. Proceso de descubrimiento del conocimiento (KDD). Fuente: (Fayyad 1996)

1. **Selección de datos.** En esta etapa se determinan las fuentes de datos y el tipo de información que se ha de utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde las fuentes de datos.
2. **Preprocesamiento.** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan

diversas estrategias para manejar datos faltantes o ausentes, en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

3. **Transformación.** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes, con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

4. **Minería de Datos.** Es la fase de modelado propiamente dicho, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles, y que están contenidos u "ocultos" en los datos.

5. **Interpretación y Evaluación.** Se identifican los patrones obtenidos, que son realmente interesantes, basándose en algunas medidas, y se realiza una evaluación de los resultados obtenidos.

Además de las fases descritas, frecuentemente se incluye una fase previa de análisis de las necesidades de la organización y definición del problema, en la que se establecen los objetivos de la Minería de Datos. También es usual incluir una etapa final, donde los resultados obtenidos se integran al negocio para la realización de acciones comerciales.

Proceso de *Data Mining* (Minería de Datos)

A grandes rasgos, el proceso de Data Mining puede dividirse en 6 pasos:

1. *Selección del conjunto de datos*: aquí se decide cuáles van a ser las variables objetivos (aquellas que se quieren predecir o inferir), las variables independientes y la selección de registros (datos) que se han de utilizar.

2. *Análisis de las propiedades de los datos:* mediante, por ejemplo, histogramas y/o diagramas de dispersión. Búsqueda de valores atípicos (outliers) y ausencia de datos.

3. Transformación o preprocesamiento del conjunto de datos de entrada: en este paso, se normalizan los datos a una misma escala. También se decide cómo se van a tratar datos faltantes, atípicos o dudosos. Una posibilidad es tratarlos como un tipo de dato especial, o bien se decide descartarlos.

4. *Selección y aplicación de técnicas de Minería de Datos:* se construye un modelo que será utilizado sobre los datos para predecir las clases mediante clasificación o para descubrir grupos similares mediante segmentación

5. *Extracción de conocimiento:* una vez aplicado el paso anterior, se buscan patrones de comportamiento en los valores de las variables del problema o relaciones de asociación entre dichas variables.

6. *Interpretación y evaluación de datos:* el modelo debe ser validado, comprobando que las conclusiones arrojadas son válidas y satisfactorias. Si el modelo final no supera esta evaluación, el proceso puede repetirse desde el principio o a partir de cualquiera de los pasos anteriores.

8.4.1 ETAPAS DE UN PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

Durante la década de los 90, el citado término Descubrimiento del Conocimiento en Bases de Datos fue un sinónimo de Minería de Datos. KDD es un proceso que utiliza métodos de Minería de Datos para encontrar información útil y patrones de datos, al contrario que el concepto ya dado de Minería de Datos, que implica utilizar algoritmos para identificar patrones de datos a través del proceso KDD.

El proceso KDD fue el primer modelo aceptado por la comunidad científica que estableció las etapas principales de un proyecto de explotación de datos. En su origen, el modelo KDD establece que la Minería de Datos es una etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos. Aunque en la bibliografía científica y profesional el término KDD y Minería de Datos son sinónimos, y se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento del conocimiento, desde un punto de vista de proceso de extracción y análisis de datos KDD es un proceso completo que comprende a la Minería de Datos como una de sus etapas.

KDD es un proceso no trivial que identifica patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos (Fayyad et al, 1996).

Etapas del proceso KDD

El proceso de extracción/descubrimiento del conocimiento en bases de datos —Fayyad et al (1996) y Dunham (2003)— consta de las siguientes fases:

- Selección de datos (Integración y recopilación)
- Preprocesamiento de datos (Preprocesamiento)
- Limpieza de datos
- Integración de datos
- Transformación y reducción de datos. Selección de la técnica de MD
- Interpretación/Evaluación de los patrones (Difusión y utilización del nuevo conocimiento)

En el proceso KDD, uno de los elementos más importantes que se han de considerar es el usuario, ya que es quien determina el dominio de la aplicación, es decir, decide cómo y qué datos se utilizarán en el proceso; el usuario debe

entender y participar activamente en el desarrollo del mismo. Las etapas en el proceso global del KDD no están claramente diferenciadas por ser un proceso iterativo e interactivo con el usuario experto.

Diferencias entre KDD y Minería de Datos
La Minería de Datos es una etapa particular del proceso KDD, donde la Minería de Datos aplica algoritmos específicos o técnicas específicas para la extracción de patrones de los datos (Fayyad et al, 1996), La Minería de Datos es la etapa más importante del KDD, que integra los procesos de aprendizaje y métodos estadísticos para la obtención de hipótesis de patrones y modelos.
De modo informal se asocia Minería de Datos con KDD.

1. Selección (integración y recopilación)

Es necesario familiarizarse con el dominio del problema y obtener conocimiento para acotar el espacio de posibles soluciones. En esta fase es preciso considerar todos los datos con los diferentes departamentos de la entidad.

Los datos pueden estar en bases de datos o en un *Data Warehouse* y será preciso unificar la información procedente de diferentes fuentes:

- Información interna (hojas de cálculo, informes, bases de datos).
- Otras bases de datos (empresas e industrias afines, adquiridas en el mercado).
- Estudios publicados (propios o externos) sobre demografía, estadísticos, catálogos, páginas web.

2. Preprocesamiento de datos

El objetivo de esta fase es seleccionar el conjunto de datos adecuado para el resto del proceso de KDD. En esta fase se busca la calidad del conocimiento descubierto, que dependerá de los algoritmos y, naturalmente, de la calidad de los datos minados.

Las actividades de esta etapa son:

- Limpieza de datos (*data cleaning*).
- Transformación de los datos.
- Reducción de la dimensionalidad.

La **limpieza de datos** requiere la selección de datos que pueden ser de diferentes categorías:

- **Datos perdidos** (*missing values*), que pueden conducir a resultados poco precisos. Se requiere analizar el motivo de la pérdida y es preciso

dar soluciones —rellenado manual, ignorarlos, eliminar el campo de la tabla (fila/columna), usar un valor especial (por ejemplo, desconocido), inferirlos utilizando técnicas estadísticas, etcétera—.

- **Datos anómalos** (*outliers*) e inconsistencias. Valores que no se ajustan al comportamiento general de los datos o bien datos erróneos o correctos pero distintos a los restantes. Se requiere en primer lugar su identificación y, a continuación, en función del problema en cuestión, se tratarán como valores perdidos y se obtendrá información de ellos.
- **Inconsistencias**. Datos inconsistentes, registros duplicados, etcétera. Normalmente, se tratan en la elaboración de la base de datos o del almacén de datos.

La transformación de los datos requiere:

- Construcción de nuevos atributos (se verán posteriormente), aplicando alguna operación a los atributos originales.
- Discretización consistente en pasar atributos continuos o discretos a casos discretos manejables o categóricos.

La reducción de la dimensionalidad requiere:

- Reducción de casos y filas; selección de variables con la elección del conjunto de atributos adecuado para la tarea específica por realizar.

3. Técnica de Minería de Datos

El objetivo de esta actividad es producir nuevo conocimiento que pueda utilizar el usuario. Se ha de construir un modelo, basado en los datos recopilados, que sea una descripción de los patrones y relaciones entre los datos con los que se puedan explicar situaciones pasadas, entender mejor los datos y hacer predicciones.

Las decisiones que se han de tomar dependerán de:

- Tipo de conocimiento buscado (predictivo, descriptivo).
- Técnica más adecuada (clasificación, regresión, agrupamiento *clustering*, asociaciones).
- Tipo de modelo *(*clasificación, reglas, árboles de decisión).
- Incertidumbre en el modelo resultante (certeza, probabilidad, lógica difusa).
- Tipo de algoritmo más adecuado.

4. Evaluación, interpretación y presentación de resultados

En esta fase de evaluación se requiere establecer cuáles son los modelos más válidos, teniendo presentes los criterios deducidos de los patrones descubiertos que han de ser precisos, comprensibles e interesantes (es decir, útiles y novedosos). Se requieren técnicas de evaluación, especialmente entrenamiento y test o pruebas. Es preciso adoptar medidas de evaluación de los modelos, que dependerá de la tarea por realizar, clasificación, regresión, agrupamiento o asociación.

Por último, se han de interpretar los mejores modelos (mediante la visualización, ventajas colaterales, posibilidad de integración), que ayuden a la selección de los modelos finales y a presentar los resultados.

5. Difusión y utilización del nuevo conocimiento

Una vez construido y validado el modelo, se puede utilizar para recomendar acciones que se han de seguir y para aplicar el modelo a diferentes conjuntos de datos. Se requieren tareas de difusión, mediante la elaboración de informes para su distribución, utilización del nuevo conocimiento de modo independiente e incorporación a sistemas ya existentes en la organización.

8.5 PROCESO DE MINERÍA DE DATOS: METODOLOGÍA CRISP-DM

El proceso o metodología **CRISP-DM** —*Cross Industry Standard Process for Data Mining* (www.crisp-dm.org)— fue propuesto en la segunda mitad de la década de los 90 por un consorcio europeo de empresas —entre sus fundadores se cuentan Daimler, Chrysler, SPSS y NCR— y se ha convertido en una metodología de Minería de Datos abierta y no propietaria. Esta metodología presenta la ventaja de haber sido diseñada y construida sobre la base de experiencias reales, y no de modo teórico, y por empresas de gran prestigio en la industria y empresas de tecnologías de la información. La figura 8.2 ilustra el proceso propuesto en CRISP-DM, que consta de etapas que comienzan con un buen conocimiento del negocio y el dominio de la aplicación y termina en el despliegue de la solución que cumpla la necesidad específica del negocio. En origen, el proceso es secuencial (avances), por naturaleza, aunque en algunas etapas se contemplan flujos de datos bidireccionales, lo que significa que algunas fases permitirán revisar parcialmente o en su totalidad las fases anteriores. La Minería de Datos se apoya en el conocimiento y, en consecuencia, es muy importante considerar la experiencia y la experimentación.

CRISP-DM actúa como una metodología y como un proceso. La metodología incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. El modelo de

proceso ofrece un resumen del ciclo vital de la Minería de Datos. Las seis fases o etapas son:

1. Comprensión del negocio.
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Despliegue/Distribución o desarrollo (Implantación).

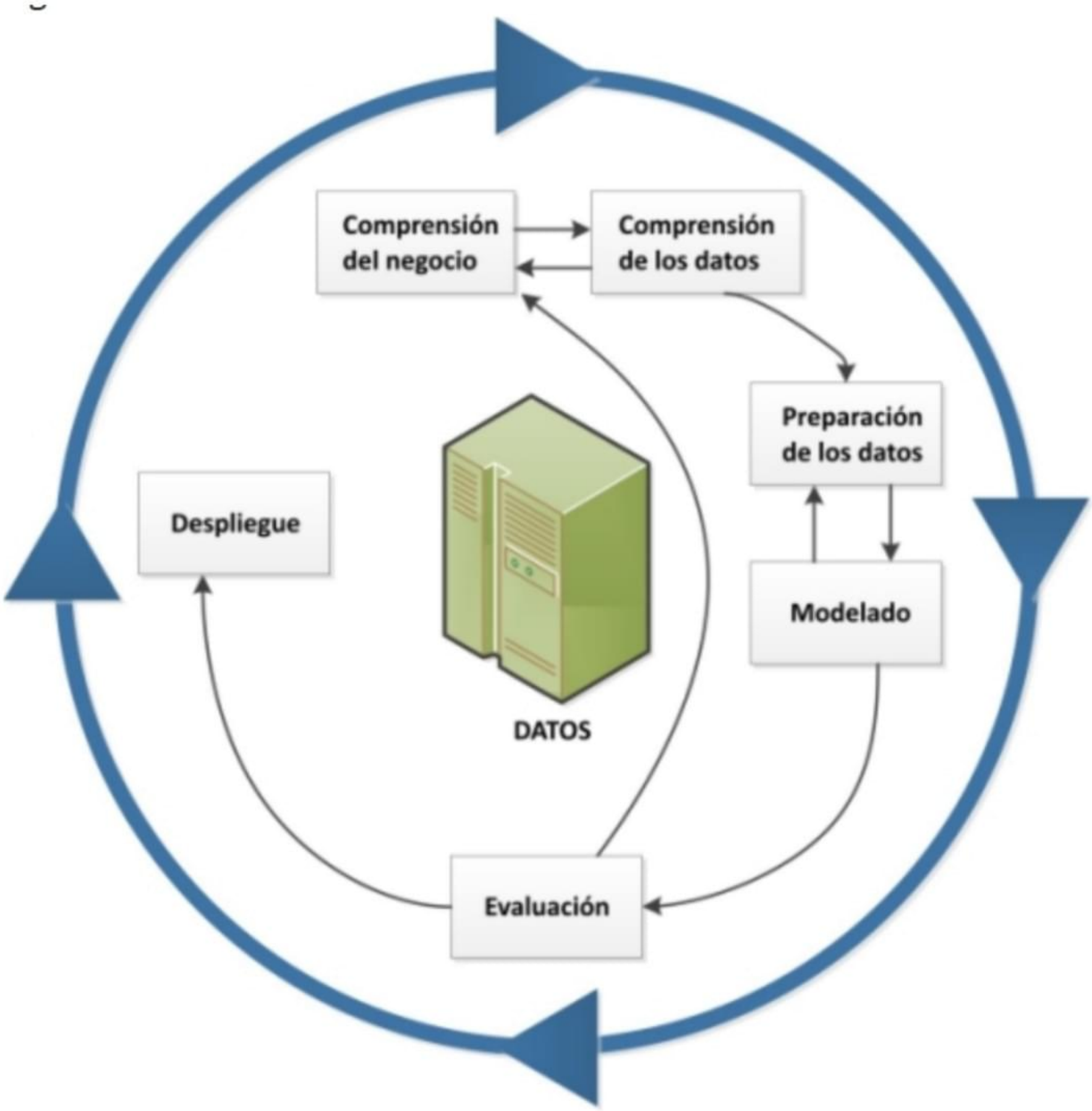


Figura 8.2. Etapas del proceso de Minería de Datos CRISP-DM. (Modelo de referencia).

El ciclo de vida del modelo tiene las seis etapas (fases) anteriores, que indican las dependencias más importantes y frecuentes entre etapas. La secuencia de

las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario. El resultado de cada fase determina que la fase, o la tarea específica de una fase, tiene que ser realizada después. Las flechas indican las más importantes y frecuentes dependencias entre fases.

El círculo externo de la figura 8.2 representa la naturaleza cíclica de la Minería de Datos y ésta no se termina una vez que la solución es desplegada.

Fase I. Comprensión del negocio

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y en la definición de las necesidades del cliente. Este conocimiento de los datos después se convierte en la definición de un problema de Minería de Datos y en un plan preliminar diseñado para alcanzar los objetivos. En esta fase se trata de entender los objetivos del negocio y requerimientos del proyecto, desde una perspectiva del negocio y no técnica.

Fase II. Comprensión de los datos

La fase de comprensión y estudio de datos comienza con recopilar y familiarizarse con los datos, descubrir conocimiento preliminar sobre esos datos y continúa con las actividades que permiten familiarizarse con ellos, identificar los problemas de calidad y analizar las primeras potencialidades, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

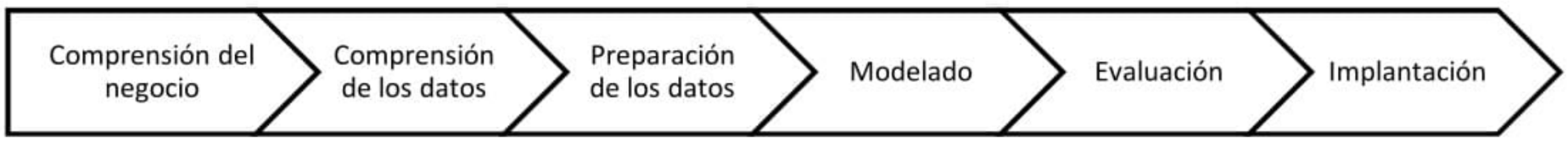


Figura 8.3 Modelo de referencia CRISP_DM. Fuente: CRISP_DM 1.0 pág. 12

Fase III. Preparación de datos

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado), a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan. En resumen, en esta fase se realiza el análisis de datos y la selección de características

El objetivo de esta fase es obtener la vista “minable”. Se realiza la selección de datos —a los que posteriormente se aplicarán las técnicas de modelado (variables y muestras)—, la limpieza de datos, la generación de variables adicionales, la integración de diferentes conjuntos de datos y los cambios de formato.

Fase IV. Modelado de datos

En esta fase se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más, mejor) y se calibran sus parámetros a valores óptimos. Típicamente, hay varias técnicas para el mismo tipo de problema de Minería de Datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Se seleccionan diversas técnicas de modelado adecuadas a un conjunto de datos ya preparado (la vista "minable"), a fin de centrarse en las necesidades específicas del negocio.

Fase V. Evaluación

En esta etapa del proyecto, se han construido uno o varios modelos que parecen alcanzar una calidad suficiente desde la perspectiva del análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo, revisar los pasos ejecutados para crearlo y comparar el modelo obtenido con los objetivos del negocio. Un objetivo clave es determinar si hay alguna cuestión importante del negocio que no haya sido considerada lo suficiente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos. El resultado final de esta fase es la obtención de resultados.

En esta fase se evalúa el modelo de la fase anterior, es decir, si el modelo nos sirve para responder a algunos de los requerimientos del negocio.

Fase VI. Despliegue

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es el de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica, y quizás automatizada, de un proceso de análisis de datos en la organización. El objetivo final de esta fase es la distribución o desarrollo (despliegue) y la puesta en producción.

La fase de despliegue (implementación, implantación o distribución) trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisiones de la organización, difundir informes sobre el conocimiento extraído, etcétera.

8.5.1 DESCRIPCIÓN DE LAS FASES DEL CICLO DE VIDA DE LA METODOLOGÍA CRISP-DM: UN ENFOQUE METODOLÓGICO

Fase 1. Comprensión del negocio

En esta fase se trata de entender los objetivos del negocio y requerimientos del proyecto desde una perspectiva del negocio y no técnica.

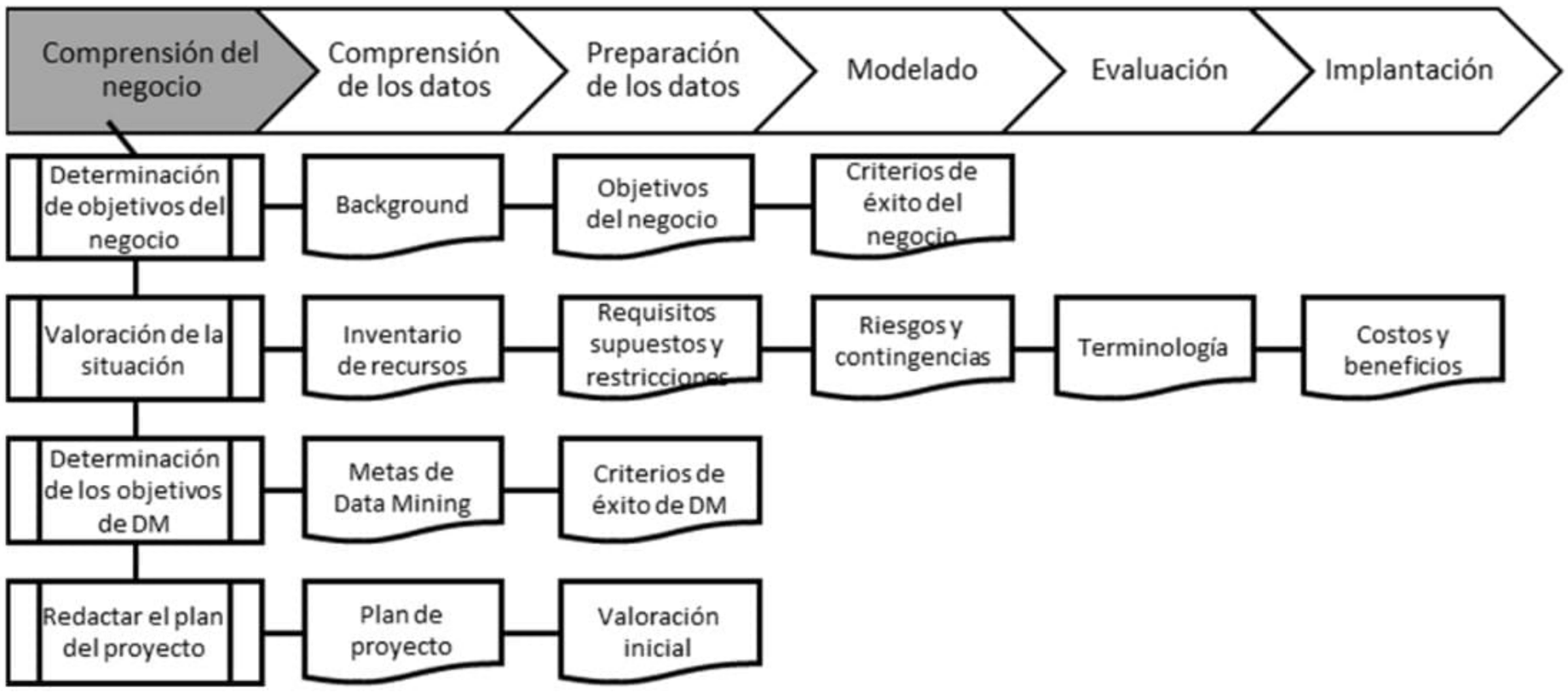


Figura 8.4. Tareas y resultados de la etapa de comprensión del negocio. Fuente: CRISP_DM 1.0 pág. 13

Las subfases son:

- Establecimiento de los objetivos del negocio
  - Desarrollar un glosario de términos para comprensión del lenguaje
  - Análisis de coste/beneficios.
  - ¿Cuál es el *statu quo* (contexto inicial)?
  - Comprender los objetivos del negocio/costes pain asociados.
  - Definir los criterios de éxito.
- Evaluación de la situación (inventario de recursos, requerimientos, suposiciones y restricciones, riesgos y contingencias, terminología (glosario de términos)
- Establecimiento de los objetivos de la minería de datos (objetivos de la minería de datos y criterios de éxito).
  - Identificar restricciones.
  - Recursos.

- Normas, leyes (por ejemplo, protección de datos).

- Generación (construcción) del plan del proyecto (plan del proyecto y evaluación inicial de herramientas y técnicas): listas, suposiciones y factores de riesgos (técnicos/financieros/negocios/organizacional).
  - Descomposición de tareas.
  - Descomponer los objetivos en subtareas.
  - Correspondencia entre subtareas con definiciones del problema de Minería de Datos.
  - Evaluación de los sistemas actuales.
  - Identificar los actores clave.
  - ¿Qué formatos debe tomar la salida?
  - Integración de la salida con el panorama de tecnologías existentes.
  - Entender normas de mercado y estándares.

Fase 2. Comprensión de los datos

Recopilar y familiarizarse con los datos, identificar los problemas de calidad de datos y analizar las primeras potencialidades o subconjuntos de datos de interés.

Las tareas son:

- Recopilación inicial de datos (informe/reporte de recopilación).
- Descripción de datos (temas de calidad de datos, calcular estadísticas básicas, informe de descripción).
- Exploración de datos (informe/reporte de exploración).
- Verificación de calidad de datos (informe/reporte de calidad).

Fase 3. Preparación de datos

El objetivo de esta fase es obtener la vista "minable". En esta fase se realiza la selección de datos —a los que posteriormente se aplicarán las técnicas de modelado (variables y muestras)—, la limpieza de datos, la generación de variables adicionales, la integración de diferentes conjuntos de datos y los cambios de formato.

En el modelo de referencia original se representan las tareas genéricas (negritas) y las salidas o resultados (cursivas).

Las tareas son:

- Selección de los datos.

  - Selección de subconjuntos de atributos.
  - Muestreo de datos (entrenamiento)/validar y conjunto de test.
- Limpieza de datos (informe de limpieza de datos).
- Construcción de datos (atributos derivados, registros derivados).
- Integración de datos (datos mezclados).
  - Unión de múltiples tablas de datos.
  - Resumen (*summarisation*)/agregación de datos.
- Dar formato a los datos.

Fase 4. Modelado de los datos

Se seleccionan diversas técnicas de modelado adecuadas a un conjunto de datos ya preparado (la vista minable), a fin de centrarse en las necesidades específicas del negocio. Las subfases son:

- Selección de la técnica de modelado apropiada (suposiciones de modelado).
- Diseño de la evaluación (generación de un diseño de comprobación: plan de pruebas).
  - Desarrollar un régimen de pruebas (*testing*/diseño del test).
  - Muestreo (verificar muestras que tengan características similares y sean representativas de la población).
- Construcción del modelo.
  - Elegir parámetros iniciales.
  - Estudiar comportamiento del modelo.
  - Descripción de los modelos.
- Evaluación del modelo.
  - Medidas del modelo.
  - Revisión de los parámetros elegidos.

Fase 5. Evaluación y pruebas

Evaluar el modelo de la fase anterior, es decir, si el modelo nos sirve para responder a algunos de los requerimientos del negocio. Las subfases son:

- Evaluación de resultados (validar el modelo).
  - Evaluación de resultados por expertos del dominio.
  - Evaluar la utilidad de los resultados desde la perspectiva del negocio.

- Revisión del proceso.
- Establecimiento de las siguientes etapas (lista de posibles acciones, decisión).
    - Arquitectura del despliegue.
    - Métricas para el éxito del despliegue.

Fase 6. De despliegue

La fase de despliegue (distribución) trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisiones de la organización, difundir informes sobre el conocimiento extraído, etcétera.

El desarrollo y evaluación de los modelos no es el final del proyecto de la Minería de Datos. Dependiendo de los requerimientos, la fase de despliegue puede ser tan simple como la generación de un informe o tan compleja como la implementación de procesos repetitivos en la empresa. Las subfases de la fase de despliegue son:

- Planificación del despliegue (desplegar el plan).
- Planificación de la monitorización y del mantenimiento (plan de la monitorización y del despliegue).
- Generación del informe/reporte final (informe y presentación final).
- Revisión del proyecto (documentación de las experiencias).

El despliegue o distribución es un proceso que consiste en utilizar sus nuevos conocimientos para implementar las mejoras en la organización. En la práctica, la distribución significa que utilice los conocimientos adquiridos en Minería de Datos para aplicar modificaciones en su organización. En general, la fase de distribución de CRISP-DM incluye dos tipos de actividades:

- Planificación y control de la distribución de resultados.
- Finalización de tareas de presentación, tales como la producción de un informe final y la revisión de un proyecto.

La etapa de despliegue puede también incluir actividades de mantenimiento para los modelos desplegados. Dado que todo lo relativo a los negocios cambia continuamente, los datos que reflejan las actividades del negocio también están cambiando. A lo largo del tiempo, los modelos (y los patrones embebidos en ellos) se construyen sobre datos antiguos que se pueden volver obsoletos, irrelevantes o erróneos (engañosos). Por consiguiente, el monitoreo o la monitorización y mantenimiento de los modelos son importantes si los resultados de la Minería de Datos se vuelven parte del día a día de los negocios y de su entorno (Turban 2011: 213).

8.6 PROCESO DE MINERÍA DE DATOS: METODOLOGÍA SEMMA

La metodología SEMMA, abreviatura de *Sample* (muestreo), *Explore* (exploración), *Modifiy* (modificación), *Model* (modelado) y *Asses* (valoración) es también muy conocida y utilizada. Se puede definir como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos". Fue desarrollada por el SAS Institute (2005).

El proceso de Minería de Datos SEMMA se compone de las siguientes etapas:

- **Muestreo**. Genera una muestra representativa de datos. Se identifican los datos (Entrada de datos, ejemplos, partición de datos).
- **Exploración**. Visualización y descripción básica de los datos. Se exploran los conjuntos de datos para observar relaciones y patrones, y se generan análisis diversos, identificación de variables importantes y análisis de asociación (Exploración distribuida, múltiples particiones, intuición, asociación, selección de variables).
- **Modificación**. Selección de variables y transformación de la representación de variables. Se preparan los datos para el análisis (Transformación de variables, filtros a los datos fuera de rango, agrupamiento, ruido).
- **Modelado**. Utiliza técnicas diversas de estadística y modelos de aprendizaje automático (Regresión, arboles, redes neuronales, etc.).
- **Evaluación (Valoración)**. Evalúa la precisión y utilidad de los modelos (Evaluación, medidas, reportes).



Figura 8.5. Etapas del proceso de minería de datos SEMMA

SEMMA comienza con una muestra representativa y estadística de los datos, lo que facilita aplicar técnicas de exploración, de estadística y de visualización; selecciona y transforma las variables predictivas significativas, modela las variables para predecir resultados y confirma una precisión del modelo.

Al evaluar el resultado de cada etapa del proceso SEMMA, el desarrollador del modelo puede determinar cómo modela nuevas cuestiones surgidas en los resultados anteriores y retroceder a la fase de exploración para refinamientos adicionales de los datos. Al igual que la metodología CRISP-DM, SEMMA se controla mediante un ciclo iterativo de experimentación. La principal diferencia

entre ambas metodologías de Minería de Datos es que CRISP-DM se centra en un enfoque muy exhaustivo de los procesos de Minería de Datos (incluyendo comprensión del negocios y datos relevantes), mientras que SEMMA supone implícitamente que los objetivos del proyecto de Minería de Datos y del negocio, junto con las fuentes de datos apropiadas, han sido identificadas y comprendidas (Turban 2011).

8.6.1 CRISP-DM *VERSUS* SEMMA

El proceso KDD es un proceso exhaustivo y muy completo, que contiene y utiliza métodos de Minería de Datos para encontrar información útil y patrones en los datos, en contraposición con Minería de Datos, que implica utilizar algoritmos para identificar patrones de datos deducidos del propio proceso KDD. La entrada al proceso KDD son datos organizacionales. Los almacenes de datos (*Data Warehouses*) de empresa facilitan que KDD sea implementado eficientemente, ya que proporciona una fuente de datos para ser minada. Dunham (2003)[2] adapta el proceso de Fayyad (1996) a las siguientes etapas: selección de datos, preprocesamiento de datos, transformación de datos, Minería de Datos e interpretación/evaluación.

Las metodologías CRISP-DM y SEMMA presentan las características citadas anteriormente, y dependiendo del proveedor de *software*, utilizarán una metodología u otra, siendo CRISP-DM la más implantada entre las soluciones de *software* de Minería de Datos.

El portal KDnuggets[3] (portal de referencia de *Business Intelligence, Data Mining*y *Data Science*) ha realizado dos estudios profundos entre empresas usuarias de herramientas de Minería de Datos, a los que denomina *KDnuggets Polls* (años 2007 y 2014), para conocer cuáles son las metodologías de Minería de Datos más utilizadas en las herramientas de proveedores de soluciones de *software*. Tanto en el *ranking* del año 2007 como en el de 2014, la metodología o proceso CRISP-DM resultó ser la más utilizada. En el *ranking* de 2014 ya contempla las soluciones en la era de *Big Data* y destaca que CRISP-DM sigue siendo la metodología más popular para proyectos de Analítica, Minería de Datos y Ciencia de Datos, con un 43% de utilización (42% en 2007), frente a metodologías propias del proveedor de Minería de Datos con un 27,5% (19% en 2007) y SEMMA con un escaso 8,5%, destacando en este caso una reducción considerable respecto del *ranking* de 2007, donde reconocía un porcentaje de uso del 13%.

8.7 MODELOS, ALGORITMOS Y TÉCNICAS DE MINERÍA DE DATOS

Existen dos modelos o métodos de Minería de Datos en función de su propósito general:

- Modelo descriptivo.

  Describe el comportamiento de los datos (conjunto de datos) de forma que sean interpretables por un usuario experto. Encuentra valores interpretables que describen los datos. Trata de proporcionar información entre las relaciones de los datos y sus características. En el modelo descriptivo se dispone de una variable (denominada respuesta) con valor desconocido y el objetivo fundamental del modelo es determinar ese valor.

- Modelo predictivo.

  Describe los datos y se utilizan para predecir el valor de algún atributo desconocido. Es decir, se utilizan algunas variables para predecir valores desconocidos de otras variables. Está orientado a estimar valores de salida. Encuentra patrones interpretables que describen los datos.

  Un modelo predictivo intenta predecir o responder a preguntas futuras sobre la base de un estudio de comportamiento pasado. Preguntas que responden a este tipo de modelo de datos (o minería):

  ¿Cómo se venderá el próximo año el producto x?

  ¿Cuántos tipos de personas comprarán el producto?

  ¿Qué riesgo tiene una cierta persona de contraer una enfermedad determinada x sobre la base de sus características personales?

  ¿Qué clientes son más propensos a darse de baja de nuestra empresa?

  El modelo predictivo requiere ser entrenado, utilizando un conjunto de datos de entrenamiento cuyo valor de variable objetivo es conocido. En esencia, el modelo entrega resultados sobre la base de un aprendizaje que se va ajustando a la realidad conocida.

8.7.1 ALGORITMOS DE APRENDIZAJE

La implementación de estos modelos requiere del uso de **algoritmos de aprendizaje supervisado y aprendizaje no supervisado**, que en la actualidad se centran en técnicas de aprendizaje automático (*machine learning*) de Inteligencia Artificial. La tendencia de mayor impacto en la actualidad en la Inteligencia Artificialy en el aprendizaje automático es el aprendizaje profundo (*Deep learning*), una disciplina inmersa en el aprendizaje automático.

Los modelos predictivos se apoyan en modelos de aprendizaje supervisado, tales como aplicaciones para la detección de aquellos clientes de una empresa (por ejemplo, operadora telefónica) que son más propensos a abandonar la empresa como clientes. Los modelos descriptivos se apoyan en modelos de aprendizaje no supervisado, donde se evalúan de acuerdo con las observaciones o datos entregados, recurriendo, normalmente, a argumentos heurísticos que permiten evaluar la calidad de los resultados.

Las **técnicas predictivas** más utilizadas son los algoritmos de:

- Clasificación.
- Regresión.
- Series temporales.
- Detección de desviaciones/anomalías.

Las **técnicas descriptivas** más utilizadas son los algoritmos de:

- Asociación (reglas de asociación).
- Agrupamiento o segmentación (*clustering*).
- Sumarización *(*resúmenes).
- Descubrimiento de secuencias de patrones.

8.8 RELACIONES DE LA MINERÍA DE DATOS CON OTRAS DISCIPLINAS: DE *BIG DATA* A *DATA SCIENCE*

La Minería de Datos tiene un enfoque integrador de múltiples disciplinas (multidisciplinar) debido a que utiliza diferentes técnicas según el tipo de información que se ha de extraer. En este sentido, se consideran disciplinas y tecnologías los métodos analíticos, estadísticos, enfoques gráficos, visualización, algoritmos, reconocimiento de patrones, bases de datos, aprendizaje máquina o aprendizaje automático (*machine learning*) y la Inteligencia Artificial.

La Minería de Datos no es una disciplina nueva, ya que se apoya en el proceso KDD que, como ya se ha comentado, tiene sus orígenes en el año 1996, pero en estos últimos años ha evolucionado considerablemente al incluir numerosas nuevas tendencias, tales como estadística avanzada, Inteligencia Artificial, aprendizaje automático y aprendizaje profundo, redes neuronales, reconocimiento de patrones, *Big Data* y *Data Science*, además de los conocimientos de sistemas de información, bases de datos y programación avanzada de algoritmos. Así, la Minería de Datos es una fusión de múltiples disciplinas, tales como:

- Estadística.
- Bases de datos y sistemas de información.

- Aprendizaje automático (*machine learning*).
- Aprendizaje profundo (*deep learning*)
- Visualización de datos.
- Computación paralela y distribuida.
- Interfaces de lenguaje natural con bases de datos.
- Sistemas de toma de decisiones.
- Inteligencia Artificial.
- Tecnologías de bases de datos y almacenamiento de datos.
- Reconocimiento de patrones.
- Modelos matemáticos.
- Algoritmia y programación
- Otras disciplinas
- Modelos matemáticos

En síntesis, en la figura se señala la relación entre la minería de datos con otras disciplinas.



Figura 8.6. Disciplinas relacionadas con la Minería de Datos.
*Machine Learning y Deep Learning.

8.9 HERRAMIENTAS DE *SOFTWARE* DE MINERÍA DE DATOS

Muchos proveedores comerciales de *software* desarrollan y comercializan herramientas de Minería de Datos. Existe un gran número de proveedores de soluciones de *software* especializados en Minería de Datos, tales como IBM, SAS, Megaputer y SAP. De igual forma, los proveedores de soluciones de software de Inteligencia de Negocios ofrecen soluciones específicas o integradas en sus herramientas (*suites*), tales como IBM Cognos, Oracle Hyperion, SAP Business Object, Microsoft, Microstrategy, Teradata, Qlik, Tableau. Además de estas herramientas de *software* propietario, han emergido numerosas herramientas de *software* libre y gratuito y de código abierto (*open source*), tales como WEKA, RapidMiner, Orange, Keel, KNIME, que ofrecen sus soluciones de *software* abierto y soporte para usuarios y desarrolladores. Asimismo, plataformas de *software* abierto de Inteligencia de Negocio ofrecen soluciones de código abierto e infraestructuras para soportar los desarrollos de Minería de Datos. Las plataformas más populares son Pentaho, Jaspersoft y Birst.

8.9.1 HERRAMIENTAS DE CÓDIGO ABIERTO

Existe un gran número de soluciones de código abierto, y también gratuitas, que pueden ser utilizadas en ámbitos académicos y de investigación, así como en ámbitos profesionales y empresariales.

WEKA. Es una de las herramientas gratuitas y de código abierto más populares. Fue desarrollada por un grupo de investigadores de la *University of Waikato* de Nueva Zelanda. La herramienta se puede descargar gratuitamente del sitio `cs.waikato.ac.nz/ml/weka`. Weka es un *software* de código abierto bajo la licencia GNU GPL (General Public License), escrito en el lenguaje de programación Java, y es una colección de algoritmos de aprendizaje automático para tareas de Minería de Datos. Los algoritmos se pueden aplicar directamente a un conjunto de datos o bien llamados desde su propio código Java. Weka contiene herramientas para preprocesamiento de datos, clasificación, regresión, agrupamiento (*clustering*) y visualización. Weka proporciona acceso a bases de datos SQL, utilizando conectividad de bases de datos en Java, y puede procesar los resultados devueltos como consultas de bases de datos.

Las versiones disponibles[4] son 3.8, la versión más estable, y 3.9, la versión para desarrolladores. Todas ellas corren en los sistemas operativos Windows, Mac OS X y Linux.

KNIME (Konstanz Information Miner). Es una plataforma de Minería de Datos (`www.knime.org`) que permite el desarrollo de modelos en un entorno visual. Está desarrollada sobre la plataforma Eclipse y programada, fundamentalmente,

en Java. Es una plataforma de análisis de datos de código abierto, de fácil uso y comprensible, para integración de datos, procesamiento, análisis y exploración. A través de *plugins*, los usuarios pueden añadir módulos de texto, imagen, procesamiento de series temporales y la integración de varios proyectos de código abierto, tales como R y WEKA.

Se desarrolló originalmente en el Departamento de Bioinformática y Minería de Datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com Gmbh es la propietaria de KNIME y está radicada en Zurich, Suiza, donde realiza labores de consultoría y formación (`knime.org`).

RapidMiner (`rapid-i.com`). Es una herramienta de *software* comercial desarrollada por Rapid-I, que se puede descargar del sitio `rapid-i.com`, y que fue creada como una herramienta de código abierto. Es una herramienta especializada en Minería de Datos y aprendizaje automático. Dispone de una interfaz gráfica de usuario muy mejorada, que utiliza un gran número de algoritmos e incorpora un conjunto de características muy potentes de visualización de datos. Se utiliza tanto en el mundo profesional como en investigación. Ha evolucionado considerablemente desde su creación y, en la actualidad, se ofrece también como una plataforma unificada de Ciencia de Datos.

El sitio web de RapidMiner ofrece una gran cantidad de recursos, guía de referencia de algoritmos, un *blog* de Ciencia de Datos, reportes, casos de estudio y un buen catálogo de *eBooks* y *whitepapers* (artículos). RapidMiner ofrece diferentes versiones, gratuitas (con comunidad de soporte) y de empresa. La versión RapidMiner Studio es la más utilizada para predicciones, validaciones y prototipados. Dispone de un programa con licencias gratuitas para uso académico, tanto para estudiantes como para profesores e investigadores.

KEEL (`keel.es`). KEEL (*Knowledge Extraction for Evolutionary Learning)* es un *software* de código abierto de aprendizaje automático, desarrollado e implementado en Java, en la Universidad de Granada, y distribuido bajo la licencia GPLv3. Está potenciado por una GUI bien organizada que permite administrar (importar, exportar, editar y visualizar) datos en diferentes formatos de archivos y experimentar con ellos (a través de sus sistemas de preprocesamiento de datos, bibliotecas estadísticas y algoritmos de Minería de Datos y aprendizaje evolutivo). Dado que KEEL está basado en Java, el entorno JVM ha de estar instalado en el sistema para ejecutar sus propios GUI y realizar proyectos de Minería de Datos. En el sitio `keel.es` se puede consultar la larga lista de algoritmos soportados. KEEL es ideal para fines académicos y de investigación; al ser un producto desarrollado en un grupo de investigación de la Universidad de Granada, es ideal para soporte y apoyo a profesores de materias como Minería de Datos, Ciencia de Datos o Inteligencia de Negocios en general.

Orange (`orange.biolab.si`). Es una solución integrada (*suite*) para Minería de Datos y aprendizaje automático, desarrollada por la Facultad de Informática de la Universidad de Ljubljana (Eslovenia). Es una herramienta que permite programación visual para el análisis exploratorio de datos y visualización, y está integrada con la biblioteca de algoritmos y programas de Python. Es una aplicación multiplataforma de *software* libre y de código abierto, que se distribuye bajo licencia GPL.

R. Programación. Es un lenguaje de programación y entorno de *software* de código abierto y gratuito para computación, estadística y gráficos. Proporciona múltiples técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásicas, clasificación, agrupación en clústeres, etcétera. Compila y ejecuta una amplia variedad de sistemas operativos (Unix, Windows y Mac OS). El lenguaje R se utiliza ampliamente entre mineros de datos para desarrollo de *software* estadístico y análisis de datos.

NLTK. Es un *pool* (grupo) de herramientas de procesamiento de lenguajes, incluyendo Minería de Datos, aprendizaje automático, *data scraping* (raspado de datos), análisis de sentimientos y otras tareas diversas de procesamiento de lenguajes. Está escrito en Python y está disponible en Windows, Mac OS X y Linux. Es un proyecto de código abierto (*open source*) y está controlado por una comunidad.

8.9.2 HERRAMIENTAS COMERCIALES PROPIETARIAS

La mayoría de los proveedores comerciales de *software* de soluciones de Inteligencia de Negocios disponen de herramientas propias de Minería de Datos o integradas en *suites* (paquetes integrados) comerciales. Algunas de las herramientas más acreditadas del mercado son comercializadas por IBM, SAS, Oracle, SAP y Microsoft.

IBM SPSS Modeler (Clementine)

`(www-01.ibm.com/software/analytics/spss/products/modeler)`

Es una herramienta de Minería de Datos de IBM (antes conocida por el popular nombre de Clementine) muy eficiente, que funciona con diferentes fuentes de datos (ASCII, ODBC), con una interfaz visual basada en procesos de datos y que manipula la mayoría de las diferentes técnicas de Minería de Datos más populares —clasificación, regresión, agrupamiento (*clustering*), redes neuronales, reglas temporales, árboles de decisión—. La interfaz gráfica facilita el descubrimiento de patrones y tendencias de grandes volúmenes de datos (estructurados y no estructurados).

Es una herramienta muy popular en análisis predictivo, es considerada como uno de los mejores proveedores de *software* de Minería de Datos, y se adapta para el análisis de Minería de Datos con *Big Data*. La herramienta SPSS Modeler

es un programa comercial con licencia de pago y será necesario que las empresas analicen previamente la puesta en marcha de la aplicación, aunque la experiencia y eficiencia de la herramienta y su gran implantación en el mercado en los últimos años, tanto en grandes empresas como en medianas e incluso pequeñas, hacen que SPSS Modeler sea una opción muy idónea como herramienta de Minería de Datos, que tiene la facilidad de integración con otras soluciones de *software* de IBM de Inteligencia de Negocios. El sistema se integra con IBM Cognos e InfoSphere. La combinación de estas tres herramientas ayuda a realizar predicciones precisas sobre cómo se desarrollará el estado de su negocio. También ayuda a mejorar los resultados comerciales en CRM, *marketing*, planificación de recursos, mitigación de riesgos y otras áreas.

SAS Enterprise Miner
(www.sas.com/technologies/analytics/datamining/miner/)

Es una herramienta de SAS (uno de los proveedores líder a nivel mundial de herramientas de Analítica de Datos o *Analytics*) que realiza el proceso de Minería de Datos y facilita la creación de modelos predictivos y descriptivos de alta precisión para grandes volúmenes de datos. La herramienta dispone de una interfaz gráfica que integra un conjunto de herramientas necesarias para la toma de decisiones y otras herramientas de análisis de datos pioneras y líderes del mercado de Analítica.

SAS Enterprise Miner está basada en la metodología de Minería de Datos SEMMA, desarrollada por SAS Institute. Es una de las herramientas con mayor implantación en el mercado para soluciones de grandes bases de datos y de *Big Data*. Ofrece un amplio conjunto de algoritmos avanzados muy completo para modelados predictivos y descriptivos, incluyendo regresión, redes neuronales, árboles de decisión, etcétera. SAS es uno de los grandes proveedores de *software* analítico a nivel mundial, por lo que su adopción como herramienta de Minería de Datos entraña también un estudio profundo sobre su coste y funcionalidades, especialmente para aquellas empresas que no sean clientes ya de otras soluciones de SAS.

Oracle Data Mining
(www.oracle.com/products/database/options/advanced-analytics/index.html)

Oracle Data Mining (ODM) es una herramienta de *software* de Minería de Datos desarrollada por la empresa Oracle —líder mundial en *software* de bases de datos— con soporte en las técnicas más avanzadas de Minería de Datos y aplicada a grandes volúmenes de datos. Es una herramienta que realiza el ciclo completo de integración de datos desde la importación de datos, preparación, desarrollo y despliegue del modelo.

La herramienta ODM integra todas las etapas del proceso de Minería de Datos y permite integrar los modelos con las bases de datos comerciales de Oracle, por lo

que no necesita exportar los archivos de usuario a paquetes de *software* externos.

Una funcionalidad muy sobresaliente de la herramienta de Minería de Datos de Oracle es la oferta de las dos versiones comerciales que ofrece ODM:

1. Herramientas tradicionales de Minería de Datos que, mediante una interfaz gráfica muy potente, les permite a los usuarios aplicar las técnicas de Minería de Datos que consideren necesarias para poder tomar las mejores decisiones en los procesos de negocios.
2. La herramienta ODM permite a los desarrolladores utilizar las API de Oracle, por lo que podrán realizar sus propias aplicaciones específicas para la empresa. Esta propiedad es una excelente herramienta para las empresas clientes de Oracle, sobre todo si utilizan otras herramientas del catálogo empresarial.

SAP Business Object

(`www.sap.com/spain/products/analytics/business-intelligence-bi.html`)

Es una herramienta integrada que, además de Minería de Datos, abarca todo el ciclo de vida de un sistema de Inteligencia de Negocios. Está disponible como una solución integrada en la empresa cliente (*onpremise*) o en la nube. Sus funcionalidades más importantes son plataforma de Inteligencia de Negocios, visualización y análisis de datos, tableros de datos (*dashboards*) e informes. Se integra fácilmente con el sistema eficiente de bases de datos "en memoria" (SAP HANA).

Microsoft. SQL Server Data Mining

(`https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-ssas`)

SQL Server es una herramienta para análisis predictivo, desde la versión en el año 2000, y facilita la Minería de Datos en Analysis Services. La combinación de la Minería de Datos con *Integration Services, Reporting Services y SQL Server* proporciona una plataforma integrada para el análisis predictivo, en la que se incluye la limpieza de los datos, la preparación, el aprendizaje automático y la generación de informes. SQL Server en Minería de Datos incluye varios algoritmos estándar, como los modelos de clústeres EM y mediana-K, redes neuronales, regresión logística y regresión lineal, árboles de decisión y clasificadores de Bayes. Todos los modelos tienen visualizaciones integradas para ayudarle a desarrollar, restringir y evaluar los modelos. Integrar la Minería de Datos en una solución de Inteligencia Empresarial le ayudará a tomar decisiones inteligentes sobre problemas complejos.

En SQL Server 2017, la Minería de Datos es eficaz y accesible, y está integrada con las herramientas preferidas de los usuarios para el análisis y la creación de informes.

CASO DE ESTUDIO

Kaggle es un portal de referencia a nivel mundial (`www.kaggle.com`) en Ciencia de Datos y, por consiguiente, en Minería de Datos y herramientas de *software* de Ciencia de Datos. Está formado por una gran comunidad internacional de científicos de datos. Es uno de los sitios más recomendados para aprendizaje, formación e investigación en disciplinas tales como Ciencia de Datos y, en consecuencia, Minería de Datos, aprendizaje automático, aprendizaje profundo y algoritmos de Inteligencia Artificial. Kaggle es una empresa *start-up* creada en Australia en 2010, con el objetivo de agrupar a la naciente comunidad de científicos de datos y especialistas de aprendizaje automátco de todo el mundo. Según diversas fuentes consultadas, tienen más de 500.000[5]socios activos que comparten sus conocimientos, ideas e incluso modelos concretos para analizar grandes conjuntos de datos (*Big Data*) y extraer información concreta de los mismos a través de técnicas de Minería y Análisis de Datos.

El sitio web de Kaggle es una plataforma para compartir conocimiento y ayudar y tutorizar a sus usuarios, tanto en su aprendizaje en las técnicas de Ciencias de Datos como también para la ayuda en la resolución de problemas concretos que se les pueden plantear. La plataforma ofrece múltiples tutoriales, artículos y competiciones para las personas que se inician en Ciencia de Datos y dispone de un foro muy activo, en el que la comunidad plantea todo tipo de temas y ayuda en su formación. Presenta una sección para que las empresas puedan publicar ofertas de trabajo para profesionales de la Ciencia de Datos.

Una sección muy interesante es la opción de competiciones, en las que los miembros de la comunidad pueden participar con sus propias soluciones, así como valorar y votar las del resto. También es posible encontrar conjuntos de datos y algoritmos, listos para ser analizados con los modelos diseñados por los usuarios, y soluciones de análisis para casos concretos, que pueden ser descargadas ya compiladas o bien acceder a su código fuente para realizar desarrollos propios de los programadores. Los lenguajes de programación más utilizados en la plataforma son aquellos que disponen de buenas librerías de funciones, tales como R y Python, aunque también se pueden encontrar desarrollos en lenguajes de programación tradicionales como C++, Java o Javascript.

Kaggle ha sido adquirida por Google, según anunciaron directivos de la empresa el 8 de marzo de 2017, con ocasión de la conferencia Cloud Next que se desarrolló en San Francisco. Esta noticia ha significado el lanzamiento de la plataforma y su potenciación dentro de la Inteligencia Artificial y de sus

disciplinas de aprendizaje automático y Minería de Datos. La fuerte intensificación del I+D+i de Google en Inteligencia Artificial apoyarán el crecimiento de Kaggle, continuando como fuente de referencia mundial en estas disciplinas.

La página de bienvenida a **kaggle.com** se anuncia como *The Home of Data Science & Machine Learning*. Las tres grandes secciones que presenta son: *Competitions* (plataforma para competiciones de Ciencias de Datos, donde se ayuda a resolver problemas difíciles, reclutar equipos potentes y aplicar su talento en el desarrollo de la Ciencia de Datos), *Datasets* (exploración y análisis de una colección de conjuntos de datos públicos de alta calidad) y *Kernels* (núcleos de trabajo, donde se pueden ejecutar código en la nube y recibir realimentación de la comunidad en su trabajo).

8.10 RESUMEN

Minería de Datos es un conjunto de técnicas que se utiliza para optimizar el desarrollo de la Inteligencia de Negocios a partir de los datos que una organización recolecta, organiza y almacena, para, posteriormente, realizar el análisis de datos y contribuir del modo más eficiente posible en la toma de decisiones.

El término Minería de Datos (*Data Mining)* se utiliza para describir el descubrimiento o "minado" (*mining*) del conocimiento a partir de grandes cantidades de datos. La Minería de Datos es una etapa dentro del proceso **KDD** (*Knowledge Discovery from Data*), **descubrimiento del conocimiento a partir de datos**, aunque el proceso KDD es un término más global y la Minería de Datos es una etapa dentro del proceso global KDD.

La Minería de Datos es una disciplina en la que confluyen muchas otras disciplinas clásicas en el mundo científico y de negocios: estadística, Inteligencia Artificial, aprendizaje máquina (*machine learning*), sistemas de información, algoritmos y bases de datos y, recientemente, *Data Science* (Ciencia de Datos) y análisis de grandes volúmenes de datos (*Big Data*).

Las aplicaciones de la Minería de Datos son muy numerosas y de impacto en una gran cantidad de sectores de todo tipo: medicina, industria, banca, seguros, *hardware* y *software* de computadoras, gestión de relaciones con los clientes, deportes, etcétera.

La extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases*, KDD), que se refiere al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información: Las etapas del proceso KDD son:

1. Selección de datos.
2. Preprocesamiento.
3. Transformación.
4. Minería de Datos.
5. Interpretación y evaluación.

Las dos metodologías más utilizadas en Minería de Datos son CRISP-DM y SEMMA.

La metodología o proceso CRISP-DM tiene seis etapas:

1. Comprensión del negocio.
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Despliegue/Distribución o desarrollo (Implantación).

El proceso de Minería de Datos SEMMA se compone de las siguientes etapas:

- Muestreo.
- Exploración.
- Modificación.
- Modelado.
- Evaluación (evalúa la precisión y utilidad de los modelos).

Los modelos de Minería de Datos más empleados, en función de su propósito general, son: descriptivos y predictivos.

Las **técnicas predictivas** más utilizadas son los algoritmos de:

- Clasificación.
- Regresión.
- Series temporales.
- Detección de desviaciones/anomalías.

Las **técnicas descriptivas** más utilizadas son los algoritmos de:

- Asociación (reglas de asociación).
- Agrupamiento o segmentación (*clustering*).

- *Sumarización (*resúmenes).
- Descubrimiento de secuencias de patrones.

La Minería de Datos tiene relación con otras disciplinas importantes en la Inteligencia de Negocios:

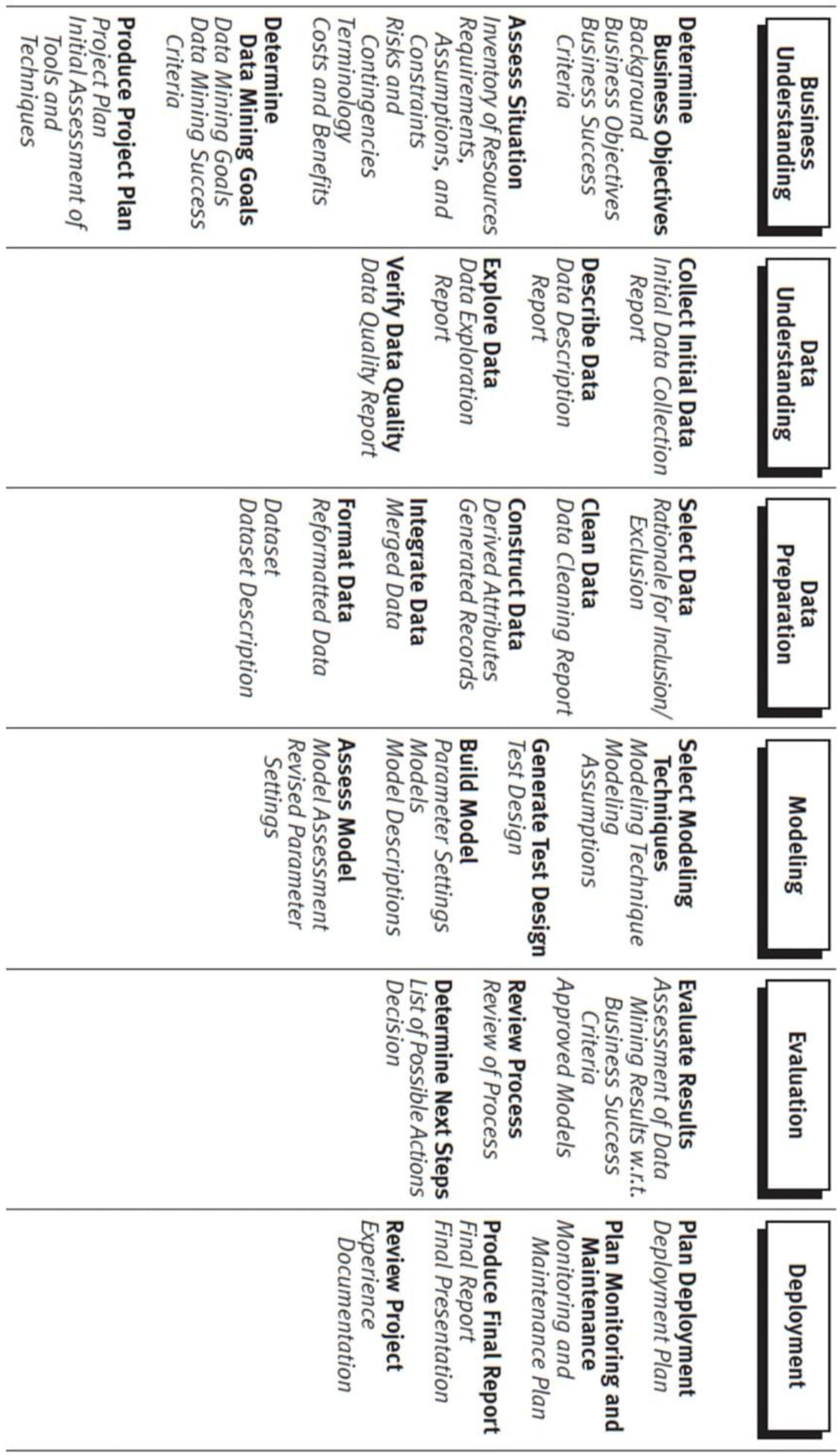
- Estadística.
- Computación paralela y distribuida.
- Aprendizaje automático (*machine learning*).
- Visualización.
- Sistemas de toma de decisiones.
- Tecnologías de bases de datos.
- Reconocimiento de patrones.
- Algoritmia y programación.
- Otras disciplinas.

Existen numerosas herramientas de Minería de Datos bien específicas o integradas en paquetes, o *suites*, con otras herramientas de Inteligencia de Negocios. Las herramientas de código abierto y gratuitas más populares son: WEKA, KNIME, RapidMiner, KEEL, Orange y proveedores comerciales de software propietario como IBM SPSS Modeler (Clementine), SAS Enterprise Miner, Oracle Data Mining.

Visualización del ciclo de vida CRISP-DM

Tareas genéricas (negritas) y salidas/resultados (cursiva). Manual original de CRISP-DM 1.0 (crips-dm.org).

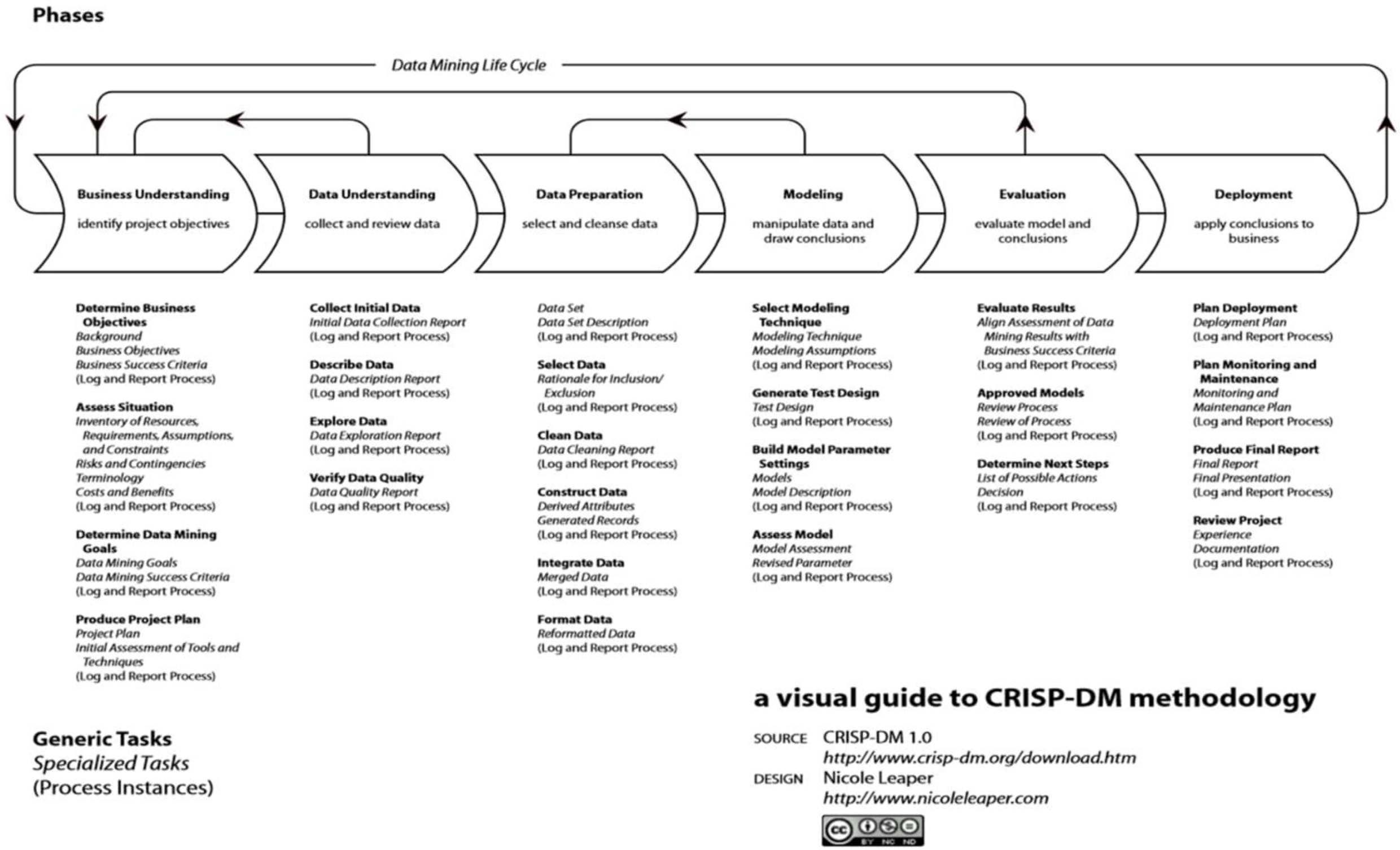
Descarga en línea: `the-modeling-agency.com/crisp-dm.pdf`



Guía visual de la metodología CRISP-DM. March 13, 2009

Ciclo de vida de la minería de datos

https://exde.wordpress.com/2009/03/13/a-visual-guide-to-crisp-dm-methodology/



Guía visual de la metodología CRISP-DM (traducción y adaptación)

Elizabeth León Guzmán. *Minería de datos*. Universidad Nacional de Colombia

http://disi.unal.edu.co/~eleonguz/cursos/md/presentacione s/Sesion5_Metodologias.pdf

Comprensión del Negocio	Comprensión de Datos	Preparación de Datos	Modelamiento	Evaluación	Despliegue
•**Determinar los Objetivos del Negocio** ✓*Antecedentes* ✓*Objetivos del Negocio* ✓*Criterio de Éxito* • **Evaluar la situación** ✓*Inventario de requerimientos de Recursos, Hipótesis y Limitaciones* ✓ *Riesgos y Contingencias* ✓*Terminología* ✓*Costos y Beneficios* • **Determinar el objetivo de Minería de Datos** ✓*Objetivos de Minería de Datos* ✓*Criterio de Éxito de Minería de Datos* • **Desarrollar el Plan de Proyecto** ✓ *Plan de proyecto* ✓ *Evaluación inicial de*	•**Obtener los datos iniciales** ✓*Reporte de la obtención de los datos iniciales* • **Describir los Datos** ✓*Reporte con la descripción de los datos* • **Explorar de Datos** ✓*Reporte de la Exploración de Datos* • **Verificar de la calidad de los Datos** ✓ *Reporte de la calidad de los datos*	• *Conjunto de Datos* • *Descripción de los Datos* • **Seleccionar los Datos** ✓*Justificación de la inclusión / Exclusión* • **Limpiar Datos** ✓*Reporte de Limpieza de Datos* •**Construir Datos** ✓*Atributos Derivados* ✓*Registros Generados* • **Integrar Datos** ✓ *Datos Combinados* • **Dar formato a los Datos** ✓*Datos Formateados*	•**Seleccionar Técnica de Modelamiento** ✓ *Técnica de Modelamiento* ✓*Modelamiento* ✓*Hipótesis* • **Generar el Diseño de Prueba** ✓*Diseño de Prueba* • **Construir el Modelo** ✓*Configuración de los parámetros del Modelo* ✓*Descripción del Modelo* • **Evaluar el Modelo** ✓*Evaluación del Modelo* ✓*Revisión de la configuración de los parámetros del modelo*	• **Evaluar Resultados** ✓*Hipótesis de Minería de Datos* ✓*Resultados* ✓*Criterio de éxito del negocio* ✓ *Modelos aprobados* • **Revisar el Proceso** ✓*Revisión del Proceso* • **Determinar los siguientes pasos** ✓*Lista de Posibles Acciones* ✓*Decisión*	• **Desplegar el Plan** ✓ *Plan de Despliegue* • **Monitorear y Mantener** ✓*Plan de monitoreo y Mantenimiento* • **Desarrollar el reporte final** ✓*Reporte Final* ✓*Presentación Final* •**Revisión del Proyecto** ✓*Documentación de las experiencias*



NOTAS

[1] *U Fayyad et al (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, 39 (11), 27-34. November 1996.*

[2] *Citado por Turban (2011). Adapta el proceso KDD original de Fayyad et al. (1996).*

[3]*KDnugetts, CRISP-DM, still the top methodology for analytics, data mining, or data science projects:*

`https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html`

[4]*La última versión de desarrollo (junio de 2018) para descarga en el sitio web es 3.9.2. Se define comoWeka 3; Data Mining software in Java.*

`www.cs.waikato.ac.nz/ml/weka/index.html`

[5]*La visita realizada, el 18 de octubre de 2017, a kaggle.com, la propia comunidad reconocía —ya en esas fechas— el millón de usuarios: "Seewhatourcommunity of overonemillion data scientists can do foryou":* `https://www.kaggle.com/host`

CAPÍTULO 9
MINERÍA WEB Y MINERÍA DE TEXTOS

CONTENIDO

9.1 Introducción
9.2 Minería de Textos
9.3 Herramientas de la Minería de Textos
9.4 Minería Web: conceptos, definiciones y categorías
9.5 Arquitectura de la Minería Web
9.6 Categorías de la Minería Web
9.7 Minería Web de Contenido
9.8 Minería Web de la Estructura
9.9 Minería Web del Uso
9.10 Herramientas de Minería Web
9.11 Motores de búsqueda (buscadores)
9.12 Posicionamiento SEO: Optimización de los motores de búsqueda
9.13 Posicionamiento SEM
9.14 RESUMEN

OBJETIVOS

- Conocer y comprender la Minería Web: sus objetivos, aplicaciones y beneficios.
- Conocer y diferenciar los tres tipos de Minería Web.
- Describir la Minería de Textos y sus aplicaciones en los procesos de Minería de Datos y de Inteligencia de Negocios.
- Comprender la necesidad de la Minería de Textos y la Minería Web.
- Conocer y comprender las diferencias entre Minería de Datos y Minería de Texto.
- Conocer y entender las aplicaciones de la Minería de Opinión.
- Aprender los conceptos fundamentales del análisis de sentimientos.

9.1 INTRODUCCIÓN

La Web es una colección de archivos interrelacionados en uno o más servidores web. La Minería Web (*Web Mining*) es la aplicación de técnicas de Minería de Datos para extraer conocimiento de los datos de la Web.

Oren Etzioni fue el primer investigador que utilizó el término para referirse al proceso de automatización del descubrimiento y extracción de información de documentos y servicios web y es considerado, en consecuencia, como "el padre de la Minería Web. En su artículo *The World Wide Web: Quagmire or Gold Miner?,* publicado en la prestigiosa revista *Communications of the ACM*, definió el término *Web Mining* y las subtareas (etapas) de la Minería Web, las que agrupó en tres categorías: **descubrimiento de recursos**, **extracción de la información** y **generalización**. A la publicación de este primer trabajo de Minería Web, le siguieron artículos en los siguientes años, que se publicaron entre 1997 y 2000 en foros, revistas y workshops.

El artículo de influencia sobre Minería Web "*Web Mining Reseach: A Survey*", que ha sido utilizado como soporte en este capítulo, fue publicado por R. Kosala, y H. Blockeel, en el año 2000, en la revista de *ACM SIGKDD Explorations Newsletter*. En este artículo, los autores se apoyan en el trabajo de Etzioni para hacer un estudio en profundidad de la Minería Web y proponen una taxonomía con tres categorías: **Minería Web de Contenido (minería de contenido de la Web), Minería Web de la Estructura (minería de estructura de la Web) y Minería Web de Uso (minería de uso de la Web).**

Se describe en detalle el ciclo de vida de la Minería Web, con sus etapas y sus tres categorías, además de sus diferentes aplicaciones y herramientas de *software* más populares y de utilización en Inteligencia de Negocios.

Otro tipo de Minería de Datos que se describirá es la Minería de Textos, muy empleada en la actualidad y con muchas semejanzas a la Minería Web de Contenido. De igual forma, se describirán herramientas populares de Minería de Textos.

9.2 MINERÍA DE TEXTOS

La era de *Big Data* en la cual vivimos se caracteriza, como ya conoce el lector, por el rápido crecimiento de la cantidad de datos —en consecuencia, información— que se recolectan, almacenan y se encuentran disponibles en las organizaciones y empresas, en Internet, y en infinidad de medios de papel y electrónicos.

Una gran mayoría de los datos de los negocios, las universidades, los centros de salud, etcétera, se almacenan como documentos de texto que, normalmente, son datos no estructurados o semiestructurados. Las innumerables fuentes de datos en formato de texto atesoran un gran conocimiento, que será muy eficiente en la toma de decisiones. Por esta razón, extraer (minar) conocimiento de los datos de texto es una necesidad vital en las entidades.

Esta inmensa cantidad de texto que se genera a diario y crece de modo exponencial es, esencialmente, texto no estructurado y no se puede procesar fácilmente por computadoras. Por consiguiente, se requieren técnicas eficientes y efectivas y algoritmos para descubrir patrones útiles. Estas técnicas incluyen preprocesamiento de texto, clasificación y *clustering*. Así nació la Minería de Textos, como la disciplina que tiene como objetivo principal la extracción de información significativa de texto. La Minería de Textos, Minería de Datos de Texto o Descubrimiento del Conocimiento en Texto, KDT (*Knowledge Discovery in Texts*) es un proceso —normalmente automatizado— de extraer patrones (información y conocimiento útiles) de grandes cantidades de fuentes de datos no estructurados.

El concepto de Minería de Textos o Descubrimiento del Conocimiento en textos, KDT (*Knowledge Discovery in Texts*) fue utilizado por primera vez por Feldman *et al* en 1995 y ha ganado una gran atención en los últimos años ante la tremenda cantidad de datos de texto que se genera cada día y que se crea en una gran variedad de formatos, tales como datos de redes sociales.

La Minería de Textos es un enfoque de la Minería de Datos, tiene el mismo propósito y utiliza los mismos procesos, pero con la Minería de Textos la entrada al proceso es una colección de archivos de datos no estructurados o semiestructurados, tales como documentos Word, archivos PDF, resúmenes de textos, archivos XML, etcétera. La Minería de Textos comienza con la recolección y almacenamiento de fuentes de datos (modelos de textos), su procesamiento y análisis para extraer información relevante y conocimiento de los datos basados en texto, mediante técnicas y herramientas de Minería de Datos. En síntesis, la

Minería de Textos es un enfoque de la Minería de Datos tradicional y métodos de descubrimiento de conocimiento.

Los beneficios de la Minería de Textos tienen gran aplicación en áreas donde existen o se generan grandes cantidades de datos, tales como bibliotecas, centros de investigación en universidades, finanzas, centros de salud (informes, recetas, análisis de enfermedades), medios de comunicación (revistas, periódicos), *marketing*—o mercadotecnia— (comentarios de clientes, socios, campañas de publicidad), etcétera.

En la actualidad, el correo electrónico, los sistemas de mensajería instantánea, sistemas de *chats*, comunicaciones electrónicas, entre otros, generan y distribuyen grandes cantidades de texto. Los conocimientos de estos contenidos serán de gran utilidad para las corporaciones propietarias de esos textos, aunque, se deberán tener siempre presentes las normativas legales del país donde resida la empresa propietaria de los textos.

9.3 HERRAMIENTAS DE LA MINERÍA DE TEXTOS

El número de herramientas de Minería de Textos ha ido creciendo a medida que crecían las aplicaciones y, al igual que sucede con otras herramientas de Minería de Datos, existe un amplio catálogo de herramientas de *software* propietario y de herramientas de *software* libre.

El reputado portal del conocimiento KDnuggets[1] —referencia obligada para todo tipo de recursos relacionados con *Business Intelligence*— publica y actualiza las herramientas de *software* de Minería de Textos[2] más populares utilizadas en la empresa. En su catálogo existe una gran cantidad de herramientas de *software* propietario (comercial) y *software* de código abierto. A continuación se describe una selección de herramientas recomendadas por KDnuggets:

Software propietario

- **Clear Forest.** Herramientas para análisis y visualización de colección de documentos.
- **IBM Intelligent Miner Data Mining Suite.**Integrado en herramientas de *software* InfoSphere Warehouse, incluye herramientas de Minería de Datos y de Textos.
- **IBM Infosphere Warehouse Enterprise Edition.**Incluye analítica avanzada, OLAP, Minería de Datos y análisis de texto.
- **IBM SPSS Predictive Analytics Suite.** Solución de Minería de Datos y Minería de Textos.

- **KNIME.** Plataforma de analítica *open source* que ofrece extensiones para análisis de texto, incluyendo Standfor de NLP, Palladin y Luguamatics.
- **Megaputer Text Analyst.** Ofrece análisis semántico de textos en formato libre, sumarización, agrupamiento (*clustering*), navegación y recuperación de lenguaje natural con enfoque en búsquedas dinámicas.
- **SAS Text Miner.** Proporciona una rica *suite* de procesamiento de texto y herramientas de análisis de texto.
- **VantagePoint.** Proporciona una variedad de vistas gráficas interactivas y herramientas de análisis con características potentes para describir conocimiento de bases de datos de texto.
- **Wordstat.** Analiza módulos de información textual tales como respuestas a cuestiones abiertas, entrevistas, artículos, etcétera.

Software de fuente abierta (*open source*)

- **Aika.** Biblioteca *open source*para minado de patrones frecuentes en texto, utilizando conceptos de redes neuronales y desarrollos gráficos.
- **GATE.** Herramienta *toolkit* líder en *open source* para Minería de Textos con un marco (*framework*) integrado *open source* (o SDK) y entornos gráficos de desarrollo.
- **LingPipe.** Es una *suite* de bibliotecas escritas en Java para el análisis lingüístico del lenguaje humano.
- **Microsoft Distributed Machine Learning Toolkit DMTK.** Incluye un marco de trabajo que soporta paralelización de datos, LightLDA, algoritmos de tópicos de modelos, Distributed (Multisense) Word algoritmos embebidos.
- **Rapid Miner.** Herramienta muy utilizada y de acceso gratuito, fácil de manejar ya que no requiere un conocimiento especializado de programación. Las aplicaciones incluyen, entre otras, Text Mining y Multimedia Mining
- **S-EM (Spy-EM).** Sistema de clasificación de texto que aprende con ejemplos positivos y sin etiquetas.

Otras aplicaciones de *software* de Minería de Textos de código abierto

En estos últimos tiempos se han consolidado tres herramientas de análisis de Minería de Textos no recogidas en las descripciones anteriores, pero de gran impacto y uso en las empresas:

R-lenguaje de programación y librerías de aplicaciones

Lenguaje de programación y *software* de Minería de Datos, cálculo y representación de gráficos. Dispone de una serie de paquetes específicos para Minería de Textos entre los cuales destacamos:

- **tm**. Específico para Minería de Textos.
- **qordcloud**. Realización de nubes de palabras.
- **readr**. Lectura y escritura de documentos.
- **cluster**. Análisis de grupos.
- **dplyr**. Funciones de manipulación y transformación de textos.

TwitterR

Es una aplicación que permite hacer minado de texto en Twitter con R a través de sus API, que permite obtener los datos que cede Twitter. El paquete posibilita conseguir los tuits de los usuarios de interés, seguidores, a quien sigue, etcétera, y además permite extraer los tuits relacionados con un tema determinado y la ubicación de los usuarios (geolocalización) que publican esos tuits.

Google Cloud Platform

Es una API de Google cada vez más utilizada por desarrolladores, que permite funcionalidades de bases de datos, almacenes de datos, aprendizaje automático y, en general, aplicaciones de *Big Data*. Alguna de las aplicaciones más eficientes de esta plataforma en Minería de Textos son análisis sintáctico de texto, análisis de opiniones, análisis de texto en varios idiomas y clasificación de documento según temas.

9.4 MINERÍA WEB: CONCEPTOS, DEFINICIONES Y CATEGORÍAS

La *World Wide Web* es un enorme depósito (repositorio) de datos de todos los formatos y de textos, y la cantidad de información existente en la actualidad crece de modo exponencial cada día. Existe un gran volumen de información existente en línea gracias a los numerosos sitios web, compuestos a su vez de un gran número de páginas web, enlazadas entre sí y con otros sitios web, tanto de empresas como de usuarios personales.

9.4.1 CRECIMIENTO EXPONENCIAL DE DATOS EN INTERNET

El volumen de datos crece exponencialmente en tiempo real: cada visitante de un sitio web, cada búsqueda en un buscador, cada clic en un enlace, cada mensaje enviado a través de sistemas de mensajería (como Whatsapp, Telegram o Snapchat), cada transacción en una tienda de comercio electrónico, crean datos adicionales y voluminosos. La fuente voluminosa de datos —especialmente no estructurados y semiestructurados— revelan patrones, tendencias y perspectivas en relación con el comportamiento de los usuarios y navegantes.

Existen numerosos estudios sobre la cantidad de información que se genera en Internet en determinados periodos de tiempo, donde se puede ver el crecimiento exponencial de los datos almacenados en la Web.

¿Qué sucede en un minuto en Internet?

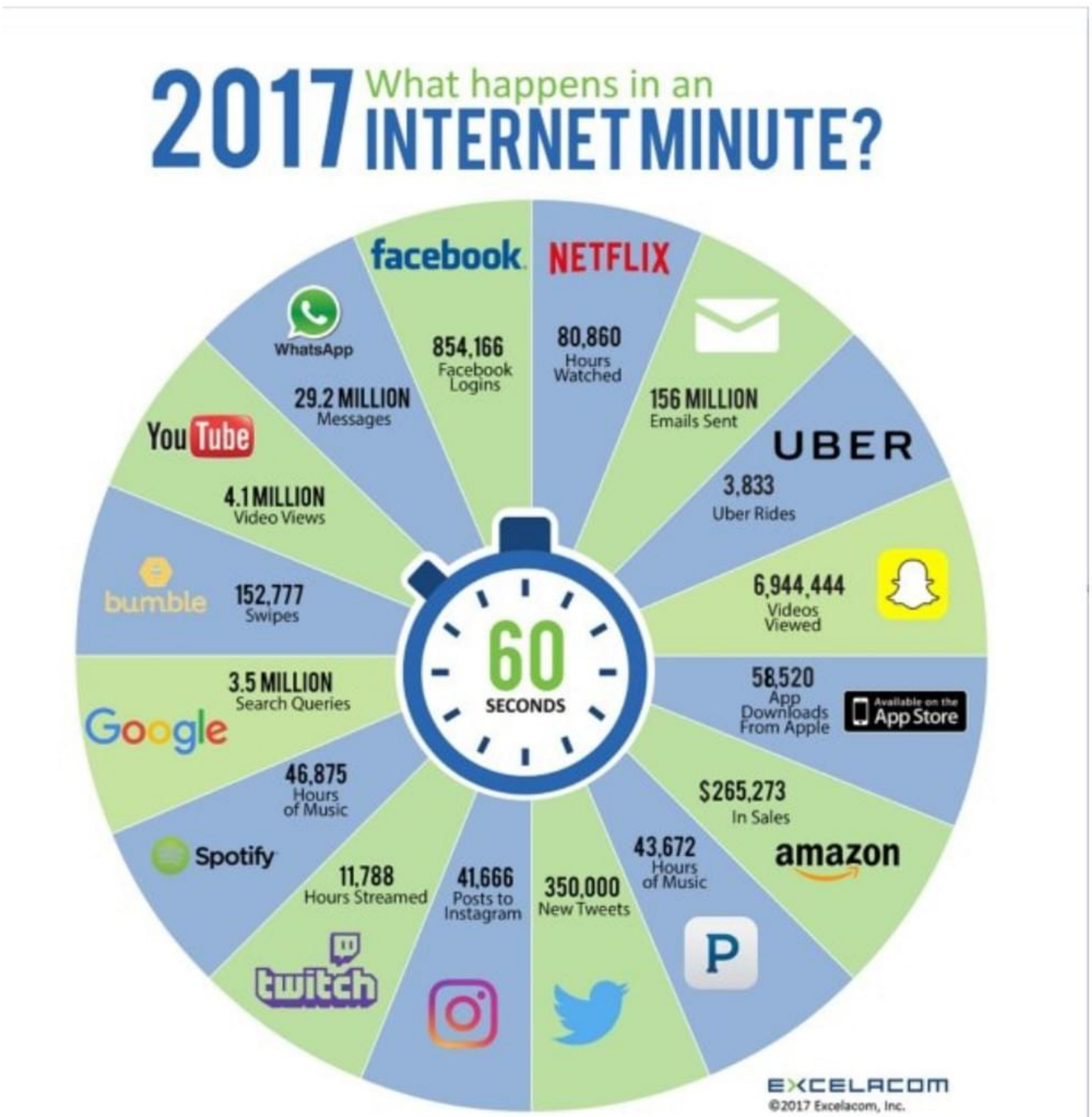




Figura 9.1 ¿Qué pasa en 1 minuto en Internet?
Fuente: Excelacom y @LoriLewis (Cumulus Media)

9.4.2. ANÁLISIS Y EVOLUCIÓN DE LOS DATOS DE LA WEB

Los datos de texto no estructurados en el formato de las páginas web, codificados en los lenguajes de programación HTML o XML, son los contenidos predominantes de la Web, aunque la infraestructura de la Web contiene información de hiperenlaces (conexiones o enlaces a otras páginas web) e información de uso (registros *logs* de interacciones de los visitantes en los sitios web). Todos estos elementos proporcionan gran cantidad de datos ricos en conocimiento, cuyo descubrimiento es una de las tareas más atractivas que pueden tener los usuarios, tanto sean personales como corporativos.

El análisis de esta información puede ayudar considerablemente a mejorar el uso de los sitios web, así como a mejorar las relaciones y el valor de los visitantes de nuestros sitios, pero el descubrimiento del conocimiento contenido en la inmensidad de sitios web no es fácil, precisamente, por las características propias de la Web. Turban (2011: 313), citando a Han y Kamber (2006),

menciona los grandes retos que afrontaba la Web para el descubrimiento de un conocimiento efectivo y eficiente:

- La Web es demasiado grande para un minado de datos eficiente.
- La Web es demasiado compleja.
- La Web es demasiado dinámica.
- La Web no es específica a un dominio.
- La Web tiene de todo.

Todos estos retos han originado grandes esfuerzos de investigación para mejorar la eficacia y eficiencia del descubrimiento y del uso de los activos datos de la Web. Un gran número de buscadores (motores de búsqueda), basados en indexación de las páginas web mediante el uso de palabras clave (*keywords*), permiten a los usuarios experimentados localizar documentos mediante la provisión de un conjunto de palabras clave o frases de texto. Sin embargo, un motor de búsqueda, aunque sea muy bueno (como es el caso de Google), tendrá varios inconvenientes para conseguir la respuesta exacta al concepto buscado. Así, por ejemplo, la búsqueda de un determinado "término" produce que el motor de búsqueda devuelva un gran número de entradas de documentos, muchos de los cuales no son relativamente relevantes para el término buscado; muchos documentos que son relevantes a un tema determinado pueden no contener las palabras clave que los definan fielmente.

En resumen, la búsqueda web basada en palabras clave, aunque es un procedimiento muy eficiente, sobre todo si se siguen las reglas dadas por los motores de búsqueda, presenta sus debilidades y, por estas razones, la minería en la Web es un enfoque más eficaz y prometedor, que se puede utilizar para mejorar la potencia de los buscadores, ya que, como veremos más adelante, la Minería Web puede identificar páginas web "autoridad", clasificar documentos web y resolver muchas ambigüedades y sutilezas planteadas por los buscadores basadas en la búsqueda por palabras clave. Servicios de Google como *Google Trends* y *Google Insights for Search*, rastrean la popularidad de las palabras y las frases utilizadas en las consultas de búsqueda de Google para conocer cuáles son los términos de interés de los usuarios y, por ejemplo, cuáles son los productos que les interesa comprar en un sitio de comercio electrónico.

Las empresas recurren a la Minería Web para que les facilite conocer el comportamiento de los clientes, evaluar la efectividad de un sitio web específico o cuantificar el éxito de una campaña de *marketing*.

La Minería Web o *Web Mining* consiste en aplicar las técnicas de Minería de Datos para descubrir y extraer automáticamente información de los documentos y servicios de la web. Minería Web o Minería de la Web (o Minería de Datos en la Web) es el proceso de descubrir relaciones intrínsecas (por ejemplo, información

útil y de interés) de los datos de la Web que se expresan en formatos de texto, vinculado o información de uso (Turban, 2011:313).

El término Minería Web (*Web Mining*) fue utilizado por primera vez por Etzioni (1996), el gran padre de esta disciplina, que definió Minería Web como el "proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web"[3]. Las técnicas de Minería de Datos, sin embargo, no son fácilmente aplicables a datos de la Web, debido a problemas relacionados tanto con la tecnología subyacente como con la ausencia de estándares en el diseño e implementación de páginas web. La información contenida en archivos *log* y otras fuentes de información debe ser procesada previamente a la obtención de los modelos.

En resumen, Minería Web es **el uso de técnicas de minería de datos aplicadas a la búsqueda, extracción y evaluación automática de información para conseguir el descubrimiento del conocimiento de los recursos de la Web: documentos y servicios.**

Las técnicas de Minería Web se apoyan, esencialmente, en las técnicas ya enunciadas de Minería de Datos, pero, sobre la base de todo lo expuesto anteriormente, presentan unas características propias, debido a las peculiaridades propias de sus sitios y páginas web:

- La mayoría de los datos de la Web son no estructurados.
- Los volúmenes de datos son muy grandes y crecen de modo exponencial, por lo que las técnicas de analítica de *Big Data* son imprescindibles.
- Los datos a nivel de páginas están relacionados mediante vínculos o hiperenlaces (*links*).
- Los datos tienen formatos muy variados, tales como HTML, PDF, imágenes, videos, audio, etcétera.
- Los datos son de muy diferente fiabilidad: se mezclan datos muy fiables con datos de escasa fiabilidad, lo que produce inconsistencia.

Por estas muchas razones, y otras que irán apareciendo, se han desarrollado técnicas específicas para gestionar los grandes volúmenes de datos presentes en la Web.

9.5 ARQUITECTURA DE LA MINERÍA WEB

Etzioni (1996), en el mencionado artículo sobre *Web Mining*, además de definir sus conceptos fundamentales, definió la organización de la arquitectura de la Minería Web en tres subtareas (etapas):

- Descubrimiento de recursos (*Resource Discovery*). Localización de los recursos de la Web, que clasificó en documentos y servicios desconocidos en la Web.
- Extracción de la información *(Information Extraction*). Extracción automática de información específica a partir de los nuevos recursos web descubiertos.
- Generalización (*Generalization*). Descubrimiento de patrones generales en sitios web individuales y en sitios múltiples.

Descubrimiento de recursos (fuentes). Etzioni, como se ha comentado antes, clasificó los recursos de la Web en dos categorías: documentos y servicios. El gran volumen de trabajos (ya en los primeros años de la Web, en 1996) en el descubrimiento de recursos (fuentes) se centraba en la creación automática de índices rastreables de documentos web. Los índices más populares en aquella época eran los robots de la Web (buscadores), tales como WebCrawler, Yahoo!y Altavista, que exploraban millones de documentos web y almacenaban un índice con las palabras de los documentos. En su trabajo, Etzioni menciona al metabuscador MetaCrawler, que en aquel entonces permitía hacer búsquedas en los buscadores existentes, tales como los ya citados y otros como Inktomi e Infoseek.

Cuando Etzioni publica en 1996 su artículo de referencia de la Minería Web, Google sólo era un algoritmo de búsqueda creado por Larry Page y Sergey Brin, estudiantes de doctorado en *Computer Science* en la Universidad de Standford, y el nombre original del buscador fue BackRuby cambiaron su nombre a Google en el año 1997 y, en 1998, crearon la empresa Google.

Esta etapa se comenzó a denominar **Recuperación de la Información** (**RI**) y busca **seleccionar documentos relevantes.**

Extracción de la información. Una vez que se ha descubierto un recurso web (recuperada la información), el reto, según Etzioni, era extraer información automáticamente del mismo. Para ello, había que acceder al recurso y analizar su respuesta. Ya anunciaba la introducción del concepto de mineros de datos de la Web que necesitaban extraer información de recursos desconocidos. Planteaba en esta etapa la selección y extracción de la información, y su posterior preprocesamiento. En aquella época, las herramientas más populares para la extracción de la información eran Harverst y Shopbot. La etapa de extracción de la información se centra en extraer información a partir de un documento (HTML, XML, texto, PDF, FAQ).



Figura 9.2. Etapas de la Minería Web[4] de Etzioni

Generalización. Una vez que se ha automatizado el descubrimiento y la extracción de la información procedente de los sitios web, el siguiente paso natural, según Etzioni, es intentar generalizar a partir de la experiencia acumulada. Ya planteaba el uso de agentes inteligentes y de técnicas de aprendizaje automático y la necesidad de que todos los sistemas virtuales de aprendizaje automático desplegados en la Web aprendan de los intereses de los usuarios en lugar de aprender de la propia Web. Ya en aquel entonces, las técnicas de Minería de Datos que proponía Etzioni eran las conocidas de *clustering* (agrupamiento), reglas de asociación, clasificación, reconocimiento de patrones, etcétera.

9.5.1 ETAPAS DE LA MINERÍA WEB (KOSALA Y BLOCKEEL)

El modelo de Etzioni fue actualizado por Kosala y Blockeel (2000), añadiendo una cuarta etapa, que implícitamente proponía Etzioni, que denominaron **Análisis: validación y/o interpretación de los patrones minados.** Las cuatro etapas de Kosala y Blockeel fueron una adaptación semántica del modelo de Etzioni, con una actualización (por los cuatro años de desfase) y la adición de la cuarta etapa antes citada.

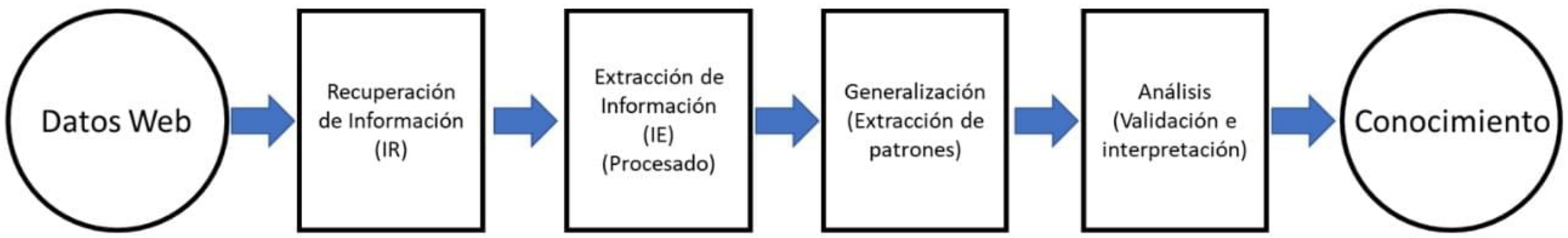


Figura 9.3. Etapas de la Minería Web
Fuente: (Etzioni 1996) y (Kosala, Blockee, 2000)

Recuperación de la información (RI). Descubrimiento de recursos: tarea de recuperar los documentos web relevantes o no usuales en la Web. Esta funcionalidad es la propia de los índices de los buscadores (extraen contenidos de palabras, idioma, partes del documento e índices temáticos para clasificar los documentos).

Extracción de la información (EI). Selección de la información y preprocesamiento. Selección automática de información específica a partir de los recursos web recuperados. Extrae determinada información a partir de un documento (HTML, XML, texto, PDF, LaTex, FAQ).

Generalización. Descubrimiento automático de patrones generales en sitios web individuales, así como en otros múltiples sitios web, mediante técnicas de agrupamiento (*clustering*), asociaciones entre documentos, etcétera.

Análisis: Validación e interpretación de los patrones o valores minados. El análisis de los datos es un problema de manipulación de datos, que requiere la

existencia de datos disponibles suficientes para que la información potencialmente útil se pueda extraer y analizar. Una vez que se han descubierto los patrones, los analistas necesitan herramientas apropiadas para comprender, visualizar e interpretar estos patrones. Se utilizan técnicas de análisis, desde las tradicionales herramientas OLAP hasta todas las nuevas técnicas de analítica de datos, analítica de big data, analítica web, analítica social que se describirán en los capítulos 10 y 11.

9.6 CATEGORÍAS DE LA MINERÍA WEB

Kosala y Blockeel (2000)[5], en su trabajo de investigación sobre Minería Web ya comentado, clasificaron a la Minería Web en diferentes tipos, que fueron agrupados en tres grandes categorías: Minería Web del Contenido, Minería Web de la Estructura y Minería Web del Uso (figura 9.4).

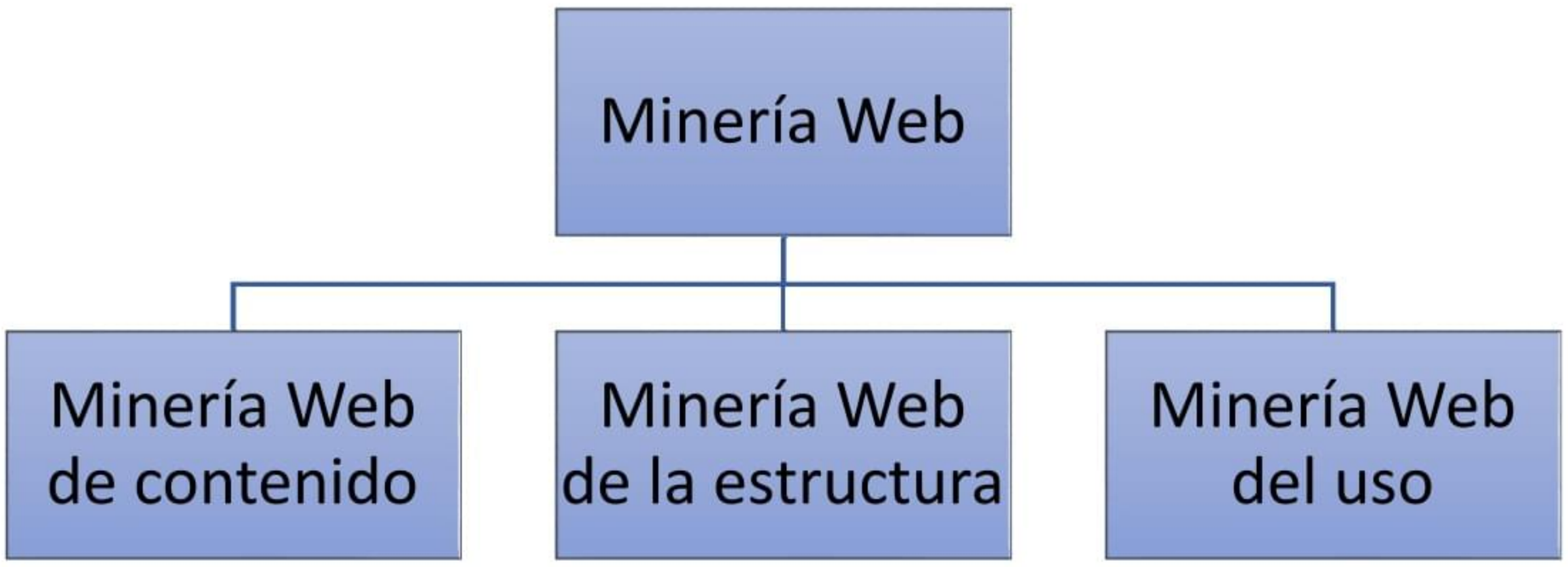


Figura 9.4. Categorías de Minería Web

Minería Web de Contenido (Web Content Mining). Se analiza texto no estructurado de las páginas web (normalmente en formato HTML: texto, imágenes y registros). Se extrae información del contenido de los documentos en la Web.

Minería Web de la Estructura (Web Structure Mining). Análisis de los enlaces contenidos en las páginas web (URL, Uniform Resource Locator: hiperenlaces, *tags*, etcétera). Se intenta descubrir un modelo a partir de la topología de los enlaces de la red. Este modelo puede ser útil para clasificar o agrupar documentos.

Minería Web del Uso (Web Usage Mining). Descripción detallada de las visitas a un sitio Web (secuencia de clics por sesión de visitas: *logs* http, *app server logs*, etcétera). Se intenta extraer información (hábitos, preferencias, etcétera, de los

usuarios o contenidos y referencia de documentos) a partir de las sesiones y comportamiento de los usuarios y navegantes.

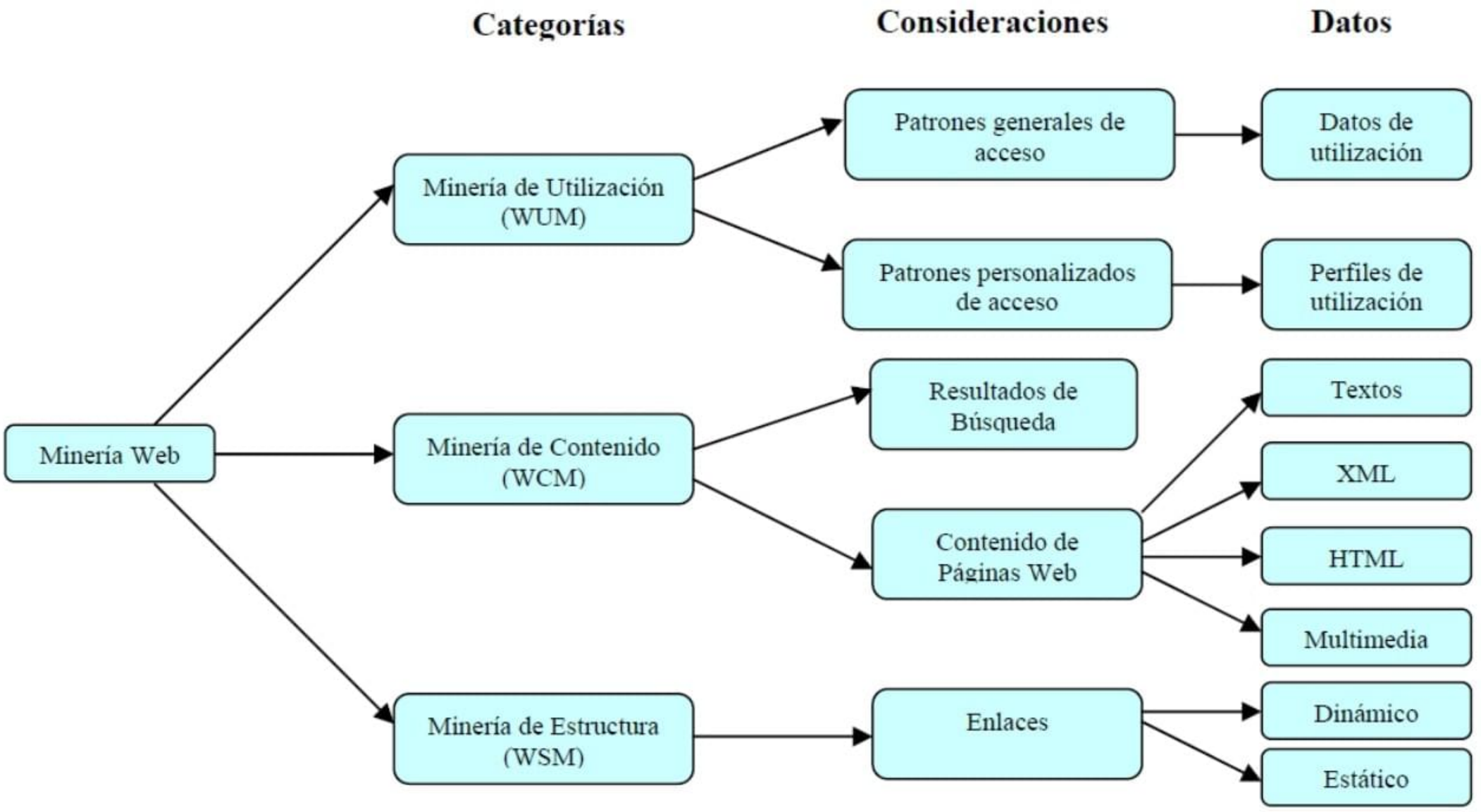


Figura 9.5. Tipologías de Minería Web
Fuente: Juan Carlos Dürsteler (Dürsteler, 2002)

La Minería Web combina objetivos y técnicas de distintas áreas:

- Recuperación de información (*IR, Information Retrieval*).
- Extracción de información (*IE, Information Extraction*).
- Procesamiento de lenguaje natural (NLP).
- Aprendizaje automático.
- Minería de datos.
- Bases de datos.
- Investigación en la *World Wide Web* y en la Web Semántica.
- Tecnologías de agentes inteligentes.
- Algoritmos de aprendizaje automático en Inteligencia Artificial.

9.7 MINERÍA WEB DE CONTENIDO

La Minería Web de Contenido o Minería del Contenido Web (*Web Content Mining*, **WCM**) se refiere a la extracción de información útil de las páginas web. Los documentos se pueden extraer en algún formato de máquina legible, de modo que las técnicas automatizadas puedan generar información sobre las páginas web. La técnica utilizada son los *Web crawlers* (robots de la Web), que se emplean para leer el contenido de un sitio web de modo automático. La información recolectada puede incluir características de documentos similares a la realizada en la Minería de Textos, pero también puede incluir conceptos adicionales, tales como la jerarquía de documentos. El minado del contenido de la Web se puede utilizar para mejorar los resultados producidos por los motores de búsqueda. Los *crawlers* (arañas, orugas) se utilizan para leer automáticamente en el contenido de un sitio web.

La Minería de Contenido Web es muy similar a la Minería de Textos y, de hecho, en ocasiones de tratan como sinónimos. La Minería de Contenido consiste en el escaneado y minado del texto, imágenes y gráficos de la página web para determinar el contenido relevante frente a la consulta acerca de las búsquedas.

Los sitios web están compuestos de colecciones de documentos de hipertexto. La recuperación de la información se realiza a través de la exploración sistemática de documentos, mediante dos enfoques: la Minería de Textos y el análisis semántico de los textos. Los sitios web son colecciones de documentos no estructurados, en los que se pueden descubrir y extraer esquemas para formularios, que capturen información semántica relevante de fuentes de datos heterogéneas. Los enfoques se basan, principalmente, en lenguajes de consulta de la Web (XML, WebSQl), múltiples bases de datos y descubrimiento de jerarquías.

9.8 MINERÍA WEB DE ESTRUCTURA

La Minería Web de la Estructura o Minería de Estructura Web (*Web Structure Mining*, **WSM**) es el proceso de extraer información útil de los enlaces embebidos en documentos web. Se utiliza para identificar **páginas web autoridad y páginas *hub***, que son las espinas dorsales de los algoritmos contemporáneos más populares de los motores de búsqueda de Google y Yahoo! Un enlace que va de una página web a otra puede indicar la popularidad del sitio la página autoridad y los enlaces dentro de la página web pueden indicar la profundidad de cobertura de un tema específico). El análisis de enlaces es un concepto muy importante en la comprensión de las interrelaciones entre grandes cantidades de páginas web y la importancia de ellas en la minería de conocimiento implícito en ellas.

Una página web contiene, además de texto, un elemento muy notable: los hiperenlaces (*hyperlinks*), que apuntan de una página a otra página. Los hiperenlaces contienen una cantidad significativa de conceptos humanos y sociales que apuntan al concepto de página autoridad. Supongamos que se desea buscar páginas web relativas a un tema determinado, tal como inversiones financieras, economía digital o economía colaborativa. Además de las páginas relevantes sobre el tema, se espera que las páginas recuperadas sean de alta calidad, o dicho en términos científicos, sean una autoridad (*authoritative of the topic*). Una **página autoridad** es una página de reconocido prestigio y con influencia en el sector al que pertenece. Los hiperenlaces generan el concepto de página autoridad. Cuando un diseñador de una página web incluye un enlace que apunta a otra página web, en general, está significando la importancia del contenido general o específico de la información de la página a la que apunta.

La búsqueda en la Web para obtener información sobre un tema específico, normalmente, devuelve unas pocas páginas web relevantes de alta calidad, pero también devuelve un gran número de páginas sin relevancia o inútiles para el usuario que realiza la consulta. De ahí que los usuarios expertos utilizan sus reglas prácticas para tratar de acertar en lo mayor posible en la entrega de resultados.

Normalmente, la cita de una página dada por diferentes desarrolladores o creadores de páginas web, creadores de *blogs*, redes sociales, etcétera, significa la importancia de esa página web y se suele considerar una página autoridad, como aquella a la que apuntan otras páginas web. El concepto de página autoridad es similar al caso del impacto de los autores de artículos de investigación (*papers*), cuya importancia se mide por el número de citas de otros artículos que hacen referencia a ese artículo en concreto.

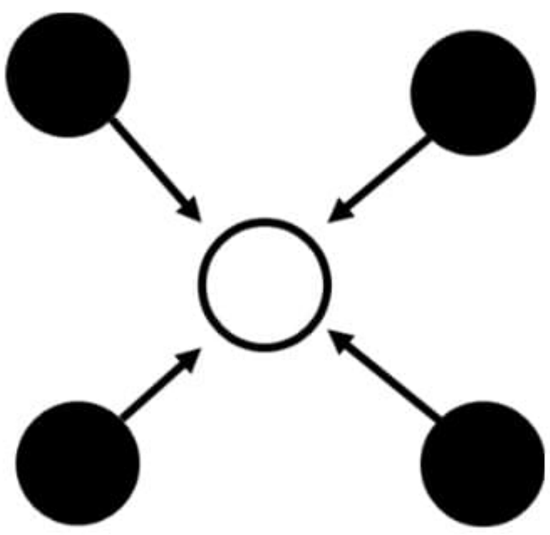
En la Minería Web de la Estructura se utiliza un índice basado en páginas autoridad, que mejorará los resultados de búsqueda, y un *ranking* de páginas importantes. Sin embargo, existe una diferencia importante entre las páginas web autoridad y una autoridad científica. En el primer caso, un hiperenlace representa una confirmación de su importancia, pero no siempre, porque en ocasiones los enlaces se crean para propósitos de navegación y en muchos casos como publicidad (de pago o gratuita); sin embargo, en el caso de la autoridad científica —casi siempre— sí se le suele citar por la importancia relevante del contenido y/o los autores. Otras diferencias residen en el caso de interés comercial o de competencia, en las referencias de páginas web, ya que, en muchos casos, las páginas web de autoridad en un sector a veces no apuntan a páginas de autoridad de la competencia en el mismo sector, o situaciones similares. Otra diferencia reside en que las páginas autoridad no siempre son descriptivas. Por ejemplo, las páginas web principales de Yahoo!, Google o Bing no contienen una descripción explícita del contenido ya que, de hecho, son motores de búsqueda y se referencia a ellos de un modo genérico.

La Minería Web de la Estructura consiste en estudiar la estructura de los enlaces "entre" e "intra" documentos. Es un grafo de enlaces donde cada página es un nodo y cada vínculo o hipervínculo de página a página constituye un arco dirigido con las siguientes reglas: los enlaces duplicados se ignoran y los enlaces entre páginas del mismo dominio también se ignoran, ya que suelen tener la funcionalidad de navegación. En estos grafos web, una página (de una persona, artículo u organización) se ve reforzada por la cantidad de referencias (amistades, citas, etcétera) que tiene.

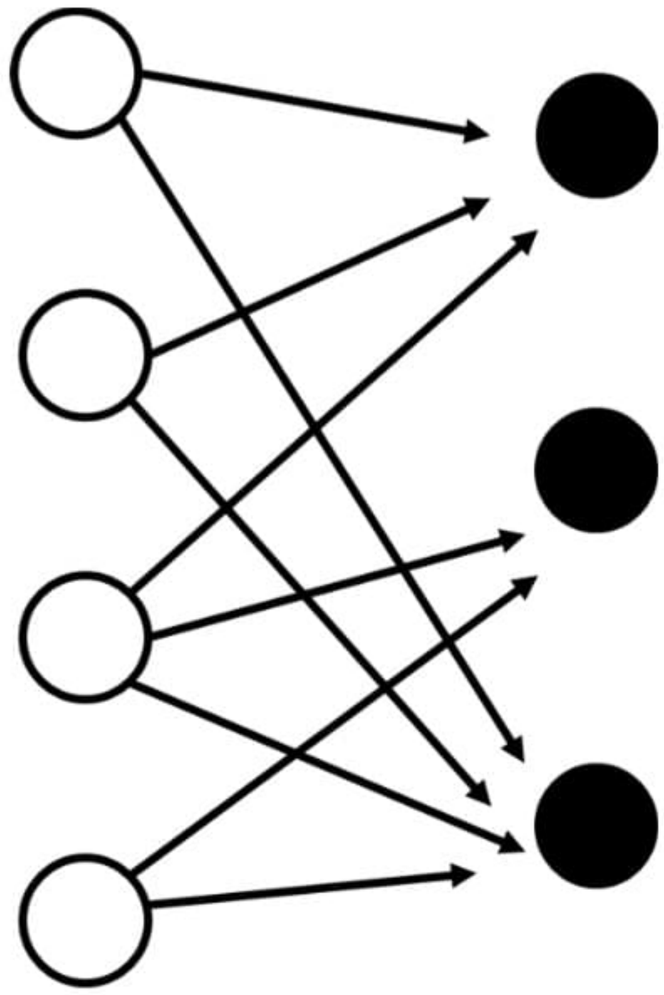
9.8.1 *HUB*

La estructura de los hiperenlaces de la Web tiene otras páginas de gran importancia, que son las páginas *hub (concentradores o agrupadores).* Una página web es un *hub* si proporciona una colección de enlaces a páginas autoridad. Las páginas web *hub* pueden no ser destacadas y sólo unos pocos enlaces apuntan a ellas; sin embargo, estos enlaces proporcionan un conjunto de sitios destacados sobre un tema específico de interés.

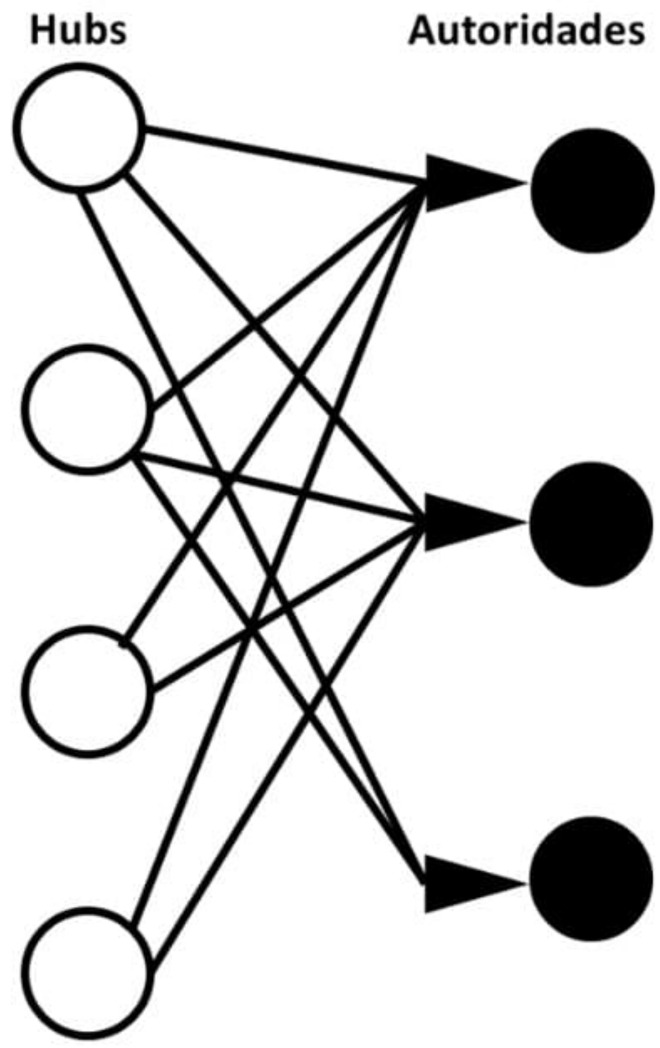
Un *hub* puede ser una lista de enlaces recomendados en una página web personal (o corporativa), sitios web de referencia o una lista de recursos recomendados (al estilo de las referencias bibliográficas o recursos web citados o recomendados en un libro). Existe una relación estrecha entre una página web *hub* y una página web autoridad. Un buen *hub* es una página que apunta a muchas buenas páginas autoridad; una buena página autoridad es una página apuntada por muchos buenos *hubs* (Han y Micheline, 2006). Estas interrelaciones entre páginas *hub* y páginas autoridad hace posible la recuperación automática de contenido de alta calidad de la Web y ayuda al minado de las páginas web autoridad y al descubrimiento automatizado de estructuras y recursos de alta calidad.



Representación del concepto autoridad



Representación del concepto hub



Representación de los conceptos autoridad y *hub*.

Figura 9.6

9.8.2 ALGORITMOS DE PÁGINAS AUTORIDAD Y PÁGINAS *HUB*

Existen numerosos algoritmos para calcular páginas autoridad y páginas *hub*. Uno de los más reconocidos es **HITS** (*Hyperlink-Induced Topic Search*). Fue originalmente desarrollado por Kleinberg en 1999 y mejorado posteriormente por diferentes investigadores. El otro algoritmo más popular, por ser el soporte del buscador Google, es **PageRank**. Según los especialistas e investigadores de Minería Web, los dos algoritmos son muy similares en su concepción científica.

El algoritmo HITS analiza las páginas web, clasificándolas mediante la información del hiperenlace contenido en ellas. En el contexto de las páginas web, el algoritmo HITS es un algoritmo recursivo que calcula los valores de las páginas web autoridad y *hub* para cada documento recuperado. Los rangos de valores obtenidos se utilizan para indexar y priorizar la recolección de documentos generados en una consulta específica.

El algoritmo de PageRank fue presentado en la séptima conferencia de la *World Wide Web*. Está basado en el algoritmo del mismo nombre, diseñado por los creadores del motor de búsqueda de Google, Sergey Brink y Larry Page (en el año 1996), como consecuencia de un trabajo de investigación para su doctorado en la Universidad de Standford.

9.9 MINERÍA WEB DE USO

La Minería Web de Uso o Minería de Uso de la Web (*Web Usage Mining*) tiene como objeto la extracción de información útil de datos, generada a través de visitas y transacción en sitios y páginas web. Se centra en técnicas que pueden predecir el comportamiento del usuario cuando interacciona con la Web. Esta información reside principalmente en clientes web (por ejemplo, *cookies)*, servidores, *proxies* y servidores de *banners* (publicidad), principalmente. Los datos recopilados de forma automática por la Web y por los servidores representan el comportamiento de navegación de los visitantes.

Existen tres tipos de datos principales generados a través de visitas a sitios web (Massand et al, 2002)[6]:

- Datos generados automáticamente y almacenados en registros *logs* de acceso a servidores, *referrers logs, agents logs* y *cookies* del lado del cliente.
- Perfiles de usuario.
- Metadatos, tales como atributos de páginas, atributos de contenido y datos de uso.



Figura 9.7. Etapas de la Minería Web del Uso

El proceso de la Minería Web de Uso comprende las siguientes etapas: recolección de datos de uso, preprocesamiento de datos de uso, descubrimiento de patrones de uso y análisis de los patrones de uso o análisis de la información recolectada (figura 9.7).

El análisis de la información recopilada por los servidores web puede ayudar a comprender el mejor comportamiento del usuario. El análisis de éstos se denomina, técnicamente, análisis de flujos de clics (*click stream analysis*). Una empresa puede ser capaz de discernir patrones interesantes a partir del flujo de clics, utilizando técnicas de Minería de Datos y de Textos.

El análisis de clics puede ser útil para conocer cuándo los visitantes acceden a un sitio. Así, por ejemplo, si una empresa sabe que el 60% de las descargas de *software* de su sitio web se produce entre las 17.00 y las 23.00, se puede planificar mejor el apoyo a esos clientes y la necesidad de ancho de banda de la red durante esas horas; y si la empresa es española y comprueba que un gran número de clientes proceden de países latinoamericanos, y que esas bandas horarias coinciden con horarios laborales de los mencionados países, deberá tener presente estas informaciones importantes para presentar posibles informaciones de interés para esos clientes personales o empresariales.

Turban (2011) muestra el proceso de la extracción del conocimiento a partir de los datos de los flujos de clic de la Minería Web de Uso y cómo el conocimiento generado se utiliza para mejorar el proceso, mejorar el sitio web y, lo que es más importante, cómo incrementar el valor del cliente. Los datos del usuario/cliente extraídos del sitio web, y de los registros *logs* correspondientes, se procesan mediante recopilación, fusión, limpieza y estructuración (identificar usuarios, identificar sesiones, identificar vistas de la página web, identificar visitas); una vez procesados, se extrae el conocimiento y se obtienen patrones de uso, perfiles de usuario, perfiles de la página, perfiles del visitante, valor del cliente; todos estos datos incrementan el valor del cliente y se actualiza el sitio web adaptándolo al conocimiento extraído de los clientes y usuarios.

Algunas aplicaciones típicas de la Minería de Uso de la Web son:

- Determinación del valor del tiempo de vida del cliente.
- Diseñar estrategias de *marketing* cruzado a través de productos.
- Evaluar campañas promocionales.
- Programar publicidad basada en patrones de acceso del usuario.
- Predecir comportamiento de uso basado en reglas aprendidas con anterioridad y de los perfiles del usuario.
- Predecir información dinámica de los usuarios basada en sus intereses y perfiles.

CASO DE ESTUDIO: AMAZON

Amazon se ha convertido en una empresa modelo de aplicación de la Minería Web y, en particular, de la Minería Web del Uso. Es un caso referente de cómo potenciar dinámicamente el uso histórico de los datos de sus clientes. El caso más patente y eficiente se muestra cuando un usuario está registrado en el sitio web —máxime si, además, es cliente habitual—.

Amazon saluda al cliente por su nombre y le presenta las últimas noticias de su interés. Esta tarea sencilla implica, simplemente, el reconocimiento del usuario a

través de la lectura de un archivo *cookie* (un pequeño archivo de texto escrito por un sitio web en la computadora del visitante). Presenta al usuario una selección de productos personalizados de su almacén, basada en compras anteriores y en un análisis de asociación con usuarios similares (en el caso de visitar el sitio de libros, suelen aparecer mensajes tales como "personas que compraron tal libro, también compraron tal otro" o, simplemente, recomendaciones similares de libros más vendidos, etcétera). Presenta a los clientes ofertas personalizadas, una selección de productos también personalizados, preferencias de usuarios de comportamientos similares por compras, zonas geográficas, horas de compra, etcétera.

Amazon utiliza las técnicas de Minería de Datos y, en particular, de la Minería Web de Uso, mediante análisis detallados del visitante así como de los grupos de clientes con comportamientos similares, mediante técnicas de *clustering*, descubrimiento de patrones de secuencias, asociación y otros datos extraídos de Minería de Textos, así como de Minería de Contenidos y de Estructura.

9.10 HERRAMIENTAS DE MINERÍA WEB

Al igual que sucede con las restantes herramientas de *software* de Minería de Datos, son numerosos los proveedores de soluciones de *software* de Minería Web. Hemos recurrido a diferentes fuentes y, en particular, al portal KDnuggets, que tiene varias páginas especializadas en Minería Web y sus diferentes categorías. KDnuggets tiene una página web excelente titulada Software for Analithics, Data Science, Data Mining, and Machine Learning (`www.kdnuggets.com/software/index.html`) con numerosa documentación. En particular las páginas dedicadas a minería web y de texto son:

1. Web Mining and Web Usage Mining Software

`https://www.kdnuggets.com/software/web-mining.html`

Esta página contiene su listado de herramientas de software de Minería Web: comercial, gratuitas y de código abierto, otros sitios de WebMining Software. Además, se pueden ver enlaces a otras páginas de su propio sitio web tales como Web Analytics Software, Web Content Mining, Web Mining Solutions y Web Searching Software.

2. Web content mining

`https://www.kdnuggets.com/software/web-content-mining.html`

3. Web Content Mining, Screen Scraping

`https://www.kdnuggets.com/software/web-content-mining.html`

4. Text Analysis, Text Mining, and Information Retrieval Software

`https://www.kdnuggets.com/software/text.html`

Dada la larga lista de soluciones de *software* que ofrece KDnuggets, hemos seleccionado algunas de las herramientas que consideramos más populares y más utilizadas por las empresas:

Comerciales:

- Angoss Knowledge Miner.
- Megaputer Web Analyst.
- Nihuo Web Log Analyzer.
- SPPS Web Mining for Clementine (herramienta específica de Minería Web de IBM SPSS)
- Web Trends (*suite* muy popular de *Data Mining* con información de tráfico web).

Código abierto y gratuitas (*open source* y *free*):

- AlterWind Log Analyzer Lite - free web analyser `http://www.alterwind.com/loganalyzer/log-analyzer-lite.html`
- htminer. `http://www.htminer.org/`

9.11 MOTORES DE BÚSQUEDA (BUSCADORES)

Los buscadores o motores de búsqueda (*search engine*) permiten a los usuarios de Internet acceder y encontrar información específica, buscando entre los miles de millones de páginas web existentes en la Red. Los motores de búsqueda son grandes bases de datos que almacenan cantidades masivas, provenientes de las diferentes partes del mundo.

Google sigue siendo el buscador más utilizado en todo el mundo y, según los países y las estadísticas, su uso puede variar de un 70 a un 90%. De igual forma, le siguen en popularidad Bing de Microsoft y Yahoo! En el ámbito asiático, y, en particular, en China, el buscador competidor de Google es Baidu, también con unos altos porcentajes de utilización.

Los motores de búsqueda actuales tienen una interfaz sencilla. En el caso de Google, un logo, un cuadro o ventana donde introducir los términos de búsqueda, dos ventanas más de "Buscar en Google" y "Voy a tener suerte" (dirige directamente al primer resultado) y opciones de menú (como Gmail, imágenes, menú de aplicaciones, notificaciones y cuentas de Google).

9.11.1 ARQUITECTURA DE UN BUSCADOR

La mayoría de los buscadores se componen de una serie de capas:

- Interfaz de usuario para realizar las consultas. Debe ser lo más amigable posible para el usuario, como lo es ahora en los buscadores citados.
- Un robot (*spider*) que busca información en Internet.
- Una base de datos donde se han indexado los contenidos recolectados por el buscador.
- Un algoritmo que conecta las peticiones de los usuarios con la base de datos del buscador

Algoritmo de búsqueda

El algoritmo del motor de búsqueda es, posiblemente, el componente fundamental y más innovador. Existen numerosos algoritmos de búsqueda en motores de búsqueda (buscadores), pero los más populares son **PageRank**, de Google, y **HITS** (*Hyperlink-Induced Topic Search*), que fue creado por Jon Kleinberg en 1997 (en aquella época, investigador de IBM) y presentado oficialmente en 1998. El algoritmo más popular entre los buscadores es el de PageRank, por ser el algoritmo del motor de búsqueda Google, el buscador más utilizado en casi todo el mundo, diseñado y desarrollado por los creadores de Google -Sergei Brint y Larry Page- como producto estrella de su tesis doctoral. El algoritmo HITS es muy similar al de PageRank de Google; ambos nacieron en la misma época y se basan en conceptos básicos de análisis de citas en publicaciones científicas, un área de la Bibliometría. PageRank define un orden estático entre las páginas web en función de su prestigio y HITS establece un orden dinámico que depende de la consulta concreta.

Ambos algoritmos se presentaron oficialmente en 1998: **PageRank** en abril de 1998, en la *Seventh International World Wide Conferencie (WWW7*), mientras que **HITS** lo hizo en enero de 1998, en el *Ninth Anual ACM-SIAM Symposium Algorithms*. Los dos algoritmos buscan medir el prestigio del rango de las páginas.

De hecho, PageRank —el algoritmo de búsqueda de Google— es, sin duda, el algoritmo de referencia que, además, se renueva continuamente. El algoritmo ha de facilitar el proceso automático de recolección, compilación e indexación y búsqueda. Los motores de búsqueda no generan contenido, sino que se limitan a mostrar los contenidos que han sido generados por terceras partes y que se encuentran publicados en los diferentes servidores de Internet, y sobre los cuales no se tiene ninguna capacidad de actuación.

Los buscadores encuentran la información utilizando un agente o robot (*web crawler)*, que entra en una página web para capturar datos, identificar los enlaces

(hiperenlaces, *links*) y seguir esos enlaces, como cualquier acción del navegador. El robot va saltando de un enlace a otro y navega por diferentes páginas, de las que ha recopilado datos, y alimenta a su base de datos. Cuando el robot encuentra la información, se crea un índice con la información esencial de la página web, que se almacena en la base de datos (creación de una indexación) y cuyos contenidos se mostrarán a los navegantes cuando realicen su petición a través de la interfaz del buscador. La información indexada se muestra al usuario dependiendo del algoritmo de búsqueda y teniendo presente una serie de indicadores que conforman las entradas de dichos algoritmos.

9.11.2 Algoritmo de PageRank

El algoritmo PageRank se rige por fórmulas que se apoyan en diferentes indicadores. La idea fundamental reside en que el algoritmo asigna un valor numérico a cada página y organiza un *ranking* de puntuaciones, que servirán de base para la presentación de los resultados al usuario (de mayor a menor peso o ponderación en el *ranking*). El algoritmo clasifica en las primeras posiciones a las páginas que poseen el mayor valor numérico y en orden descendente. El algoritmo de PageRank mide si las páginas son enlazadas por otra página, de forma que la página mejor clasificada será no sólo la que cuente con mejores y óptimos contenidos sino, y sobre todo, la página web apuntada (enlazada) por otras páginas web. Mientras más páginas apunten a una página web determinada, aumentará el índice del *ranking* de la página. Es decir, las páginas con enlaces entrantes tienen mayor peso que las páginas con gran número de enlaces externos.

En resumen, Google dio prioridad desde sus orígenes al factor de importancia o popularidad de la página. Ordena los resultados de la búsqueda utilizando su propio algoritmo de PageRank, donde a cada página se le asigna un número en función de la cantidad de enlaces de otras páginas que le apuntan y, esencialmente, el valor o peso de estas páginas, además de otros criterios no públicos que tiene el diseño del algoritmo. La figura 9.8 muestra un ejemplo del funcionamiento del algoritmo, donde se pueden ver los pesos de páginas que apuntan a otras páginas y cómo los pesos de las páginas aumentan en función de la cantidad de páginas que apuntan a ella, y el peso de cada una de estas páginas entrantes.

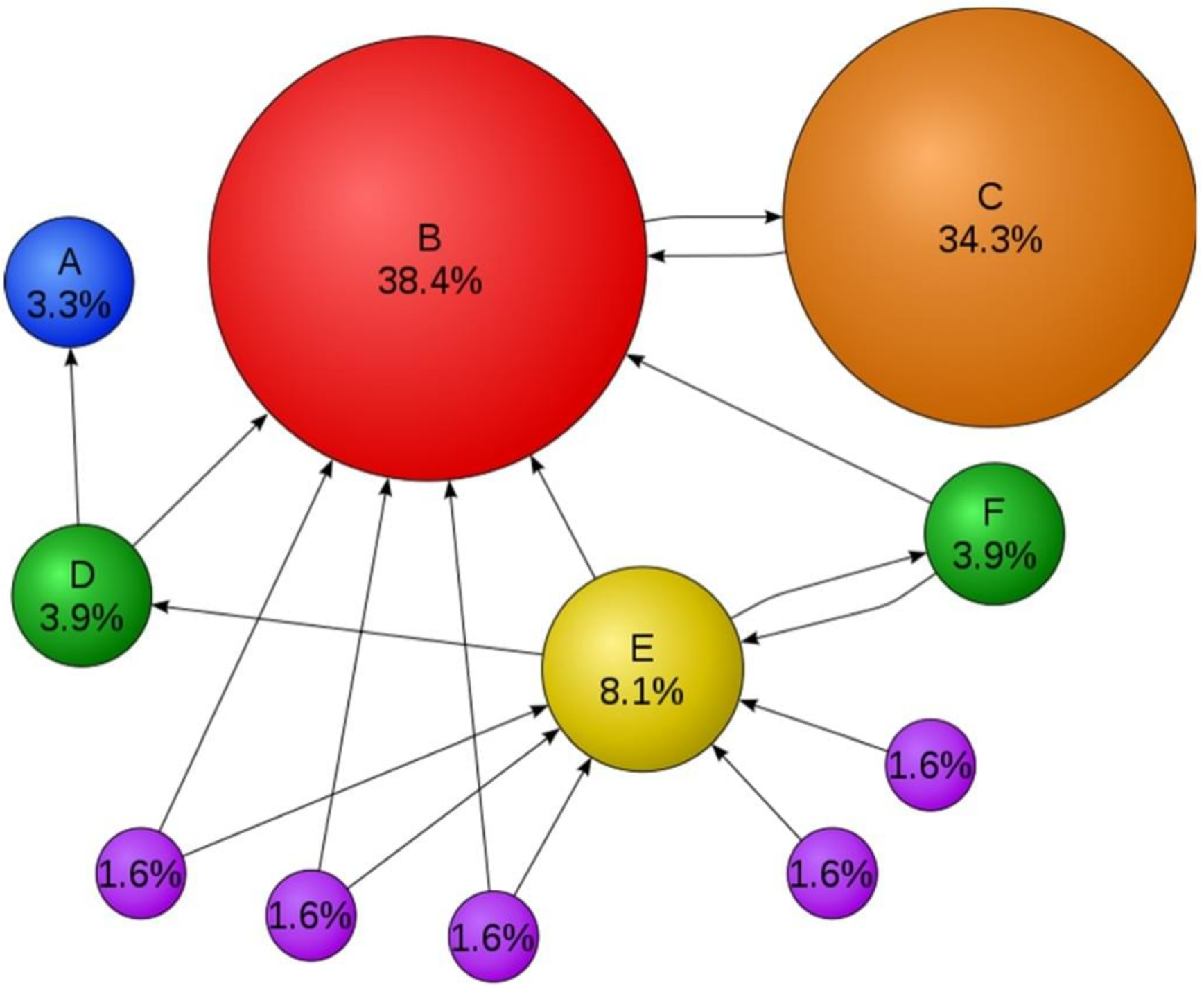


Figura 9.8. *Ranking* de páginas web en función de su peso (algoritmo PageRank)
Fuente: Wikipedia (https://es.wikipedia.org/wiki/PageRank)

En el caso de PageRank, Google lo ha ido actualizando de modo continuo, desde su primera versión en el año 1998. El algoritmo ha ido adoptando diferentes nombres según se actualizaba. Así, algunas de las actualizaciones más significativas del algoritmo en los últimos años son:

- *Allegra*, 2005.
- *Caffeine*, en 2009.
- *Panda*, 2011.
- *Venice*, 2012.
- *Penguin1*, 2012.
- *Pigeon,* 2014.
- *Mobilegeddo*n, 2015.

9.11.3 ALGORITMO HITS

El algoritmo HITS[7] propone que, dada una consulta de cada página existente en el grafo de la Web, existirán dos tipos de puntuaciones asociados: una puntuación *authority* (**con autoridad**), a la cual apuntan o enlazan muchas páginas, y una puntuación *hub* (**concentrador**), que consta de colecciones muy completas de enlaces a páginas valiosas sobre el tema específico. Para cada consulta se construye un subgrafo de la Web, donde se computa una puntuación de *hub* y otra de autoridad para el documento correspondiente donde:

1. Una página con alto puntaje de autoridad proveerá información relevante para la consulta.
2. Una página con alto porcentaje de *hub* proveerá enlaces a sitios relevantes para consultar.

Un buen *hub* incrementa el peso de autoridad de las páginas a las que apunta. Una buena autoridad incrementa el peso de la autoridad de las páginas a las que apunta. Una buena autoridad incrementa el peso del *hub* de las páginas a las que apunta. En síntesis, un buen *hub* apunta a buenas *authorities* y buenas *authorities* son apuntadas por buenos *hubs*.

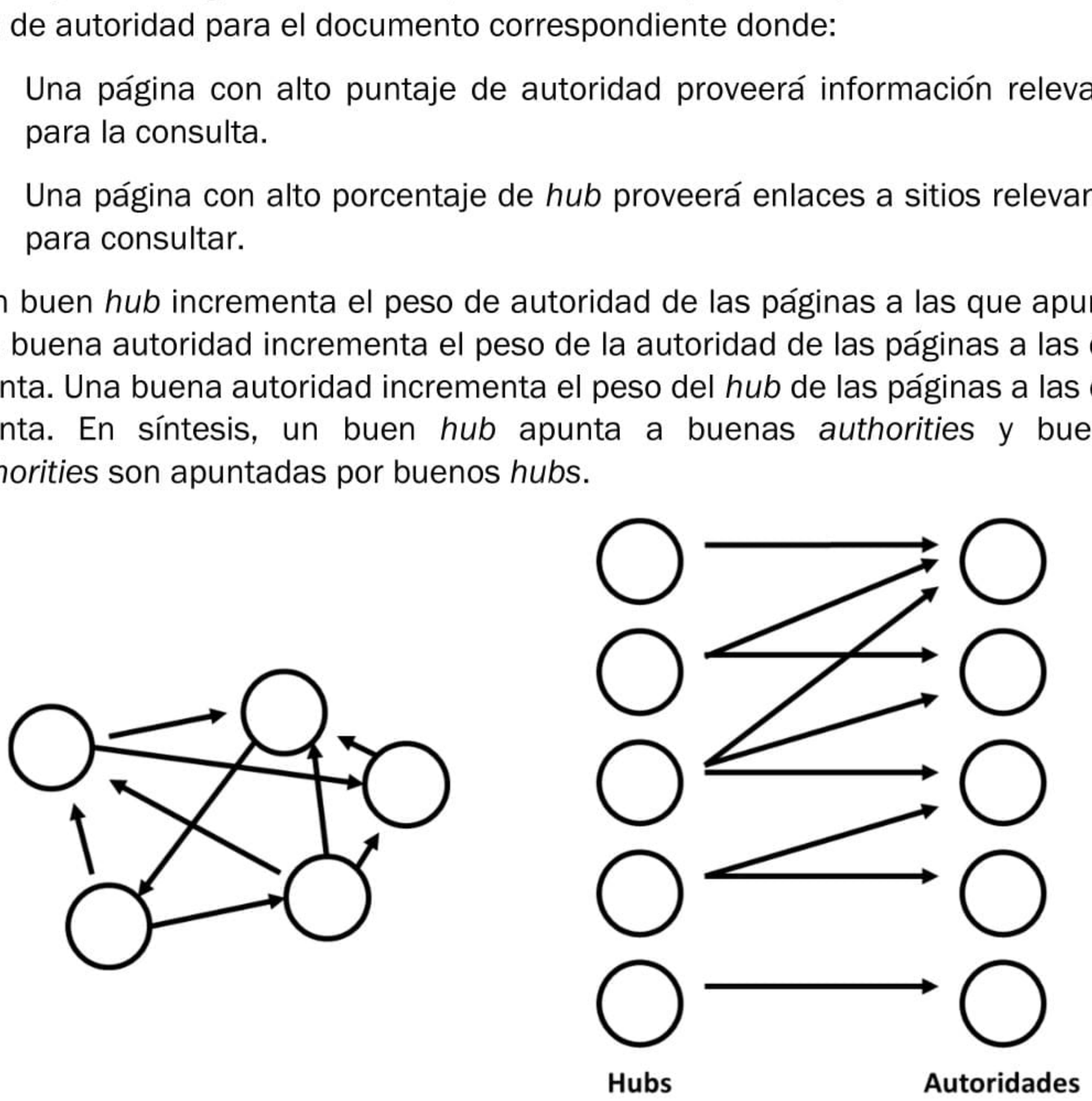


Figura 9.9. Relaciones entre páginas concentradoras (*hubs*) y autoridad (*authorities*)

Aunque existen diferencias técnicas entre ambos algoritmos, los tipos de medidas o puntuaciones se definen recursivamente: el prestigio de los nodos depende de otros nodos y la medida de ser un buen *hub* depende de la buena autoridad que sean los nodos vecinos (y viceversa).

9.12 POSICIONAMIENTO SEO: OPTIMIZACIÓN DE LOS MOTORES DE BÚSQUEDA

SEO es un término que tiene dos acepciones:

1. SEO (*Search Engine Optimization*). Optimización de motores de búsqueda o posicionamiento en motores de búsqueda.
2. SEO (*Search Engine Optimizer).* Optimizador de motores de búsqueda. Es la actividad o profesional dedicado a optimizar la presencia de un sitio web en los buscadores.

En ambos casos, el nexo común son los términos de búsqueda y el objetivo, aumentar el tráfico de calidad en un sitio web. El tráfico procede de la búsqueda orgánica —aquella en la que no se realiza ningún pago a los buscadores—. SEO se ha convertido en una estrategia que se centra en maximizar la visibilidad de un sitio o página web en los resultados orgánicos cuando un usuario busca por una palabra, o grupo de palabras de interés, para una empresa o un usuario a título personal.

El objetivo principal del SEO es mejorar el posicionamiento orgánico del sitio web, *blog*, red social, wiki, en los resultados de un buscador, cuando, como ya se ha comentado, un usuario busca una palabra, grupos de palabras, frases o preguntas. El SEO se caracteriza por centrarse en optimizar en los sitios web los elementos o contenidos que los buscadores toman en cuenta para posicionar dichos sitios web en sus resultados y conseguir que mejoren su *ranking* y visibilidad en los resultados de búsquedas orgánicas, de forma que el público objetivo pueda contactar más fácilmente con la página web.

El sitio web debe estar optimizado de la mejor manera para tratar de aparecer lo más arriba posible en los resultados de los buscadores de Internet, cuando una persona realiza una determinada búsqueda relacionada con la actividad de la empresa u organización, en una página web, *blog*, wiki, red social de la misma, bien a título corporativo o bien como usuario personal. Las páginas web mediante técnicas de SEO deben ser diseñadas y construidas de modo que puedan ser encontradas fácilmente en los buscadores. Las técnicas SEO deben ayudar a estar mejor situadas en los motores de búsqueda y, en consecuencia, aumentar su visibilidad.

9.12.1 ESTRATEGIAS DE POSICIONAMIENTO

Las estrategias de posicionamiento tradicionales para conseguir los mejores resultados en los buscadores y aparecer entre las primeras posiciones son:

- **Posicionamiento orgánico SEO.** Es el posicionamiento web natural y se consigue por diferentes acciones, entre las que destacan el trabajo realizado en el contenido publicado, los términos de búsqueda más interesantes (palabras clave), optimización interna de los sitios web, obtención de enlaces externos y conversaciones generadas. El posicionamiento SEO requiere una labor constante de optimización y análisis y medición de resultados.
- **Posicionamiento de pago SEM.** No responde a criterios orgánicos o naturales de los buscadores, sino a la creación de campañas específicas de publicidad, que posicionan un determinado sitio mediante el pago de una cantidad variable por parte del contratante en función de un sistema, normalmente, de subastas. En el caso de Google, las aplicaciones Google AdWords y Google AdSense facilita esta tarea. La publicidad consiste en colocar anuncios patrocinados para las claves de mayor interés para la empresa.
- **Posicionamiento SEO local.** Consiste en posicionar los negocios cercanos al usuario que disponen de una dirección física o área de servicio. Según el ámbito del negocio, el posicionamiento web será diferente. El posicionamiento a nivel local adapta los negocios que tienen presencia física y desarrollan sus actividades en un local o en una zona geográfica determinada. El posicionamiento SEO local optimiza la presencia de los negocios físicos en los resultados de búsqueda local para estar visible ante los consumidores más cercanos a ellos. El SEO local busca aparecer en las primeras posiciones de la página de resultados, no sólo como resultado orgánico sino también con información de posición geográfica (nombre, dirección, teléfono, redes sociales, horarios de atención, rutas al negocio, etcétera) mediante técnicas de geoposicionamiento.

Una estrategia de geoposicionamiento local requiere de un diseño local del sitio web, que incluya datos tales como el nombre de la ciudad del negocio en la estructura de la página web (en el título, en la descripción, URL, imágenes, contenidos), de modo que los buscadores localicen dónde está ubicado el negocio y cómo llegar allí. Una característica importante es incorporar un mapa embebido en el sitio web de la empresa en Google+, Facebook o Twitter. Se trata de registrar la página web como negocio local, permitir que los usuarios puedan puntuar y opinar sobre el establecimiento, etcétera (en la actualidad, incorporar asistentes virtuales de la empresa y que se visualicen de inmediato).

El SEO local es un conjunto de técnicas dirigidas a aumentar la visibilidad de una página web en las búsquedas de información relevantes para los usuarios, en función de la ubicación geográfica actual de los mismos. Se trata de que el sitio web de una empresa con un local comercial físico preste un servicio en un área geográfica específica, sea encontrado por los potenciales clientes que se encuentran en esa zona y realicen búsquedas relacionadas con el negocio.

9.12.2 CRITERIOS DE OPTIMIZACIÓN

Un motor de búsqueda se basa en varios factores o criterios de optimización para posicionar una página web, principalmente:

- **Accesibilidad.** El buscador debe poder acceder con facilidad al sitio web y a todas las páginas que contiene. Debe poder moverse con facilidad por el sitio web (**rastreabilidad**) para comprender bien el contenido de cada página web y poder mostrarlo en los resultados. En consecuencia, se ha de garantizar que el contenido de la web es accesible y comprensible para que pueda ser encontrado, valorado y guardado por los buscadores.

- **Indexabilidad.** El sitio web debe haber sido indexado previamente para que pueda aparecer como resultado en un motor de búsqueda. La estrategia SEO debe asegurar que el sitio web sea **indexable**, es decir, la web corporativa debe reunir las características necesarias para que los buscadores puedan leer su contenido (palabras clave, etiquetas, enlaces, etcétera), enlazado interno del sitio, estructura del sitio web, configuración adaptada para dispositivos móviles, velocidad de carga del sitio, carga rápida y que "no se caiga".

- **Autoridad.** La autoridad de un sitio web es la cantidad de enlaces externos de relevancia (o importancia) que apuntan al mismo. La autoridad es también la popularidad del sitio, ya que cuanto más popular sea una página más valiosa será la información que contiene. La compartición de un contenido se debe, esencialmente, a la utilidad y valor de sus usuarios (**experiencia del usuario**). Se deben crear campañas de consecución de enlaces entrantes, citas y referencias que potencien las señales de popularidad, indicador muy importante para los buscadores. Es importante el desarrollo de la autoridad del sitio web en su sector de negocios o comercial y la temática que se ha de posicionar para potenciar la campaña de enlaces externos.

  Este criterio está muy relacionado con la relevancia de la página.

- **Relevancia.** El sitio web ha de ser relevante para los usuarios. La relevancia de un sitio web se mide también por la valoración por parte de los usuarios: cuanta más gente acceda, cuanto más y mejor lo enlacen y cuantos más votos favorables reciba de ellos, el sitio web será más relevante. Debe existir coincidencia del contenidos y elementos del mismo con los términos empleados por los usuarios en su búsqueda.

- **Recomendación.** Garantizar que los textos, las imágenes, las fotografías, los videos incluyan contenido con los términos relevantes, información única y actualizada; informar sobre el significado del contenido utilizando técnicas de marcado estructurado, así como técnicas de buenas prácticas.

- **Confiabilidad.** Existe un gran número de características que debe cumplir un sitio web para que sea confiable para los buscadores. En las numerosas

obras dedicadas a *marketing* digital se recoge una gran cantidad de características. Algunas de las más enunciadas son:

- Carga rápida. El sitio web debe estar en un buen alojamiento o *hosting* que garantice un alto nivel de disponibilidad y su diseño debe estar optimizado para que se cargue rápidamente.
- El nombre del dominio es antiguo.
- Compatible con todos los navegadores.
- No se "cae" (se desconecta) con frecuencia.
- Carga bien en dispositivos móviles.
- Los sitios web más confiables tienen dominios .edu y .org.

9.12.3 FACTORES DE POSICIONAMIENTO

El posicionamiento orgánico (de pago) en buscadores es una tarea compleja que requiere de estrategias y profesionales especializados para llevarlas a cabo. En la consecución de un buen posicionamiento en los buscadores influyen numerosos factores que, a su vez, contemplan dos escenarios principalmente.

- **SEO *on-site*** *(on-page).* Acciones que se realizan en la página o sitio web para mejorar su optimización y que el buscador comprenda bien el contenido. Algunas de estas acciones son: optimización de palabras clave, tiempo de carga, experiencia del usuario y formato de las direcciones URL.
- **SEO *off-site*** *(off-page).* Acciones que se centran en factores externos a la web y que se han de realizar para aumentar la presencia en Internet. Los factores más importantes son: número y calidad de los enlaces, presencia en redes sociales, menciones en medios de comunicación, autoridad y relevancia de la marca, resultados de búsqueda. En resumen, promoción de la marca que ayude a aumentar el tráfico al sitio web.

Otros factores a considerar en el posicionamiento SEO y de impacto en el *ranking* de resultados en buscadores son.

- Calidad del contenido del sitio (texto, imágenes, perfil). Ofrecer contenidos que aporten valor y proporcionen información útil para los usuarios es la clave para ser visitados y posicionar nuestra marca en Internet.
- Conocer cuáles son las palabras clave que más conviene utilizar para captar tráfico cualificado. Las palabras clave han de ser términos relacionados con el nicho de mercado que faciliten a los buscadores identificar la temática de los contenidos y poder ofrecerlos a su público objetivo.
- Factores técnicos de la Web (elección de un buen dominio, etiquetas meta o *meta tags*, URL amigable, optimización de imágenes).

- Experiencia de usuario: diseño, usabilidad, señales de usuario. Google considera proporción de clics en relación con el número de usuarios, páginas vistas, tasas de rebote, tiempo de permanencia en el sitio web, etcétera.
- Número y calidad de los enlaces entrantes (enlazabilidad).
- Señales sociales (interacción, recomendaciones, "me gusta", retweets[8], +1=).

9.13 POSICIONAMIENTO SEM

SEM, siglas de *Search Engine Marketing*, es el posicionamiento de pago que no responde a criterios orgánicos o naturales de los buscadores, sino a la creación de campañas específicas que posicionan un determinado sitio web y que, normalmente, es un sistema de subastas. La principal diferencia entre las acciones de SEO y de SEM es que las acciones de SEO son gratuitas y orgánicas o naturales, mientras que las acciones de SEM son de pago.

El sistema básico de funcionamiento de SEM se apoya en un método conocido como **PPC** (*Pay Per Click*) o **CPC** (coste por clic) o **PPC** (pago por clic). En este sistema, el propietario de un sitio web sólo paga cuando un usuario hace clic en su anuncio y se dirige al sitio o página web. Si el anuncio no recibe ningún clic y, por consiguiente, ninguna visita, no se realiza ningún pago y el coste será cero.

SEM es un sistema de publicidad creado por Google y apoyado en herramientas de *software* específicas, en concreto Google AdWords, que permite al cliente mostrar anuncios en las páginas de resultados del buscador. Estos anuncios se publican junto a los resultados naturales, normalmente en las primeras posiciones y en el lateral de la pantalla, aunque últimamente, en las versiones más actuales de Google, los resultados se publican en el mismo formato que los resultados naturales, pero en las primeras posiciones. El sistema SEM permite poner anuncios en sitios de la red de buscadores y asociados a Google.

Google AdWords es un programa de Google utilizado para realizar campañas de publicidad y anunciarse en el buscador o en su red de contenidos. El método de pago por clic sólo se paga por cada clic que recibe su anuncio, con independencia del número de veces que se muestre. Están identificados con la palabra "Anuncios". El pago por clic en los anuncios de AdWords es variable y se calcula en forma de subasta, de forma que el precio de una palabra será mayor cuanto más solicitada o popular sea la palabra, grupos de palabras o frases.

El sistema de presentación de resultados del buscador Google ha variado sensiblemente desde sus primeros anuncios. En la actualidad, la versión en español suele presentar los resultados de pago con el término "Anuncio", antes de la URL del sitio web publicitado, bien en la parte alta de la página de resultados o bien en la parte baja, y en número diferente de direcciones URL (cinco o seis anuncios por página).

9.14 RESUMEN

La Minería de Textos, Minería de Datos de Texto o Descubrimiento del Conocimiento en Texto, KDT (*Knowledge Discovery in Text*) es un proceso —normalmente automatizado— de extraer patrones (información y conocimiento útiles) de grandes cantidades de fuentes de datos no estructurados.

El concepto de Minería de Textos es similar al de Minería de Datos, tiene el mismo propósito y utiliza los mismos procesos, pero con la Minería de Textos la entrada al proceso es una colección de archivos de datos no estructurados o semiestructurados, tales como documentos Word, archivos PDF, resúmenes de textos, archivos XML, etcétera. La Minería de Textos comienza con la recolección y almacenamiento de fuentes de datos (modelos de textos), su procesamiento y análisis para extraer información relevante y conocimiento de los datos basados en texto, mediante técnicas y herramientas de Minería de Datos.

El número de herramientas de Minería de Textos ha ido creciendo a medida que crecían las aplicaciones y, al igual que sucede con otras herramientas de Minería de Datos, existe un amplio catálogo de herramientas de *software* propietario y de herramientas de *software* libre.

La Minería Web, o *Web Mining*, consiste en aplicar las técnicas de Minería de Datos para descubrir y extraer automáticamente información de los documentos y servicios de la web. El término Minería Web (*Web Mining*) fue utilizado por primera vez por Etzioni (1996), el gran padre de esta disciplina, que definió Minería Web como el "proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web". Las técnicas de Minería de Datos, sin embargo, no son fácilmente aplicables a datos de la Web, debido a problemas relacionados tanto con la tecnología subyacente como con la ausencia de estándares en el diseño e implementación de páginas web.

En síntesis, la Minería Web es el uso de técnicas de Minería de Datos aplicadas a la búsqueda, extracción y evaluación automática de información para conseguir el descubrimiento del conocimiento de los recursos de la Web: documentos y servicios. La Minería Web se organiza en tres subtareas (etapas):

- Descubrimiento de recursos (*Resource Discovery*).
- Extracción de la información (*Information Extraction*).
- Generalización (*Generalization*).

Las categorías de la Minería Web son: Minería Web de Contenido, Minería Web de la Estructura y Minería Web de Uso.

Los algoritmos más utilizados en los buscadores son PageRank, de Google, y HITS, de Klinberg.

Un concepto muy importante relacionado con la Minería Web es SEO (Search Engine Optimization, posicionamiento natural en buscadores, técnica utilizada para mejorar el posicionamiento orgánico del sitio web, blog, red social, wiki, etcétera, en los resultados de un buscador, cuando un usuario busca una palabra, grupos de palabras, frases y preguntas.

SEM, siglas de Search Engine Marketing, es el posicionamiento de pago que no responde a criterios orgánicos o naturales de los buscadores, sino a la creación de campañas específicas que posicionan un determinado sitio web en función del abono de una cantidad variable, normalmente, mediante un sistema de subastas. La principal diferencia entre las acciones de SEO y de SEM es que las acciones de SEO son gratuitas y orgánicas o naturales, mientras que las acciones de SEM son de pago.

NOTAS

[1] *KDnuggets. Text Analysis, Text Mining, and Information Retrieval Software*

`https://www.kdnuggets.com/software/text.html`

[2] `http://www.kdnuggets.com/software/text.html`

[3]Oren Etzioni. The World Wide Web: Quagmire or God Mine?. Communications of the ACM. Vol 39, 11, 85-68, noviembre 1996.

[4] *KOSALA, R. y BLOCKEEL, H. (2000). Web Mining Reseach: A Survey. ACM SIGKDD Explorations Newsletter of the SCMA Special Interest Group on Knowledge Discovery and Data Mining. Vol. 2 (1). pp. 1-15. En su profundo trabajo de investigación, citaron como precursores a L. Etzioni (1996), Madria et al (1999) y a Borges y Levene (1999).*

[5] *Ibidem. Raymond Kosala y Hendrik Blockee. Web Mining Research: A Survey . ACM SIGKDD Explorations Newsletter. Vol 2 (1). Julio 2000.*

[6] *Citado por Turban, p. 316.*

[7]`http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html`

[8] *La palabra retuit no está aceptada por la RAE. Sin embargo si admite retuitear (reenviar un Tweet a un número determinado de personas)*

CAPÍTULO 10
ANALÍTICA DE DATOS (*BIG DATA & ANALYTICS*)

CONTENIDO

10.1 Introducción
10.2 ¿Qué es Analítica de Datos? (*Data Analytics*)
10.3 Analítica de Negocios (*Business Analytics/Analytics*)
10.4 Una visión global de Analítica de *Big Data*
10.5 Categorías prácticas de Analítica
10.6 Analítica de *Big Data*
10.7 Características de una plataforma de integración de Analítica de *Big Data*
10.8 Analítica Digital
10.9 Analítica Web
10.10 Proliferación de datos sociales
10.11 Analítica Social
10.12 Análisis de Sentimientos
10.13 Analítica Móvil
10.14 RESUMEN

OBJETIVOS

- Definir los conceptos fundamentales de Analítica de Datos.
- Revisar los conceptos globales de Analítica de Big Data.
- Conocer las características fundamentales de una plataforma de integración de Analítica de Datos.
- Conocer y diferenciar los diferentes tipos de Analítica: Digital, Web, Móvil, Social, de Sentimientos.
- Introducción a las métricas y los KPI.
- Analizar el impacto del crecimiento de los datos sociales y la necesidad de una Analítica Social.
- Conocer los casos de estudio de gran impacto en las empresas: Google Analytics e IBM Watson Data Platform.

10.1 INTRODUCCIÓN

El crecimiento exponencial de datos en la última década ha de ser explotado de manera eficaz y eficiente por las organizaciones. Hoy en día, los datos no estructurados, que pueden llegar al 80% o más de la información de la empresa, afectan a las infraestructuras de cómputo (computadoras y servidores). El problema es que suelen ser difíciles de analizar y, en cualquier forma, el proceso puede durar mucho tiempo si no se tiene una formación adecuada.

En primer lugar, se requiere recolectar los datos y, a continuación, se deben organizar; una vez realizadas estas operaciones, hay que llevar a cabo los procesos de análisis (con procesos de descubrimiento, consultas e informes, Minería de Datos, etcétera) y la toma de decisiones mediante la planificación y predicciones adecuadas.

La analítica de *Big Data* y sus herramientas permiten a los usuarios analizar los datos masivos con tamaños desde terabytes hasta petabytes —y, cada vez más, exabytes— de un modo rápido y económico. Por su parte, los usuarios deben ser capaces de explorar y visualizar datos masivos mediante gráficos interactivos, Cuadros de Mando Integral, CMI (*balanced scorecards*), tableros de control o cuadros de mando (*dashboards*) y visualizadores de informes de resultados en tiempo real cuando sea necesario.

El tratamiento y análisis de grandes volúmenes de datos requiere de una gran potencia analítica. El análisis de *Big Data* debe ayudar a tomar mejores decisiones y evaluar las medidas que se han de tomar del modo más eficiente y rentable posible. En este capítulo, se estudiará el análisis de datos en sentido

general y una descripción más detallada de la Analítica de *Big Data*, Analítica Web, Analítica Móvil, Analítica Social y Analítica de Sentimientos.

10.2 ¿QUÉ ES ANALÍTICA DE DATOS? (DATA ANALYTICS)

Existen numerosas definiciones del término Analítica de Datos (*Data Analytics*), pero hemos decidido utilizar la definición de ISACA, organización profesional de impacto mundial en sistemas de información (gobierno, auditoría y seguridad), que publicó en agosto de 2011 un artículo (*whitepaper*) con el mismo nombre, para dar su opinión sobre un término de tanto impacto en el mundo corporativo. **Analítica de datos** (AD, *Data Analytics, DA) "implica los procesos y actividades diseñados para obtener y evaluar datos para extraer información útil"*[1].

Los resultados de la AD (DA) se pueden utilizar para identificar áreas clave de riesgos, fraudes, errores o mal uso; mejorar los procesos de negocios; verificar la efectividad de los procesos e influir en las decisiones del negocio. Existen muchos temas para considerar cuando arranca un nuevo programa de AD, incluyendo la maximización del retorno de inversión (ROI), cumplimiento de presupuesto de proyectos, gestión de falsos resultados, aseguramiento de la protección y confidencialidad de las fuentes de datos y resultados.

Hay una gran variedad de herramientas y métodos de *software* que se utilizan en Analítica de Datos. Las técnicas que más se emplean son: consultas e informes (*quering* y *reporting*), visualización, minería de datos, análisis predictivo de datos, lógica difusa, optimización, *streaming* de audio, video o fotografía, etcétera.

Analítica de Datos se considera también la ciencia de examinar datos en bruto (crudos), con el propósito de obtener conclusiones acerca de la información contenida en ellos. Se usa en muchas industrias para que las organizaciones y empresas tengan mejoras en la toma de decisiones. Este término suele emplearse en el campo de la Inteligencia de Negocios (*Business Intelligence*) y, según los fabricantes de herramientas de *software*, puede abarcar una gran variedad de términos: OLAP, CRM, *dashboard* (tableros de control), etcétera.

Hoy en día, la Analítica de Datos está influenciada por todo tipo de dispositivos y medios sociales, como los datos procedentes de GPS, chips NFC y RFID, códigos de barra y códigos QR, sensores ZigBee, y otros dentro de Internet de las Cosas, o datos procedentes de redes sociales (Facebook, Twitter, LinkedIn o Instagram), todos ellos unidos al tránsito de datos en todo tipo de negocios como banca, grandes almacenes, medios de comunicación, industrias, etcétera.

10.2.1 TIPOS DE ANALÍTICA DE DATOS

La Analítica de Datos consta de varias etapas, que van desde la captura de datos hasta la producción de resultados, para que las empresas obtengan un valor empresarial óptimo. Existen tres clases o tipos de Analítica de Datos: descriptiva, predictiva y prescriptiva (ya introducidos en el capítulo 2).

Analítica Descriptiva (*Descriptive Analytics*). Consiste en preparar y analizar datos históricos para identificar patrones y tendencias. Logra un profundo conocimiento a partir de dichos datos con informes, cuadros de mando, agrupaciones, etcétera. Utiliza los datos para explicar lo que sucedió en el pasado. Por esta razón, el análisis descriptivo responde a la pregunta ¿qué sucedió? La Analítica Descriptiva usa técnicas tales como modelos de regresión, modelados y visualización de datos. Algunas escuelas, como la citada consultora Gartner, consideran una cuarta categoría denominada **Analítica de Diagnóstico**, que detalla análisis y visualizaciones de diagnósticos y responde a la pregunta ¿por qué sucedió?

La Analítica Descriptiva está muy ligada a la Inteligencia de Negocios y, como ya se dijo, utiliza técnicas tales como modelos de regresión, modelado y visualización de datos. Por estas razones nos permite conocer alguna de las siguientes actividades:

- Detectar cuáles son los productos que se venden más y en qué zonas geográficas se despliega el negocio.
- Observar la evolución histórica de la demanda de un producto o servicio en determinados periodos de tiempo.
- Conocer las personas o empresas más influyentes en el sector de negocios de la empresa.
- Visualizar las noticias de prensa, radio, TV, redes sociales de mayor repercusión, así como su situación geográfica, segmentos de población, etcétera.

Analítica Predictiva (*Predictive Analytics*). Permite descubrir patrones ocultos en datos que el humano experto no puede apreciar. Es el resultado de aplicar matemáticas y estadística a los datos. Consiste en utilizar los datos para determinar lo que sucede, o puede suceder, en el futuro. La Analítica Predictiva permite determinar la probabilidad asociada a eventos futuros a partir del análisis de la información disponible (presente y pasado); además, hace posible descubrir relaciones entre los datos que, normalmente, no se detectan con un análisis menos complejo. Los modelos predictivos emplean técnicas de aprendizaje automático, Minería de Datos, etcétera. Responde a la pregunta ¿qué sucederá?

La Analítica Predictiva trata de extraer conocimiento de los datos en forma de patrones, tendencias o modelos que nos proporcionen una cierta tendencia sobre potenciales situaciones futuras o el resultado de diferentes alternativas sobre las que tomar una decisión. El objetivo clave es pronosticar lo que sucederá en el futuro a partir del análisis de los datos históricos; es decir, la Analítica Predictiva se construye sobre la Analítica Descriptiva y usa modelos estadísticos avanzados y de aprendizaje automático, así como, cada día con mayor profusión, aprendizaje profundo.

Algunas de las tareas que permite la Analítica Predictiva son:

- Anticipar demandas de clientes en diferentes puntos de venta.
- Detectar transacciones bancarias fraudulentas.
- Descubrir grupos de interés de clientes que comparten características comunes: demográficas, geográficas, preferencias de productos, etcétera.

Analítica Prescriptiva (*Prescriptive Analytics*). Utiliza los datos para prescribir aquellas acciones que incrementan nuestras posibilidades de obtener los mejores resultados. Determina nuevas formas de operar que permiten alcanzar los objetivos del negocio. Responde a las preguntas ¿por qué va a pasar?, ¿qué hacer para que pase? Sugiere opciones de decisión acerca de la manera de aprovechar una oportunidad de futuro, o mitigar un riesgo futuro, y muestra las consecuencias de cada decisión; nos permitirá saber qué debemos hacer para optimizar nuestro negocio.

10.3 ANALÍTICA DE NEGOCIOS (BUSINESS ANALYTICS/ANALYTICS)

Analítica de Negocios o Analítica Empresarial (*Business Analytics* o, simplemente, *Analytics*) es el descubrimiento y la comunicación de patrones significativos de datos. La Analítica de Negocios posibilita lograr una ventaja competitiva para las organizaciones, sobre todo en aquellas que son más ágiles e innovadoras. Es un subconjunto de la Inteligencia de Negocios y así se suele considerar; sin embargo, la Analítica está teniendo mucha fuerza. Así, el reputado informe anual de la consultora Gartner *Cuadrante mágico de Gartner de plataformas de Inteligencia de Negocios* pasó a denominarse *Cuadrante mágico de Gartner de plataformas de Inteligencia de Negocios y Analítica* desde el año 2013.

La Analítica ha emergido, según Gartner, como un término que engloba a diferentes iniciativas y aplicaciones relacionadas de Inteligencia de Negocios. Sin embargo, muchas otras escuelas consideran la Analítica como el proceso de analizar información de un dominio determinado, como puede ser Analítica Web o Analítica Social. Otras escuelas consideran el término como la aplicación de las capacidades de Inteligencia de Negocios (IN) a un área de contenido específico

(por ejemplo, ventas, servicios, cadena de suministro, etcétera.). *Analytics* ha ganado un creciente interés de los profesionales de negocio de TI que tratan de explotar los grandes volúmenes de datos disponibles generados interna o externamente.

Los métodos utilizados por la Analítica de Negocios se apoyan en las diferentes categorías de Analítica de Datos examinadas en el apartado anterior.

DATOS EN TIEMPO REAL

Hoy en día, existe una enorme cantidad de datos que procede de las tecnologías más comunes existentes y que produce grandes volúmenes de datos, tales como datos espaciales, de sistemas de información geográfica (SIG/GIS), de geolocalización, generados por máquinas (M2M o Internet de las Cosas) como chips móviles (NFC, RFID), sensores, robots, códigos QR, antenas, sistemas de medios de comunicación, datos de acontecimientos o eventos, etcétera.

Por otra parte, el flujo continuo y creciente de datos (*streaming Big Data*) —sobre todo en datos de texto, video, fotografía y audio— producen su procesamiento en tiempo real, es decir, a medida que se producen, se capturan y almacenan; este flujo continuo es crucial para encontrar datos significativos y de calidad.

La creciente marea de grandes volúmenes de datos procedentes de dispositivos de geolocalización, realidad aumentada y telefonía móvil (especialmente de teléfonos inteligentes) requiere, en numerosos negocios y actividades de la vida diaria, la toma de decisiones rápida y en tiempo real o con el menor retardo.

10.4 UNA VISIÓN GLOBAL DE ANALÍTICA DE BIG DATA

El análisis de *Big Data* es el proceso de examinar, a una gran velocidad, grandes volúmenes de datos de una amplia variedad de tipos y de gran valor (el modelo de las 4V) para descubrir patrones ocultos, correlaciones desconocidas y otra información útil, de modo que los resultados del análisis puedan proporcionar ventajas competitivas a las organizaciones en relación con la competencia y producir beneficios para el negocio, tales como un *marketing* más efectivo y eficaz y mayores ingresos.

Los grandes volúmenes de datos procederán de bases de datos relacionales tradicionales, así como de otras fuentes de datos (capítulo 5) tales como son registros del servidor web, de seguimiento de clics en Internet (*click stream*), informes de actividades sociales, medios de comunicación, datos de teléfonos móviles inteligentes, registros detallados de llamadas en las centralitas de la empresa o en sus *call centers*, la información captada por sensores,

etcétera. Recordemos que los grandes datos no sólo se asocian a los datos no estructurados y semiestructurados, sino también a los datos estructurados procedentes de transacciones comerciales o almacenadas en bases de datos relacionales. Algunas personas asocian exclusivamente grandes datos a análisis de datos no estructurados. Sin embargo, lo técnicamente correcto es asociar la Analítica de *Big Data* a la integración de datos estructurados y no estructurados/semiestructurados.

El análisis de grandes datos se puede hacer con herramientas de *software* tradicionales dentro de las técnicas de analítica avanzadas, como la Minería de Datos o el Análisis Predictivo. Sin embargo, las fuentes de datos no estructurados utilizados en el análisis de grandes cantidades de datos pueden no encajar en los almacenes de datos tradicionales (las bases de datos o los almacenes de datos empresariales, EDW) y, además, estos almacenes pueden no ser capaces de manejar las demandas de procesamiento de grandes datos. En consecuencia, han surgido nuevas tecnologías que incluyen bases de datos NoSQL y "en memoria", Hadoop y MapReduce, SAP Hana, Oracle *in-memory*, etcétera.

Los grandes retos que enfrentan las organizaciones es la necesidad de integrar las nuevas infraestructuras de *Big Data* con las infraestructuras de datos existentes y, tal vez más complicado, la contratación de profesionales con experiencia en analítica de *Big Data,* tales como analistas y científicos de datos. Ya se ha comentado también la dificultad de que los *Data Warehouses* convencionales puedan escalar hasta terabytes de datos o soportar Analítica Avanzada.

La tecnología, no obstante, sigue avanzando y comienzan a verse actualizaciones en torno a plataformas NoSQL que tienen información estructurada y no estructurada. Así, en la galería de soluciones disponibles, actualmente se encuentra la aplicación Greenplum, de EMC, Hadoop y MapReduce, la nueva plataforma Vertica, de HP, la oferta por separado de Smart Analytic System y Netezza, de IBM, basadas en DB2, y Microsoft Parallel Data Warehouse. Existen otros jugadores más pequeños, de nicho, como Infobright y Kognitio. Oracle ha entrado en el mercado y está ofreciendo magníficas soluciones y Teradata sigue siendo una de las soluciones líderes.

En el entorno de *Big Data*, las organizaciones se encuentran ante el desafío de incorporar información en crudo, sin procesar, que se actualiza en tiempo real y que presenta una enorme complejidad. Pero la cuestión clave no tiene que ver con la capacidad para recolección y almacenamiento de los grandes datos. No basta con capturar y almacenar una gran cantidad de datos, es necesario saber organizarlos, refinarlos y convertirlos en información relevante que permita ganar posiciones en el mercado. La información en crudo sólo tiene valor potencial, es su análisis y sistematización lo que permite incrementar la capacidad de innovar de las organizaciones. De manera que el tratamiento de los grandes volúmenes de datos requiere de las etapas de tratamiento de Big Data, ya analizadas:

Adquisición o ingesta de datos. Los datos procederán de fuentes de datos tradicionales (almacenes de datos de empresa EDW, bases de datos relacionales y archivos con datos transaccionales) y de una gran cantidad de fuentes de datos no estructurados que se podrán almacenar en bases de datos NoSQL y "*en memoria*" (*in memory*).

Procesamiento de la información. Preparar y tratar la información para obtener de ella los mejores resultados posibles y sobre los cuales se puedan aplicar, lo más eficientemente posible, las técnicas de Analítica Avanzada.

Análisis. Analizar toda la información con acceso a todos los datos con herramientas estadísticas avanzadas, como puede ser la Minería Social y de Opinión, o aplicar técnicas desarrolladas con el lenguaje de programación **R**, específico para el diseño de estadística avanzada. Desde un punto de vista global, sería conveniente que el proveedor de Analítica pueda ofrecer herramientas de consultas (*quering*) e informes o reportes (*reporting*), Minería de Datos, visualización de datos, modelado predictivo y optimización.

Decisión, resultados y visualización. Tomar decisiones en tiempo real, o lo más rápido posible, de modo que pueda afectar positivamente en los negocios de la empresa. Esta etapa se encuentra indisolublemente unida a la etapa de análisis; de hecho, muchos vendedores ofrecen estas herramientas integradas con las de decisión (este es el caso de Oracle). La decisión se ha de realizar en tiempo real sobre la base de los resultados obtenidos en el análisis, de modo que los datos en crudo se conviertan en **conocimiento accionable** para integrarlo en los tableros de control (*dashboards*), cuadros de mando integral (*balanced scorecards*) y herramientas de visualización, y así predecir el comportamiento de un producto o servicio a los consumidores.

10.5 CATEGORÍAS PRÁCTICAS DE ANALÍTICA

Se pueden considerar diferentes categorías de Analítica de Datos en función de los dispositivos y plataformas que se han de utilizar en el proceso completo de Analítica:

- **Analítica de Datos tradicional** en organizaciones y empresas que analizan datos tradicionales: transaccionales y operacionales.
- **Analítica Web** o analítica del tráfico de datos en un sitio Web. Tecnologías y herramientas de análisis Web, utilizadas en servidores, PC y *laptops*, que han conformado los componentes de los sistemas de información tradicional, base y fundamento de la disciplina de Analítica Web.
- **Analítica Social (*Social Analytics*).** Análisis de datos de los medios sociales (*blogs*, wikis, redes sociales, RSS). A medida que los medios sociales comienzan a implantarse en organizaciones y empresas como servicios, al

igual que cualquier otro *software* o servicio, aparece la necesidad de analizar y gestionar los datos procedentes de estas aplicaciones.

Surgen nuevas profesiones asociadas a los medios sociales, tales como especialistas en SMO (*Social Media Optimization*) al estilo del SEO, especialistas en Gestión de Comunidades, conocido popularmente como *Community Manager* y el SMM *(Social Media Manager),* gestor o administrador de medios sociales (similar al *Community Manager*, pero con tareas de administración o dirección).

- **Analítica móvil** en dispositivos móviles, con el objeto de analizar los datos que envían, reciben o transitan desde dichos dispositivos (***Mobyle Analytics***). El despliegue imparable, primero de teléfonos inteligentes (*smartphones*) y posteriormente de las tabletas (*tablets*) —especialmente desde el lanzamiento de la tableta iPad de Apple en 2010—, ha creado la necesidad de analizar los datos producidos por estos dispositivos móviles, dado que se están convirtiendo en los puntos de acceso a Internet más empleados en organizaciones y empresas, y también por particulares.
- **Analítica de *Big Data*** o analítica de los grandes volúmenes de datos. La gran tendencia que se está produciendo en estos dos y tres últimos años es la explosión de los *Big Data.* Esta creciente tendencia ha traído de modo irreversible la necesidad de realizar estudios de Analítica Web sobre el inmenso tráfico de grandes datos que se están produciendo día a día.

10.6 ANALÍTICA DE BIG DATA

La Analítica de *Big Data* (*Big Data Analytics*) es el uso de técnicas analíticas aplicadas a conjuntos de grandes volúmenes de datos. Por consiguiente, Analítica de *Big Data* es realmente dos cosas: **Analítica** y ***Big Data***. La primera ayuda a descubrir aquellos datos que han cambiado en el negocio para saber cómo reaccionar; los grandes datos deben ayudar a convertir en oportunidades los retos producidos por el crecimiento espectacular de los *Big Data*. La Analítica es el mejor medio para descubrir nuevos segmentos de clientes, identificar a los mejores proveedores, asociar productos por afinidad, entender las ventas por la estacionalidad, etcétera.

Analítica de *Big Data* es el proceso de examinar grandes cantidades de datos, de una variedad de tipos, para descubrir patrones ocultos, correlaciones desconocidas y otra información útil. Dicha información puede proporcionar ventajas competitivas sobre organizaciones rivales y brindar beneficios en los negocios, tales como un *marketing* más eficiente y un aumento de los ingresos.

La Analítica es una manera de descubrir qué ha cambiado y cómo reaccionar ante ese cambio. Está compuesta por una colección de técnicas relacionadas y tipos de herramientas que normalmente incluyen Analítica Predictiva, Minería de

Datos, análisis estadísticos y programación compleja de SQL. Se puede extender la lista para cubrir visualización de datos, Inteligencia Artificial, procesamiento de lenguaje natural y capacidad de bases de datos que incluyan soporte de Analítica (como MapReduce, analítica *in database*, bases de datos *in memory*, bases de datos o almacenes de datos columnares).

Este tipo de analítica es conocida en las organizaciones como Analítica Avanzada. Sin embargo, está naciendo un término mejor para reflejar este tipo de analítica y es la denominada **Analítica de Descubrimiento o Analítica Exploratoria**, que hace uso del descubrimiento de datos y su análisis para la toma de decisiones. En otras palabras, con Analítica de *Big Data* el usuario es, normalmente, un analista de negocios que está intentando descubrir nuevos hechos que nadie en la empresa conocía antes. Para hacer eso, el analista necesita grandes volúmenes de datos con gran profusión de detalle que, por lo general, la empresa no ha aprovechado todavía para analizar.

El análisis de *Big Data* se realiza con herramientas de *software* utilizadas, normalmente, como parte de la disciplina de la Analítica Avanzada. Así, las herramientas usuales son:

- Consultas avanzadas en SQL.
- Consultas e informes (*quering y reporting*).
- Análisis estadístico avanzado.
- Visualización de datos.
- Minería de Datos, Minería de Textos, Minería Web y Minería Social.
- Análisis y modelado predictivo.
- Optimización.
- Sensibilización.
- Cuadros de control y de mando (*dashboard* y *scorecards*).

Las tecnologías asociadas con Big Data incluyen, fundamentalmente, *Data Warehouses, Data Marts,* bases de datos NoSQL y "en memoria", marcos de trabajo (*frameworks*) Hadoop, Spark y MapReduce.

Big Data, como reconocen todos los estudios serios realizados sobre el tema por las grandes consultoras y fabricantes de herramientas, es una oportunidad más que un problema. La analítica de grandes datos es una necesidad ineludible, pero su adopción está supeditada y controlada por una variada gama de tecnologías, la gestión o administración de los negocios y la economía de la organización. Las infraestructuras de *Big Data* están soportadas por almacenamiento de datos, técnicas "en memoria", aplicaciones NoSQL y el soporte de grandes anchos de banda.

ANALÍTICA DE *BIG DATA*: UNA NECESIDAD

Analítica de *Big Data* es la aplicación de técnicas de Analítica Avanzada para operar sobre grandes conjuntos de *Big Data*. En realidad, lo que se hace es unir dos áreas con entidad propia: *Big Data*, como cantidades masivas de información detallada, y *Analítica Avanzada*, que en realidad es una colección de diferentes tipos de herramientas, incluyendo aquellas que están basadas en Analítica Predictiva, Minería de Datos, estadística, Inteligencia Artificial, lenguajes de procesamiento natural, etcétera. Se unen los dos conceptos y se obtiene la analítica de *Big Data*.

Algunas organizaciones comienzan a gestionar los *Big Data* en sus almacenes de datos empresariales (*Data Warehouses*) y **EDW** (*Enterprise Data Warehouse*), aunque otras han diseñado sus propios *Data Warehouse* para las nuevas necesidades, y algunas más recurren a enfoques híbridos, como es el caso de los ya mencionados Lagos de Datos (*Data Lakes*).

La Analítica (*Analytics*) nos ayuda a descubrir lo que ha cambiado y cómo debemos reaccionar y la Analítica Avanzada es el mejor medio para descubrir nuevos segmentos de clientes, identificar los mejores proveedores, asociar productos por afinidad, comprender la estacionalidad de las ventas, etcétera, (TDWI, 2011)[2]. En esencia, la Analítica Avanzada son las implementaciones de formas específicas de Analítica que constan de una colección de técnicas relacionadas y tipos de herramientas; normalmente, incluyen Analítica Predictiva, Minería de Datos, análisis estadístico y SQL complejo, aunque la lista cubre visualización de datos, Inteligencia Artificial, lenguajes de procesamiento natural y capacidades de bases de datos analíticas como MapReduce, analítica *in database,* bases de datos *in memory* y almacenes de datos columnares.

TDWI también utiliza otro término, ya citado anteriormente, en lugar de Analítica Avanzada: **Analítica de Descubrimiento**, que procede de lo que están intentando descubrir los usuarios como resultado del análisis exploratorio (conocido también como **Analítica Exploratoria**). Recurriendo de nuevo a TDWI, esta organización define Analítica de *Big Data* como el conjunto de técnicas de Analítica Avanzada que operan sobre *Big Data*. La Analítica Avanzada tiene ya mucha presencia en las organizaciones y la Analítica de *Big Data* comienza a tener también presencia significativa.

El multicitado número de la revista *Harvard Business Review* de octubre del 2012 publicó un tercer artículo de Barton y Court (2012: 79-83), dedicado precisamente a la Analítica Avanzada, con un subtítulo sugerente: *Una guía práctica para la capitalización de Big Data*, que comienza con la frase: "*Big Data* y *Analytics* se han disparado a la cima de la agenda corporativa". Los ejecutivos ven con admiración, según los autores, a empresas como Amazon, Google y otras que han eclipsado a las competidoras con poderosos nuevos modelos de negocios, derivados de una gran capacidad para explotar los datos. La tendencia

está generando mucho ruido, pero sí es cierto que los líderes *senior* están comenzando a prestar atención a estas nuevas estrategias del mercado.

Los autores del artículo, reconocidos analistas de la consultora McKinsey (líder en implantación de soluciones en *Big Data*) con gran experiencia en asesoramiento a empresas, recomiendan que, ante la explosión de los grandes volúmenes de datos y de la Analítica, las empresas y las industrias requieren de tres capacidades que se soportan mutuamente.

En primer lugar, las empresas deben ser capaces de identificar, combinar y gestionar múltiples fuentes de datos. En segundo lugar, ellas necesitan la capacidad para construir modelos de Analítica Avanzada para la predicción y optimización de resultados. Por último, y la capacidad más crítica, la gestión debe poner toda su potencia para transformar la organización, de modo que los datos y los modelos produzcan realmente mejores decisiones.

Otras características importantes que sustentan estas actividades son: una clara estrategia de cómo utilizar los datos, la analítica para competir, el despliegue de las capacidades, y la arquitectura tecnológica adecuada.

La era de los *Big Data* está evolucionando rápidamente, y la experiencia sugiere que las compañías deben actuar ya y deben concentrar sus esfuerzos en las fuentes de datos, construcción de modelos y transformación de la cultura organizacional. Esta disposición corporativa es esencial, dado que la información, junto con la tecnología para su gestión y análisis, continuará creciendo y produciendo un flujo constante de oportunidades. La destreza en el buen uso de *Big Data* se convertirá pronto en un activo competitivo decisivo.

10.7 CARACTERÍSTICAS DE UNA PLATAFORMA DE INTEGRACIÓN DE ANALÍTICA DE BIG DATA

Una plataforma integrada de Analítica de *Big Data* debe ser innovadora y se ha de integrar en la infraestructura de TI de la organización. Además, debe ser de última generación. Peter J. Jamack, consultor de IBM, ha publicado en la plataforma oficial de la empresa un excelente análisis de cómo integrar la infraestructura de analítica de *Big Data* y la infraestructura de Inteligencia de Negocios de la empresa[3]. Jamack, plantea que la inteligencia de negocios, la gestión de datos maestros, *Big Data* y la analítica deben integrarse en una plataforma y agruparse en una solución visualmente innovadora.

Una inteligencia de negocios y analítica de *Big Data ha de ser innovadora. Debe ser de última generación.* Debe utilizar tecnologías NoSQL y "en memoria" *(in memory)* o configurar un sistema para utilizar herramientas como Hadoop y Apache Cassandra como área de transferencia, recinto de seguridad, sistema de almacenamiento y ser un sistema nuevo y mejorado de ETL (extracción,

transformación y carga). Debe integrar datos estructurados, no estructurados y semiestructurados. Si las operaciones de ETL no se realizan correctamente, de repente recibirá datos incorrectos y poco confiables. Los datos poco confiables se convierten en un sistema poco confiable y no utilizado.

Una solución es desarrollar un sistema completo de código abierto utilizando el marco de trabajo Hadoop (HDFS y MapReduce) o Spark, y herramientas tales como Zookeeper, Avro, Sqoop, Hive, HBase, Ambari o Storm. Otra solución sería desarrollar un sistema usando herramientas propietarias e inyectores a Hadoop, como puede ser el caso de IBM con las herramientas InfoSphere, BigInsights e IBM Netezza. Otras compañías, tal vez, quieran separar datos estructurados y sin estructura y desarrollar una capa de interfaz gráfica de usuario (GUI) para usuarios, usuarios avanzados y aplicaciones. A veces, se puede emplear herramientas como Sqoop, gran herramienta para ingerir datos de sistemas de gestión de base de datos relacionales. Añadir otras herramientas de código abierto como Flume o Scribe puede ayudar con los sistemas de registros.

El **almacenamiento de datos** es un factor enorme y puede requerir que use diversas tecnologías. En el sistema de Hadoop se encuentra HBase. Pero algunas compañías usan Cassandra, Neo4j, Netezza, HDFS y otras tecnologías, dependiendo de lo que se necesite. HDFS es un sistema de almacenamiento de archivos. HBase es un almacén de datos por columnas similar a Cassandra. Muchas compañías utilizan Cassandra para analíticas más cercanas al tiempo real.

El **sistema de gestión de bases de datos** puede considerar a HBase o Cassandra cuando desee emplear un sistema de código abierto para Analítica de *Big Data.* En lo que se refiere a plataformas de almacenes de datos, Netezza es una de las principales tecnologías en la industria de la Analítica y la IN. La mejor opción para la integración de *Big Data* es usar una plataforma integrada, compuesta de Hadoop y Cassandra para datos sin estructura o semiestructurados y Netezza para datos estructurados.

La **interfaz gráfica de usuario** (GUI) se puede realizar con herramientas como SPSS Statistics de IBM, o el lenguaje R de estadística o herramientas de Minería de Datos, modelado predictivo, aprendizaje de máquina (como Apache Mahout) y desarrollo de algoritmos y modelos complejos, con lenguaje de consulta estructurado, como Apache Hive.

10.8 ANALÍTICA DIGITAL

La Analítica de Datos descrita en los apartados anteriores tiene un enfoque muy especial cuando se habla de los datos de Internet y de la Web, y todas las categorías de Analítica definidas serán de utilidad en este campo específico. Este enfoque de la Analítica se conoce como Analítica Digital.

Descargado en: eybooks.com

Analítica digital[4] se centra en analizar la actividad de un sitio web a partir de los datos extraídos de la navegación de los usuarios, así como estudiar también la presencia de una marca, persona u organización en Internet, con un objetivo señalado previamente: ventas, reputación, influencia o visibilidad.

La *Digital Analytics Association*[5] define Analítica Digital como "una parte de la ciencia del análisis que usa datos para entender los patrones históricos con miras a mejorar el desempeño y predecir el futuro. El análisis de datos digitales se refiere a la información recogida en canales interactivos (en línea, móvil, social, etcétera.)".

La Analítica Digital permite hacer el seguimiento y el análisis de los clientes y usuarios que entran en contacto con la parte digital del negocio: sitio web, redes sociales, correo electrónico, aplicaciones móviles, etcétera, con el objetivo de obtener conocimiento de los datos digitales de la Web y de Internet para una toma de decisiones adecuada y óptima.

Los tipos de análisis de datos más empleados en la actualidad, para realizar el seguimiento de los diferentes sitios y páginas web, son:

- Análisis de texto.
- Análisis de sentimientos.
- Análisis de localización.
- Análisis de movimiento.
- Análisis de reconocimiento facial.
- Análisis de voz.
- Análisis de imágenes.

Todos estos tipos de análisis se agrupan en tres grandes categorías de Analítica Digital: **Analítica Web**, **Analítica Social** y **Analítica Móvil**. El estudio de estos tres tipos de Analítica Digital constituye el soporte de gran parte de los datos manejados en los sistemas de Inteligencia de Negocios y de Analítica de Negocios, así como herramientas indispensables para afrontar la digitalización y la transformación digital de organizaciones y empresas.

10.9 ANALÍTICA WEB

Analítica Web es una categoría de Analítica Digital que, a su vez, es una rama de la Analítica de Datos y Analítica Empresarial, que se centra en el análisis de los datos que fluyen a través de sitios y páginas web. En realidad, el análisis de datos en la Web es más bien análisis del tráfico Web. La Web ha ofrecido datos, más datos, llegando (como sabe el lector) a los grandes volúmenes de datos (*Big Data*). Sin embargo, se trata de encontrar los datos significativos, y ésta es la

tarea difícil. Avinash Kaushik, uno de los padres de la Analítica Web (tal vez el más significativo), en el capítulo 1 de su libro *Analítica Web 2.0* habla de la paradoja de los datos. Señala que "para la Web, la paradoja de los datos es una lección de humildad: sí, hay una gran cantidad de datos, pero a la hora de tomar decisiones inteligentes existen obstáculos fundamentales". En realidad, en su análisis inicial, Kaushik plantea que, tal vez, los datos no son el problema real y que probablemente lo sea la gente; eso le lleva a considerar que los elementos nucleares del análisis del tráfico web o análisis web, sean las personas (los analistas web como profesionales y los directivos que han de tomar las decisiones) y las herramientas empleadas en el análisis web.

La referencia mundial de Analítica Web es el ya citado Avinash Kaushik y su obra *Web Analytics 2.0* (en español, *Analítica Web 2.0*) es el fundamento más importante de la disciplina. En ella, su definición de Analítica Web es: "el análisis de datos cuantitativos y cualitativos de su sitio Web y de la competencia, para impulsar una mejora continua de la experiencia *online* que tienen tanto los clientes habituales como los potenciales y que se traduce en unos resultados esperados (*online* y *offline*)" (Kaushik 2011: 24). En la definición se destacan sus dos palabras básicas: el analista y las herramientas que se utilizarán. El análisis web se soporta fundamentalmente en el *click stream* (flujo o secuencia de clics). Este flujo de clics permitirá conocer casi todo acerca de los usuarios o consumidores, así como disponer de datos suficientes para analizar lo que está sucediendo y las acciones que se han de realizar para mejorar.

La secuencia de clics permitirá recopilar, almacenar, procesar y analizar los datos a nivel de clic de su sitio web. Esta tarea se podrá obtener con herramientas de Analítica Web como Google Analytics, Yahoo! Analytics, Twitter Analytics, Facebook Analytics, Webtrends, etcétera, y se podrá obtener la información, ya sea en su propio sitio web o en el servidor web, dependiendo de la herramienta de *software* instalada.

Analítica Web es una disciplina nueva que se ha ido desgajando de las teorías de Inteligencia de Negocios, de los sistemas de información, especialmente desde el auge del *marketing* digital y de la gestión de relaciones con los clientes (CRM). Fue durante los primeros años de la primera década del siglo XXI, justo después de la explosión de las empresas "puntocom", cuando la Analítica Web comenzó a ver la luz en el reino de sistemas de información de empresas con gran presencia tecnológica.

En 2011 y 2012, tras la publicación en 2010 de la citada obra de impacto mundial *Web Analytics 2.0*, se comenzó a propagar la Analítica Web como una disciplina a nivel mundial. En esta segunda década del siglo XXI, siguen las noticias comerciales de Analítica Web, pero sobre todo se han asentado las herramientas en los sistemas de información de las compañías, al igual que sucede con el *software* antivirus o la actualización de la última versión del sistema operativo.

10.9.1 MÉTRICAS

El análisis del flujo secuencial de clics se apoya en dos componentes fundamentales: las métricas y los **KPI** (indicadores clave de rendimiento). En este apartado nos centraremos en el concepto de métrica.

Una **métrica** es una valoración cuantitativa de estadísticas que describen tanto los eventos como las tendencias de un determinado sitio web (Kaushik, 2011). Una métrica, en realidad, es una medida cuantitativa que permite conocer el estado de un sitio web[6], de una página web o un proceso que se realiza en un sitio web, para un atributo o parámetro determinado. Existen numerosas métricas que aportan abundante información. En el capítulo 11 se describen las métricas más empleadas, aquellas que Kaushik denomina "las ocho métricas cruciales", y alguna otra más de gran impacto en el análisis web de un sitio o página web, aunque es preciso constatar que puede existir algún matiz diferenciador en el contexto de la herramienta web utilizada y en la de una determinada métrica, que puede variar de una herramienta a otra, por lo que será conveniente, una vez decidida la herramienta de análisis, conocer fielmente el concepto de cada métrica en su herramienta. Esta confusión de términos de métricas es especialmente acertada en las métricas **visitante** y **visitante único**, aunque cada proveedor tratará de medir el proceso de personas reales que de verdad han visitado el sitio web y, normalmente, no tendrá problemas en la identificación numérica de estos conceptos.

10.9.2 INDICADORES CLAVE DE RENDIMIENTO (KPI)

De todas las métricas existentes es preciso seleccionar aquellas que sirven para planear los objetivos empresariales específicos, que llevan a obtener mayor productividad y conseguir el cumplimiento de los objetivos del sitio web o del medio social correspondiente. Las métricas que permiten medir el progreso del sitio web (portal o medio social) en relación con sus objetivos se denominan **indicadores clave de rendimiento (KPI)**.

Los **KPI** (*Key Performance Indicators*), **indicadores clave de rendimiento** (también se les conoce en algunas zonas de Latinoamérica como **indicadores clave de desempeño**), son métricas utilizadas para cuantificar objetivos, es decir, reflejan el rendimiento y la progresión para conseguir los objetivos. Los KPI deben ayudar a definir y medir el progreso hacia los objetivos de la empresa y tienen que mostrar si está consiguiendo sus propósitos de negocio. En consecuencia, **todos los KPI son métricas, pero no todas las métricas son KPI**. Dependiendo de los objetivos estratégicos de la organización o empresa, se diseñan los objetivos del sitio web y se deberán deducir, dependiendo del modelo de negocio, cuáles son las métricas idóneas que deberán convertirse en KPI; es decir, se deben seleccionar y determinar cuáles son las métricas que ayudan a conseguir los

objetivos del negocio. Los KPI son muy variados y dependerán de las métricas que pueden obtenerse del análisis del tráfico de la web.

No todas las métricas se revisan cada día; sin embargo, aquellas que dan una visión global de la empresa son candidatas a KPI y, en consecuencia, los KPI deberían ser la primera métrica por comprobar cuando se inicia una sesión diaria, y tendrían que supervisarse constantemente para determinar el estado de sus programas.

Diferencias entre métricas y KPI
Todos los KPI son métricas, pero no todas las métricas son KPI, o mejor dicho, no son lo suficientemente importantes para ser consideradas KPI. En la práctica, una métrica es una medida o estadística de un evento y un KPI es una métrica que permite cuantificar cuánto se está haciendo respecto de los objetivos del sitio web o medio social.

10.10 PROLIFERACIÓN DE DATOS SOCIALES

Las explosiones de los grandes volúmenes de datos proceden de numerosas fuentes de datos, pero, sin lugar a dudas, los medios sociales (*social media*) son responsables de grandes porcentajes en un sentido amplio. Si analizamos las estadísticas de acceso a Internet por dispositivos móviles (Internet móvil), y sobre todo las tendencias y cifras previstas para los próximos años, el aluvión de datos debido a los *social media* crecerá con cifras espectaculares hasta el punto que los petabytes y exabytes serán las cifras que se han de considerar.

En el caso de los medios sociales, los datos que se están acumulando proceden de multitud de fuentes (YouTube, Twitter, Facebook, Google+, LinkedIn, Amazon, eBay, Instagram, Foursquare, Tuenti, Pinterest, WhatsApp, Spotify y Snapchat) y cientos de miles de *blogs*, *wikis*, chats, foros, etcétera. Esta situación lleva a una sobrecarga de datos y a la necesidad de descubrir los que son realmente significativos para organizaciones y empresas; es decir, se necesita conocer y ponderar la relación señal/ruido. Por esta razón, el análisis de datos sociales es hoy en día una necesidad vital para las organizaciones y empresas, pero también para el usuario individual, llámese empleado, estudiante, ingeniero, profesor, directivo o científico.

La Analítica Social permite integrar y analizar los datos no estructurados que se encuentran en el correo electrónico, la mensajería instantánea, los portales web, los *blogs* y otros medios sociales, usando las herramientas de obtención de datos existentes, los informes de Inteligencia de Negocios o Empresariales y otras herramientas como los cuadros de mando integral. El análisis de la información generada en los medios sociales y dispositivos móviles permite obtener información en tiempo real sobre las tendencias de consumo. En el apartado

10.11 se define analítica social y se describen los componentes necesarios para utilizar esta disciplina como herramienta de gestión eficiente.

A medida que aumentan los *Big Data*, y en particular los procedentes de los medios sociales, el análisis de datos se hace más completo y se requerirá conocer las fuentes de datos o los canales de comunicación por donde se envían o reciben esos grandes volúmenes.

La proliferación de los datos sociales supone un gran reto para las organizaciones y empresas, y en particular para los directores de *marketing* (**CMO**), de tecnologías de la información (**CIO**)[7], *community manager* (administrador de comunidades), analistas de la Web, especialistas SEO, etcétera; es decir, una pléyade de profesionales, además de los directores y estrategas de la compañía.

A principios de septiembre de 2012, IBM publicó un informe[8] sobre las principales preocupaciones de los directores de *marketing*. Lo presentaba como resultado de un estudio mundial sobre los principales desafíos y preocupaciones de los CMO. Gran parte de los ejecutivos consultados señaló la explosión de datos, las redes sociales, la proliferación de canales y dispositivos y los cambios demográficos de los consumidores como los cuatro factores que afectarán a su negocio en los próximos años. El estudio fue realizado entre más de 1.700 directores de *marketing* de 64 países y 17 industrias.

En relación con la explosión de datos, 71% de los CMO afirmaban, ya en 2012, que era uno de los temas que más le preocupaban porque deben ser capaces de obtener información de valor entre los miles de millones de datos, estructurados y no estructurados, existentes en la actualidad. El volumen de información digital, como ya conoce el lector, es abrumador: 8 zettabytes se esperaban para 2017 (recordemos que 1 zettabyte de información equivale a 1.000 millones de discos duros de 1 terabyte o 75 mil millones de iPads de 16 gigabytes).

El estudio revelaba que, por aquel entonces, el 90% de la información que se creaba en tiempo real representaba datos no estructurados, y un porcentaje muy alto procedía de redes sociales y otros medios como *blogs*, *wikis*, chats o mensajes de texto y video. Evidentemente, este inmenso arsenal de conocimiento, si es aprovechado por los CMO, conseguirá un valor añadido y un alto posicionamiento en relación con la competencia. Otro factor por considerar es el incremento y proliferación de canales y dispositivos vinculados al, ya tantas veces repetido, uso creciente de tabletas y teléfonos móviles.

El estudio concluye con una recomendación al CMO: hay que ser capaz de pensar analíticamente y aprovechar la información disponible para averiguar los deseos del cliente antes que la competencia.

10.11 ANALÍTICA SOCIAL

A medida que las empresas aumentan su presencia en la Web y, especialmente, en las redes sociales, es imprescindible conocer las posibilidades que brinda la Analítica Social. La **Analítica Social** o **Analítica de Medios Sociales** (***Social Analytics*** o ***Social Media Analytics***) está comenzando a ser una disciplina muy necesaria en organizaciones y empresas, y es una de las áreas más impactantes dentro de la Analítica de Datos. Para referirse al análisis de datos en medios sociales, también se la conoce simplemente como **Análisis Social** (***Social Analysis***). Dada la fuerza que está adquiriendo, comienza también a considerarse una disciplina autónoma dentro de la Inteligencia de Negocios.

¿Qué es *Analítica Social*? Analítica, de acuerdo con el diccionario de la Real Academia Española (www.rae.es) es aquello "perteneciente o relativo al análisis" o "que procede descomponiendo, o que pasa del todo a las partes". En esta definición no existe ninguna acepción de la categoría **informática** o **computación** ni **social**, lo que sí sucede en la definición del término **análisis**: "estudio mediante técnicas informáticas, de los límites, características y posibles soluciones de un problema al que se aplica un tratamiento por ordenador (computadora)".

Análisis Social se puede considerar la disciplina que ayuda a las empresas a analizar, calcular y explicar el rendimiento de las iniciativas de *social media* en el contexto de objetivos empresariales específicos (Lovett, 2012: 164). **Analítica Social** se podría considerar al proceso de medir, analizar e interpretar los datos sociales que se presentan a través de los diferentes canales, medios de comunicación y dispositivos. También ha nacido un nuevo término asociado y acotado a las redes sociales como parte muy importante de los medios sociales: el término **Análisis de Redes Sociales** (**SNA, *Social Network Analysis***). El término ARS o SNA es un término ligado a las ciencias sociales y a la teoría general de redes en el campo de las comunicaciones.

> **Analítica Social** es una disciplina que ayuda a las organizaciones y empresas a analizar, medir y explicar el rendimiento de las iniciativas y proyectos sociales (de los *social media*) dentro del contexto de sus metas y objetivos.

El Análisis Social se basa en la teoría de redes sociales, en técnicas estadísticas y en la buena gestión de los medios sociales de la empresa. El Análisis de Datos Sociales debe proveer la capacidad de vincular la información a otras métricas de la gestión empresarial y a indicadores clave de rendimiento (KPI).

El Análisis Social permite analizar métricas cuantitativas para calcular éxitos, fracasos y situaciones críticas de los negocios; asimismo, le permitirá examinar interacciones con los clientes y el modo de recepción de los mensajes enviados por los departamentos de *marketing* a sus clientes, ayudándoles a entender cómo los usuarios perciben su marca y responden al lanzamiento de productos

corporativos, servicios y las diferentes campañas de *marketing*. El Análisis Social proporciona los datos necesarios para una acertada toma de decisiones, permitiendo el uso de los datos para efectuar recomendaciones a la empresa sobre cómo están funcionando los diferentes modelos de negocio y cómo se pueden mejorar; en la práctica, recopila, organiza y actualiza la información poniéndola disponible para las personas o grupos de interés (*stakeholders*) de la organización, de modo que ayude en la toma de decisiones.

La Analítica Social debe proporcionar una apertura al aumento de consumidores, empleados o negocios, así como a la capacidad de visualizar mejor patrones, tendencias y oportunidades. Las tecnologías y herramientas sociales deben facilitar el Análisis Social, que se ayuda, a su vez, de las personas y de los procesos de negocios.

La mayoría de las empresas utilizan las métricas y los indicadores clave de rendimiento o desempeño (KPI) para cuantificar, medir e informar sobre la actividad de los medios sociales.

Las redes sociales obligan a la creación de nuevas métricas distintas de las tradicionales (métricas sociales), precisamente por la importancia que han ido adquiriendo, ya que se han convertido en un elemento equivalente y complementario a los medios tradicionales de comunicación. Las métricas tradicionales siguen siendo muy importantes en las organizaciones y empresas y entregan datos relevantes para la toma de decisiones, pero es vital que se realicen estrategias de convergencia de medios para aprovechar las sinergias entre ambas.

10.12 ANÁLISIS DE SENTIMIENTOS

Actualmente, **Análisis de Sentimiento** o **Sentimientos** (***Sentimental Analysis***), también conocido en algunos ambientes como **Minería de Opinión**, se refiere al análisis automático del sentimiento que trata de traducir, a indicadores más o menos medibles, las emociones humanas inmersas en los datos sociales, tanto en fuentes externas y autónomas (redes sociales, *blogs*, *microblogs*, foros, medios de comunicación, wikis, etcétera) como internas o propias de la empresa (interacciones almacenadas en el CRM, transcripciones de conversaciones registradas en el sistema de soporte de incidencias, encuestas realizadas a clientes y empleados). Desde la perspectiva de una organización o empresa, el Análisis de Sentimientos permite analizar de modo rápido y eficiente qué se dice sobre una marca o producto, seguir las opiniones o conversaciones de determinados usuarios influentes, detectar tendencias en Internet, etcétera.

Se realiza mediante la monitorización y el análisis de datos sociales y de otro tipo, tanto procedentes de fuentes internas como externas a la empresa. El Análisis de Sentimientos tiene una aplicación muy importante en la

monitorización de las redes sociales, y de su análisis se puede obtener el grado de empatía de los internautas hacia una organización, así como permite a las empresas conocer en forma certera el grado de simpatía o rechazo que tienen ante la marca y/o producto. El Análisis de Sentimientos tiene diferentes indicadores, y los de mayor impacto son: *positivo/negativo/neutro* o, dicho de otro modo, *buenos/malos/neutro*.

En esencia, el Análisis de Sentimientos ha pasado a primer plano y existen numerosas herramientas de tendencias que obtienen los datos de millones de sitios y redes sociales para un mejor entendimiento de lo que se está comentando sobre las empresas, marcas u otros temas, lo que a su vez permite identificar oportunidades de inversión.

El Análisis de Sentimientos es un método más de intento de traducción de las emociones humanas en datos, pero con el uso de las herramientas modernas se puede conseguir que la espontaneidad e inmediatez de la opinión en medios sociales haga que dichos sentimientos sean más auténticos y preserven su contenido emocional. El Análisis de Sentimientos relativos a contenidos no estructurados se puede medir con tres características fundamentales*:* **polaridad** (¿la opinión o expresión emitida sobre un tema es positiva, negativa o neutra?), **intensidad** (¿cuál es el grado de emoción que se expresa?) y **subjetividad** (¿la fuente que emite la expresión o comentario es objetiva, es parcial o imparcial?).

A medida que el concepto de Análisis de Sentimientos se va asentando (especialmente en medios sociales y en aplicaciones de *software* empresarial como CRM social), el número y el uso de aplicaciones va creciendo en numerosos sectores de los negocios y de la sociedad en general. Algunas aplicaciones son:

- Medida de la satisfacción de los empleados y del clima laboral.
- Medida de la satisfacción del cliente.
- Prevenir abandono de clientes mediante la detección de situaciones de riesgo de pérdida de un cliente (opiniones negativas que se interpreten como posible abandono del cliente). Esta aplicación es muy utilizada en operadoras de telefonía para tratar de evitar las ofertas de los competidores.
- Comparación con la competencia mediante la evaluación de la opinión acerca de la misma (marca, empresa, productos).
- Detección de fortalezas y debilidades en diferentes áreas de nuestra empresa (opiniones positivas o negativas de impacto).
- Medida del impacto en la reputación corporativa.
- Predicción de la evolución de determinadas acciones (lanzamiento de productos, por ejemplo).

- Análisis de la opinión del electorado en el caso de votaciones políticas (presidenciales, regionales, municipales, etcétera). En el caso de las últimas votaciones presidenciales de diferentes países europeos y americanos, se publicaron numerosos Análisis de Sentimiento con diferentes herramientas y aplicaciones.

El Análisis de Sentimientos se encuadra dentro del procesamiento de lenguaje natural (PLN), de la Inteligencia Artificial y de la Minería de Textos (entre otras técnicas), ya que, fundamentalmente, busca extraer información subjetiva de un texto (un tuit, un *post*, un *blog*, etcétera). El analista de sentimientos se está convirtiendo en una profesión emergente dentro del área de analistas de datos y analistas web, que requiere de una formación multidisciplinar como lingüística, ingeniería de sistemas (informática), psicología, e incluso matemáticas o física.

10.13 ANALÍTICA MÓVIL

Analítica Web para dispositivos móviles o **Analítica Móvil** (***Mobile Analytics***) nace de la creciente necesidad de las empresas de conocer el retorno de la inversión de su canal móvil. A medida que aumenta la adopción de la telefonía móvil para el acceso a Internet, las empresas requieren estar presentes en el canal móvil a través de su sitio web (*web app*) o aplicaciones nativas. La **Analítica Móvil** se refiere al campo específico de la Analítica Web, en el canal móvil de la organización, y comprende el conjunto de prácticas y tecnologías para coleccionar y analizar los datos de la presencia en Internet desde móviles *(smartphones*) y tabletas, con el objetivo de tomar decisiones de negocio. Una de las primeras decisiones que ha de tomar la empresa, antes de proceder a la fase de Analítica de Datos, es seleccionar una aplicación web o nativa. El objetivo es conocer cuál es el modo en que el usuario interactúa con la aplicación o con la página web desde el móvil.

La Analítica Móvil debe poder responder a preguntas tales como:

- ¿Quién está utilizando las aplicaciones móviles de mi empresa o del profesional individual?
- ¿Cuál es la mejor aplicación: aplicación web o aplicación nativa?
- ¿Qué productos y servicios demandan los usuarios de teléfonos móviles o tabletas?
- ¿Cómo funciona mi sitio web desde mi dispositivo iPhone, Android o Windows Phone?

El servicio de Analítica Web Móvil[9] debe incluir al menos:

- Recolección de datos.
- Análisis (exploración y recomendaciones de negocio).

- Entrega y gestión de informes (*reporting*).

En la recolección de datos existen dos tipos de dimensiones: tipos de dispositivos que se han de medir y plataformas sobre las que se medirán. En la etapa de análisis será preciso contemplar cuáles son los KPI para medir los objetivos de atención al cliente, ventas, contenidos de las páginas, etcétera. Y en la etapa de *reporting,* entrega y gestión de informes, deberá estudiarse la integración de los datos con todos los canales de la empresa.

INFORMACIÓN DE HERRAMIENTAS DE ANALÍTICA MÓVIL

Las herramientas de Analítica Web Móvil, al igual que las herramientas de escritorio, deberán proporcionar información que pueda ser de utilidad para la empresa o el profesional. Algunas de las informaciones de interés que ha de proporcionar la herramienta de Analítica Móvil son:

Relativas a la actividad de los usuarios

- Usuarios activos.
- Duración de la sesión.
- Participación del usuario (duración media de sesión, *screens*/sesión, etcétera).
- Visitantes nuevos y recurrentes.
- Fidelización de los usuarios.
- Datos demográficos (idioma, país/territorio, ciudad).
- Flujo de interacción.

Relativas al dispositivo y a la aplicación

- Nombre y versión de la *app.*
- Número de instalaciones, bloqueos y excepciones.
- Ingresos de la *app*.
- Proveedor de servicios.
- Marca del dispositivo móvil.
- Resolución de pantalla.
- Versión del SO (sistema operativo).

Herramientas de analítica móvil

El sistema normal de medición de estadísticas en los sitios web de escritorio no se puede aplicar en su totalidad a las páginas de móviles, y es necesario tener en cuenta otras consideraciones. En primer lugar, el acceso a Internet en el caso de los dispositivos móviles no suele ser tan seguro como es el caso de las aplicaciones de escritorio. El acceso a las aplicaciones no está garantizado al 100%, bien por falta de cobertura 3G o 4G (y en el futuro cercano **5G**, previsible despliegue comercial en 2020), inexistencia de red Wi-Fi, bajas velocidades de acceso, ausencia de sincronización, etcétera.

Existen numerosas herramientas de Analítica Web Móvil, tanto gratuitas como de pago, así como con la modalidad *freemium.* Una selección de las herramientas más empleadas en Analítica Web Móvil (con los tres tipos de versiones: propietarias, códigos abiertos o gratuitos) es la siguiente:

- **Google Mobile Analytics**. Funcionalidades específicas para aplicaciones web (*apps*) y marcación (HTML 5).
- **Flurry Analytics**. Aplicación específica para Analítica Móvil.
- **Adobe Omniture/Adobe Site Catalyst**. Permite trabajar con librerías para *apps* y marcación de HTML 5.
- **Countly**. Es una aplicación de Analítica en tiempo real para la medición de aplicaciones móviles. Es una aplicación de código abierto (*open source*).
- **Localytics**. Plataforma para medición de *apps*.
- **Piwik Mobile**. Similar a Piwik de escritorio.
- **Apsalar**. En mayo de 2018 se ha fusionado con Singular para formar una plataforma analítica fusionada.
- **Woopra Mobile**. Igual a Woopra de escritorio.

CASO DE ESTUDIO: GOOGLE ANALYTICS

Google Analytics[10] es una herramienta gratuita de Analítica Web de Google, alojada en la nube, y reconocida como una de las herramientas más eficientes y de referencia en el área de Análisis Web que ayuda a la creación, actualización y mantenimiento de los sitios web de una empresa o de un usuario particular. Es también una herramienta idónea para conseguir mejorar el retorno de inversión de su negocio o empresa (ROI).

En marzo de 2005 Google compró la empresa Urchin —en aquella época, una de las herramientas más acreditadas del mercado en Análisis Web—. Google hizo una adaptación para su buscador y herramientas asociadas, así como una nueva

interfaz, y en noviembre de 2005 lanzó la nueva herramienta con el nombre de Google Analytics.

La herramienta Google Analytics es, a su vez, un conjunto de herramientas de medición de Analítica Web para la empresa que ha integrado características de Analítica Web, Analítica Móvil y Analítica Social. Proporciona las siguientes funcionalidades:

- **Estadísticas de varios canales**[11]. "Sus clientes investigan, comparan y toman decisiones de compra en distintos momentos y lugares. Los embudos multicanal de Google Analytics le permiten ver las interacciones entre distintos medios digitales y muestran cómo se combinan estos canales para crear ventas y conversiones. Lo más importante es que esta herramienta le ayuda a tomar decisiones de *marketing* fundamentales sobre las inversiones en publicidad en el nivel superior (los canales concretos) y en el inferior (por ejemplo, las palabras clave y las ubicaciones de anuncio específicas)."

- **Soluciones para dispositivos móviles**[12]. "Analítica de móviles (analizar sus anuncios, aplicaciones y sitios para móviles; evaluar los sitios web y las aplicaciones para móviles, así como las visitas desde dispositivos móviles con conexión a Internet, incluidos las tabletas y los teléfonos, tanto de gama alta como básicos). Ofrece la mejor experiencia de usuario en diferentes pantallas y dispositivos."

- **Informes sociales**. "Ayudan a medir el impacto de las redes sociales en los objetivos de su empresa y en las conversiones relacionadas. Mide las métricas que más interesan a la empresa.[13]"

Google Analytics proporciona informes estándar e informes personalizados sobre cómo los visitantes utilizan su sitio, cómo han llegado a él y qué puede hacer para que sigan visitándolo. Para la obtención de los informes, la herramienta de Google utiliza:

- Herramientas de análisis.
- Analítica de Contenido.
- Analítica de Redes Sociales.
- Analítica de Móviles.
- Analítica de Conversiones.
- Analítica de Publicidad.

CASO DE ESTUDIO: IBM WATSON DATA PLATFORM

El 28 de octubre de 2016, IBM[14] presentó en Las Vegas su plataforma **IBM Watson Data Platform,** de su computadora cognitiva Watson, con una aplicación directa a *Big Data* y al análisis de datos. El objetivo fundamental del proyecto, anunciado por IBM, es ayudar a las organizaciones a obtener el máximo valor de los datos, transformando los datos en conocimiento para la ayuda en la toma de decisiones.

Según IBM, se trata de una plataforma con el sistema de ingestión de información más veloz desarrollado hasta el momento en el mundo. Además, incorpora tecnología cognitiva para la toma de decisiones empresariales y permite a los profesionales de datos trabajar con cualquiera de los servicios disponibles en la nube de IBM.

La plataforma Watson Data Platform "integra todas las capacidades de sus tecnologías cognitivas, analíticas y *cloud* para ponerlas al servicio de los profesionales. Posibilita la colaboración entre todos los profesionales que trabajan con información (científicos de datos, ingenieros de datos, analistas de negocio y desarrolladores), ofreciéndoles distintos servicios y herramientas y una visualización sencilla que se puede compartir fácilmente con el resto de la organización. La nueva plataforma de IBM aprovecha además las posibilidades que le ofrece la tecnología *open source* **Apache Spark** y contempla la opción de incorporar datos meteorológicos recogidos por The Weather Company.[15]"

En la presentación oficial, IBM recordó cómo la digitalización está transformando el mundo y la proliferación de datos, con una generación diaria de 2.500 millones de gigabytes a escala global. En España, la presentación oficial se realizó el 23 de marzo de 2017[16], coincidiendo con la presentación del evento Watson Data Platform Summit, que se celebró en Madrid el 26 de abril de 2017.

En la página oficial[17] de la plataforma Watson Data de España se reseñan sus principales funcionalidades y sus destinatarios actuales:

Principales funcionalidades

- Herramientas específicas de usuario.
- Entorno que permite la colaboración entre todas las funciones.
- Ingesta rápida de datos de todo tipo —más rápido que cualquier otra plataforma—.
- Analítica de autoservicio sobre datos fiables y gobernados.
- Despliegue automatizado con *machine learning* cognitivo y Apache Spark.
- Acceso a múltiples tecnologías de terceros gracias a un ecosistema abierto.

Los destinatarios actuales

- **Analistas y experto de negocio** (necesita datos para tomar decisiones estratégicas de negocio).
- **Desarrollador de aplicaciones** (accede a datos y modelos, desarrollando código para construir aplicaciones API).
- **Ingeniero de datos** (gestiona, integra, gobierna y protege los datos).
- **Científico de datos** (explora y analiza los datos desarrollando modelos algorítmicos).

10.14 RESUMEN

El Análisis de Datos tiene el objetivo fundamental del estudio de los datos de una organización, con la finalidad de extraer conocimiento de éstos y tomar decisiones correctas y eficientes en beneficio de la mencionada organización. La **Analítica de Datos** (***Data Analytics***), según ISACA, "implica los procesos y actividades diseñados para obtener y evaluar datos para extraer información útil"[18]. Analítica de Datos se considera también a la ciencia de examinar datos en bruto (crudos), con el propósito de obtener conclusiones acerca de la información contenida en ellos.

1. Existen tres tipos de Analítica de Datos: Analítica Descriptiva, Analítica Predictiva y Analítica Prescriptiva.
2. La Analítica de *Big Data* permite a los usuarios analizar los datos masivos de las organizaciones, con tamaños desde terabytes hasta petabytes, de modo rápido y económico.
3. Existe gran variedad de herramientas de *software* que se usan en Analítica de Datos. Las técnicas más empleadas son: realización de consultas e informes (*quering y reporting*), visualización, Minería de Datos, análisis de datos predictivos, lógica difusa, optimización, *streaming* de audio, video o fotografía, etcétera.
4. Las herramientas de Analítica deben permitir a los usuarios analizar los grandes datos de un modo rápido y económico. Los usuarios deben ser capaces de explorar y visualizar datos masivos mediante gráficos interactivos, cuadros de mando integral (*balanced scorecards*), tableros o cuadros de control (*dashboards*), herramientas de *reporting* y *query* (informes y consultas) de resultados, así como herramientas de visualización, en tiempo real cuando sea necesario.
5. El tratamiento de los grandes volúmenes de datos requiere de las siguientes etapas: adquisición o ingestión, almacenamiento y organización de la información, análisis y presentación o visualización de resultados.

6. En la era de los grandes volúmenes, podemos considerar cuatro grandes categorías en Analítica de Datos:
   - **Analítica Web** o analítica del tráfico de datos en un sitio web.
   - **Analítica social** o análisis de datos de los medios sociales (*blogs*, wikis, redes sociales, RSS, etcétera).
   - **Analítica móvil** en dispositivos móviles, con el objeto de analizar los datos que envían, reciben o transitan en dichos dispositivos.
   - **Analítica de *Big Data*** o analítica de los grandes volúmenes de datos.
   - **Analítica Web**, según Avinash Kaushik, es "el análisis de datos cuantitativos y cualitativos de su sitio web y de la competencia, para impulsar una mejora continua de la experiencia *online* que tienen tanto los clientes habituales como los potenciales y que se traduce en unos resultados esperados (*online* y *offline*)"[19].
7. Una **métrica** es una valoración cuantitativa de estadísticas que describen tanto los eventos como las tendencias de un determinado sitio web; en realidad, es una medida cuantitativa que permite conocer el estado de un sitio web, de una página web o un proceso que se realiza en un sitio para un atributo o parámetro determinado.
8. Métricas cruciales: visitas, visitantes, visitantes únicos, tiempo en la página y en el sitio, tasa de rebote, tasa de salida, tasa de conversión, compromiso.
9. Otras métricas importantes que se deben considerar: visitas provenientes de buscadores o directas, *ranking* de páginas más vistas/páginas por visita, procedencias de las visitas.
10. Un indicador clave de rendimiento (KPI) es una métrica que ayuda, del modo más eficiente posible, a conseguir los objetivos previstos del sitio web.
11. Todas las KPI son métricas, pero no todas las métricas son KPI.
12. Los informes que emiten las herramientas de Analítica Web son variados. En el caso de Google Analytics se destacan: informes estándar, personalizados y sociales.
13. Herramientas de Analítica Web. Existe gran cantidad de herramientas de Analítica Web de pago y también gratuitas. Una breve selección puede ser: Coremetrics, Omniture, Piwik, Woopra, Google Analytics, WebTrends, Unica, etcétera.
14. Analítica Web para móviles se refiere al campo específico de la Analítica Web en el canal móvil de la organización, y comprende el conjunto de prácticas y tecnologías para coleccionar y analizar los datos de presencia en Internet desde móviles y tabletas, con el objetivo de tomar decisiones de negocio.

15. La información de las herramientas de Analítica Web debe proporcionar resultados similares a las métricas ordinarias y, además, información relativa a la actividad de los usuarios, al dispositivo y a la aplicación.

16. **Análisis de Sentimiento** o **Sentimientos** (*Sentimental Analysis*), también conocido como **Minería de Opinión**, se refiere al análisis automático del sentimiento que trata de traducir a indicadores más o menos medibles las emociones humanas inmersas en los datos sociales, tanto en fuentes externas y autónomas (redes sociales, *blogs*, *microblogs*, foros, medios de comunicación, wikis, etcétera) como internas o propias de la empresa (interacciones almacenadas en el CRM, transcripciones de conversaciones registradas en el sistema de soporte de incidencias, encuestas realizadas a clientes y empleados).

NOTAS

[1] *ISACA: Data Analytics. A Practical Approach (white paper), agosto 2011. En* <www.isaca.org/dataanalytics> encontrará el artículo citado y un buen número de recursos adicionales.

[2] *The Data Warehouse Institute (www.tdwi.org)*

[3] *Peter J. Jamack: "Analítica de inteligencia de negocios de Big Data". Disponible en:* <http://www.ibm.com/developerworks/ssa/library/ba-big-data-bi/index.html>. *IBM 2003.*

[4] *Pere Rovira en* http://www.postgradomarketingonline.com/blog/que-es-la-analitica-digital/

[5] *Digital Analytics Association. What is Digital Analytics?* http://www.digitalanalyticsassociation.org/

[6] *Recordemos que un sitio Web es el componente Web de un dominio, y un sitio Web se compone o tiene varias páginas Web. El sitio Web tiene la dirección URL de la organización y cada página tiene su propia dirección URL que se deriva de la página del sitio.*

[7] *CIO: Chief Information Officer. CMO: Chief Marketing Officer.*

[8] *The IBM Global Chief Marketing Officer Study (The IBM 2011 Global CMO Study): "Del reto al éxito. La transformación de marketing en la era digital". Disponible en:* <http://www-05.ibm.com/services/es/c-suite/cmo/cmo-study-registration-2011.html>. *[Consulta: 8 de septiembre de 2012].*

[9] *El portal Analítica Web (www.analiticaweb.es) ha publicado diferentes artículos sobre Mobile Analytics de Juan Manuel Elices, donde se analizan estrategias de analítica móvil, algunas de las cuales recogemos en esta sección.*

[10] https://www.google.es/intl/es/analytics

[11] https://www.google.es/intl/es/analytics/features/multichannel-funnels.html

[12] https://www.google.es/intl/es/analytics/features/mobile.html

[13] *-La descripción de las funcionalidades de Google Analytics se incluyen en la web del sitio de Google:* www.google.es/intl/es/analytics

[14] https://www-03.ibm.com/press/es/es/pressrelease/50924.wss

[15] www.ibm.com/analytics/es/es/watson-data-platform/

[16] https://www.ibm.com/blogs/think/es-es/2017/03/23/watson-data-platform-y-el-cientifico-de-datos/

[17] https://www.ibm.com/analytics/es/es/watson-data-platform/platform.html

[18] *Op. Cit. ISACA (2011)*

[19] *Avinash Kaushik. Analítica Web 2.0 . 2ª edición. Barcelona: Gestión 2000 , 2011. Pág 24.*

CAPÍTULO 11
ANALÍTICA WEB Y ANALÍTICA SOCIAL

Contenido

11.1 Introducción
11.2 Primeras consideraciones empresariales sobre Analítica Web
11.3 Breve historia de la Analítica Web
11.4 Métricas
11.5 Indicadores clave de rendimiento (KPI)
11.6 Informes (Google Analytics)
11.7 Herramientas de Analítica Web
11.8 Analítica Web Móvil (Mobile Analytics)
11.9 Analítica Social
11.10 Herramientas de Analítica Social
11.11 Herramientas de monitorización
11.12 Herramientas de reputación e influencia social
11.13 RESUMEN

Objetivos

- Explicar los diferentes enfoques de Analítica Web.
- Examinar las características fundamentales del concepto de métricas y KPI.
- Identificar las características fundamentales de la Analítica Web Móvil.
- Localizar herramientas de Analítica Web y encontrar las más populares, gratuitas y de pago.
- Analizar el origen y causa de la proliferación de datos.
- Analizar en detalle el análisis de sentimientos.
- Profundizar en el concepto de Analítica Social.
- Examinar métricas sociales populares.
- Conocer las KPI más notables en Analítica Social.
- Identificar herramientas de Analítica Social.
- Identificar herramientas de monitorización, reputación e influencia social y ver sus usos y aplicaciones más frecuentes.

11.1 INTRODUCCIÓN

En el capítulo 10 se hizo una introducción a las disciplinas de Analítica Web y Analítica Social y en este capítulo se pretende profundizar en estas disciplinas desde un enfoque muy práctico y analizando las herramientas de software de Analítica más populares a disposición de profesionales y empresas.

Analítica Web es una rama o disciplina de la Analítica de Datos o Analítica Empresarial, que se centra en el análisis de los datos que fluyen a través de sitios y páginas web. En realidad, el análisis de datos en la Web es más bien un análisis del tráfico web. La Web ha ofrecido datos, más datos, llegando (como conoce el lector) a los grandes volúmenes de datos (Big Data). Sin embargo, se trata de encontrar los datos significativos, y esta tarea es la difícil. Avinash Kaushik, uno de los padres de la Analítica Web, tal vez el más significativo, en el capítulo 1 de su libro Analítica Web 2.0 habla de la paradoja de los datos. Kaushik señala que "para la Web, la paradoja de los datos es una lección de humildad: sí, hay una gran cantidad de datos, pero a la hora de tomar decisiones inteligentes existen obstáculos fundamentales"[1]. En realidad, en su análisis inicial, Kaushik plantea que, tal vez, los datos no son el problema real, sino la gente; eso le lleva a considerar que los elementos nucleares del análisis del tráfico web o análisis web sean las personas (los analistas web como profesionales y los directivos que han de tomar las decisiones) y las herramientas empleadas en el análisis web.

En este capítulo, se analizarán los temas clave en la Analítica Web: conceptos fundamentales, métricas, indicadores clave de rendimiento (KPI), informes, segmentación, factores y embudos de conversión y estadísticas en tiempo real. Se detallarán las herramientas de Analítica Web más utilizadas, gratuitas y de pago (profesionales), con énfasis especial en la herramienta gratuita Google Analytics.

La explosión de los grandes volúmenes de datos, que llevamos considerando proceden de numerosas fuentes de datos; pero, sin lugar a duda, los medios sociales (social media) son responsables de grandes porcentajes en un sentido amplio. Si analizamos las estadísticas de acceso a Internet por dispositivos móviles (Internet móvil) y, sobre todo, las tendencias y cifras previstas para los próximos años, el aluvión de datos debido a los social media crecerá con cifras espectaculares, hasta el punto de que los petabytes y exabytes serán las cifras que se han de considerar.

En el caso de los medios sociales, los datos que se están acumulando proceden de multitud de fuentes (Twitter, Facebook, Google+, Amazon, eBay, Picasa, Foursquare, Instagram, Pinterest, entre otras) y cientos de miles de blogs, wikis, chats, foros, etcétera. Esta situación lleva a una sobrecarga de datos y a la necesidad de descubrir los realmente significativos para organizaciones y empresas; es decir, se necesita conocer y ponderar la relación señal/ruido. Por esta razón, el análisis de datos sociales es hoy en día una necesidad vital para organizaciones y empresas, pero también para el usuario individual, llámese empleado, estudiante, ingeniero, profesor, directivo o científico.

La Analítica Web es la parte de la Analítica que permite integrar y analizar los datos no estructurados que se encuentran en el correo electrónico, la mensajería instantánea, los portales web, los blogs y otros medios sociales, usando las herramientas de obtención de datos existentes, los informes de Inteligencia de Negocios o Empresariales y otras herramientas como los cuadros de mando integral. El análisis de la información generada en los medios sociales y dispositivos móviles permite obtener información en tiempo real sobre las tendencias de consumo. Se define el Análisis Social y se describen los componentes necesarios para utilizar esta disciplina como herramienta de gestión eficiente.

La Analítica Social o de Medios Sociales (Social Analytics o Social Media Analytics) está comenzando a ser una disciplina muy necesaria en organizaciones y empresas y constituye una de las áreas más impactantes dentro de la Analítica de Datos. Por referirse al análisis de datos en medios sociales, también se la conoce, simplemente, como Análisis Social (Social Analysis). Dada la fuerza que está adquiriendo, comienza también a considerarse una disciplina autónoma dentro de la Inteligencia de Negocios.

11.2 PRIMERAS CONSIDERACIONES EMPRESARIALES SOBRE ANALÍTICA WEB

Como ya se ha comentado, Analítica Web es una disciplina que ya está consolidada y son numerosas las obras que se refieren a ella, así como suele ser el título de numerosas asignaturas que se imparten en carreras de grado de ingeniería, economía de empresa, marketing o comunicaciones y, naturalmente, en titulaciones de máster y maestrías.

Una de las primeras decisiones que deberán tomarse en la empresa es si el Análisis Web se realiza en la propia empresa o con un proveedor externo de software de Analítica Web, que en muchas ocasiones será el propio proveedor de servicios de Internet (PSI).

Las empresas y organizaciones deben afrontar la implementación con éxito de su Análisis Web. Para ello, se deberá plantear una evaluación de las infraestructuras en TI y verificar su solidez, de modo que parece razonable pensar que si la empresa no dispone de una infraestructura sólida sería conveniente subcontratarla. En este caso, será necesario optar por contratar consultores autorizados o expertos independientes.

Dada la importancia de la elección de una herramienta de Analítica Web, será preciso examinar con detenimiento las diferentes opciones: costes, tiempo de implementación, facilidad de mantenimiento y notoriedad de los resultados.

Una primera clasificación de las herramientas de Analítica Web fue realizada por Kaushik en su ya citada obra (Kaushik 2011: 48). Él realizó una selección de las herramientas de Analítica Web existentes en esa época y las dividió en tres grandes grupos (clasificación que, a grandes rasgos, consideramos que sigue teniendo vigencia):

Grupo 1: Omniture, Coremetrics, Webtrends.

Grupo 2: AffiniumNetinsights de Unica, XiTi, Nedstat, Click Tracks.

Grupo 3: Google Analytics, Yahoo! Web Analytics.

Kaushik considera que las herramientas de este grupo son soluciones analíticas, robustas y eficientes y, además, son gratuitas, por lo que recomienda que sólo se debe pagar por la Analítica Web cuando sus necesidades sean lo suficientemente complejas como para requerir una herramienta especial (Kaushik, 2011:48-49). La herramienta más acreditada y, tal vez, la más utilizada es Google Analytics.

Otra decisión que será necesario tomar es si se realiza el análisis en el servidor o se llevan a cabo estadísticas en tiempo real. El análisis de la actividad del servidor permite medir adecuadamente una serie de datos importantes sobre el rendimiento de la presencia del sitio web en Internet, mientras que las

estadísticas en tiempo real aportan datos más exactos sobre el número real de visitantes únicos y de páginas vistas.

11.3 BREVE HISTORIA DE LA ANALÍTICA WEB

La Analítica Web es una disciplina nueva que se ha ido desgajando de las teorías de Inteligencia de Negocios y de los sistemas de información, especialmente desde el auge del marketing digital y de la Gestión de Relacione con los Clientes (CRM). Durante los primeros años de la primera década del siglo XXI, justo después de la explosión de las empresas "punto com", la Analítica Web comenzó a ver la luz en el reino de los sistemas de información de empresas con gran presencia tecnológica.

Sin embargo, la historia de la Analítica Web nació, casi en paralelo, con el nacimiento y evolución de Internet y de la Web[2]. El año 1993 se considera el despegue comercial de la Web; en este mismo año, se publicó la primera herramienta comercial de Análisis de Datos en la Web y, por ende, de la Analítica Web: Webtrends. Esta herramienta exigía amplios conocimientos informáticos para su uso y sólo manipulaba un número escaso de métricas: visitas, consumo de contenidos y duración de las sesiones. Gorostiza y Barainca (2016) señalan el nacimiento de Hit Counters —el primer contador de visitas— y otras herramientas de Analítica, tales como Accrue, Omniture o WedSideStory. En 1997, los citados autores destacan la implantación de nuevas herramientas de recolección de datos basadas en JavaScript, que "permiten recoger todas las acciones del usuario en la página web sin atender a las peticiones que se hacen a nuestro servidor"[3].

En 2004, coincidiendo con el rápido advenimiento del marketing digital, se publicó uno de los primeros libros dedicados a la Analítica Web y que mayor impacto ha tenido en la evolución actual de la disciplina: Web Analytics Demystified, de Erict Peterson. También, en 2014 nace la Asociación de Analítica Web, WAA (Web Analytics Association), referencia mundial en el sector.

Sin embargo, otra fecha clave fue noviembre de 20052, cuando Google (que anteriormente había comprado Urchin, una de las mejores herramientas de Analítica Web de aquellos años) presenta Google Analytics, un servicio que se utiliza para medir el impacto de los sitios web y de las campañas de marketing.

La siguiente fecha de impacto en Analítica Web fue el 2007, donde, al igual que en 2004, los fenómenos desencadenantes fueron el firme asentamiento de la publicidad en línea y la publicación del primer libro de Avinash Kaushik, Web Analytics: An Hour a Day, que supuso el punto de partida para el nacimiento profesional de la Analítica Web como disciplina y del analista web como profesión.

En septiembre de 2009, Adobe compró Omniture, plataforma líder mundial en Analítica Web profesional. En 2010, se producen otras adquisiciones

importantes: ComScore, uno de los servicios de medición de audiencias en Internet más reputados, compra NedStat, el primer proveedor europeo; IBM compra otras dos empresas de Analítica Web, Unica y Coremetrics. También en 2010, Avinash Kaushik publicó su siguiente obra de impacto mundial, Web Analytics 2.0, aunque fue en 2011 y 2012 cuando se inició su penetración a nivel global.

Ya en los primeros años de la segunda década del siglo XXI, Google Analytics se convierte en la herramienta de referencia de Analítica Web. En 2011 lanza su versión 5.0 y presenta una nueva interfaz con grandes cambios y nuevas funcionalidades, tales como informes en tiempo real, datos de redes sociales, embudos multicanal y paneles de control o cuadros de mando personalizados.

En 2013, Google provoca una nueva "revolución" en la Analítica Web con la presentación de su nueva herramienta, Google Universal Analytics, dotada de un gran número de funcionalidades, entre ellas soporte del seguimiento de tráfico de los dispositivos móviles con sistemas operativos IOS y Android, y la presentación de un nuevo sistema de medición denominado Universal Analytics.

En esta segunda década del siglo XXI, seguirán las noticias comerciales de Analítica Web, pero, sobre todo, se han asentado las herramientas de Analítica Web en los sistemas de información de las compañías, al igual que el software antivirus o la actualización de la última versión de sistema operativo.

11. 4 MÉTRICAS

El análisis del flujo secuencial de clics se apoya en dos componentes fundamentales: las métricas y los KPI (indicadores clave de rendimiento). En este apartado, nos centraremos en el concepto de métrica.

Una métrica es una valoración cuantitativa de estadísticas que describen tanto los eventos como las tendencias de un determinado sitio web (Kaushik, 2011). Una métrica, en realidad, es una medida cuantitativa que permite conocer el estado de un sitio web[4], de una página web o un proceso que se realiza en un sitio web, para un atributo o parámetro determinado. Existen numerosas métricas que aportan abundante información; vamos a describir las métricas más utilizadas, aquellas que Kaushik denomina "las ocho métricas cruciales", y alguna otra más de gran impacto en el Análisis Web de un sitio o página Web, aunque es preciso constatar que puede existir algún matiz diferenciador en el contexto de la herramienta Web utilizada y en la de una determinada métrica, que puede variar de una herramienta a otra, por lo que será conveniente, una vez decidida la herramienta de análisis, conocer fielmente el concepto de cada métrica en su herramienta. Esta confusión de términos de métricas es especialmente acertada en las métricas **visitante** y **visitante único**, aunque cada proveedor tratará de medir el proceso de personas reales que de verdad han visitado el sitio web y,

normalmente, no tendrá problemas en la identificación numérica de estos conceptos.

11.4.1 VISITAS

Las visitas son la cantidad de veces que una persona entra en un sitio web durante un cierto tiempo, navegando por él antes de abandonarlo. Las visitas indican el número de veces que los usuarios han estado en un sitio web en un período de tiempo determinado. Técnicamente, a este proceso se lo denomina sesión (session).

En la mayoría de las herramientas de Analítica Web, una sesión o una visita se define como aquello que ocurre entre una primera petición y la última. Es decir, si un usuario accede a un sitio web y permanece inactivo un determinado tiempo (30 minutos, en el caso de Google Analytics), la sesión se da por finalizada y cualquier actividad que lleve a cabo a continuación en el mismo sitio se contabilizará como otra visita. De igual forma, si un usuario abandona un sitio web antes de los 30 minutos (por ejemplo, 5 minutos), y vuelve a visitarlo dentro de esos 30 minutos (por ejemplo, al minuto 27), no se contabilizará como una segunda visita.

Será preciso verificar en su herramienta de análisis (o que se lo aclare su proveedor) el concepto de sesión o visita y, como veremos en los apartados siguientes, visitante y visitante único.

También se debe tener en claro que una visita conlleva entrar en una o varias páginas dentro de un sitio web, de forma que si un visitante ha visto cuatro páginas del sitio web, a efectos estadísticos, es una única visita, aunque es posible también conocer el número de visitas de una página concreta; este caso es de mucho interés para el Web Master y los responsables de marketing, porque les permitirá conocer la importancia de las diferentes páginas web del sitio y evaluar si está bien diseñado y construido.

11.4.2 VISITANTE

Este término, como tal, suele dar lugar a confusión y, además, no es una métrica importante. Si un visitante (visitor) entra cinco veces en un sitio web, se contabilizará como un único visitante.

11.4.3 VISITANTE ÚNICO

Esta métrica es una de las más importantes y significativas en el análisis de un sitio web. Los visitantes únicos (unique visitor) es el número de personas diferentes que han visitado (accedido) a un sitio web. Es decir, si un visitante

entra cinco veces a un sitio, se contabiliza como un único visitante, y lo mismo se refiere a períodos de tiempo: si un mismo visitante o persona entró en un sitio web cincuenta veces en un mes, se considera que es el mismo visitante. Como señalamos antes, es una de las más empleadas y de mayor éxito, sobre todo en el posicionamiento en buscadores y en el éxito de las campañas de marketing de las empresas, ya que si éstas triunfan, entrañarán, con toda probabilidad, un aumento del número de visitantes únicos.

Sin embargo, el indicador de visitantes únicos no es una medida exacta, sino aproximada, al menos por ahora. Es probable, aunque no siempre sea cierto, que cada visitante único sea una única persona. Por consiguiente, como señala Kaushick, hay que entender que, aunque el indicador de visitantes únicos sea un representante válido del número de personas únicas que visitan su sitio web, no es una medida perfecta. Veamos algunos casos que ilustrarán el concepto aproximado de visitante único.

- Mi amigo Luis Mackoy accede a un sitio web por la mañana en la computadora del trabajo, por la tarde accede desde su iPad (a través de 4GWi-Fi) y por la noche, desde la computadora personal de su casa. En la mayoría de las herramientas, se contabilizarán las visitas como tres visitantes únicos cuando en realidad es una única persona.

- Mi amigo Luis Mackoy está con su alumno de doctorado visitando el sitio de la revista Business Week con su laptop. Cuando termina, y sin cerrar el sitio, le presta a su alumno su computadora para que continúe visitando el sitio web porque le ha solicitado ver la sección “Technology”. Es una única visita y, por el contrario, son dos personas diferentes. El navegador utilizado en cada caso puede influir también, dependiendo de que el navegador no admita cookies o rechace las de terceros, aunque cada vez más, las herramientas de analítica modernas utilizan cookies de primer nivel que no suelen rechazar. Así se pueden dar numerosos casos.

- Acceso a un sitio web desde una computadora con el navegador Explorer y, posteriormente, en la misma computadora con el navegador Firefox. Normalmente, se contabilizará como dos visitantes únicos. Éste es el caso de Google Analytics, que interpreta que cada navegador es un visitante único y lo señala mediante una cookie para reconocerlo en visitas posteriores.

Pese a estos inconvenientes que los proveedores de soluciones de analítica trabajan para corregir, la métrica de visitante único sigue siendo un indicador excelente para determinar el número de personas físicas que visitan un sitio web. Algunas herramientas (Omniture) denominan a la métrica visitante único y absoluto (absolute unique visitor).

11.4.4 TIEMPO EN LA PÁGINA Y TIEMPO EN EL SITIO

El indicador tiempo (*time*) junto con las visitas y visitantes son indicadores muy importantes en analítica web. Este indicador mide el tiempo que los visitantes pasan en una página determinada así como el tiempo que pasan en el sitio durante una determinada sesión o visita. Se suele utilizar también el término *promedio de tiempo en una página web* como la métrica que nos permite conocer cuánto tiempo pasan de media los usuarios en cada página web y también el término promedio de tiempo en un sitio web para indicar cuánto tiempo pasan de media en un sitio web

Al igual que en el caso de las métricas anteriores, los tiempos efectivos de la visita dependerán de la herramienta de Analítica utilizada, y es un valor muy difícil de estimar con exactitud. El registro del tiempo es difícil, porque se conoce bien el momento de la entrada en el sitio, pero no tanto el momento de salida, y no es lo mismo estar seis segundos en un sitio o en una página web que estar diez minutos, ya que en el primer caso no da tiempo a realizar ninguna lectura apreciable, y sí en el caso de diez minutos. Lógicamente, no es lo mismo visitar una página de un sitio web que cinco páginas. En resumen, es necesario conocer el sistema para medir la permanencia en la página y en el sitio, o al menos un tiempo aproximado que permita utilizar la métrica para obtener un beneficio importante para la empresa.

11.4.5 TASA DE REBOTE

La tasa de rebote (*bounce rate*) es el porcentaje de visitas que llegan a un sitio o página web cuando se ha visto una única página y lo abandonan rápidamente sin hacer un solo clic ni ver nada más en ese sitio web. Los porcentajes de abandono alto, normalmente, presuponen que las visitas que han llegado no eran cualificadas o no estaban interesadas en los contenidos de la página web. Google Analytics considera un abandono o tasa de rebote cuando un usuario accede a una página o sitio web, por cualquier procedimiento, y abandona la página en uno o dos minutos, sin hacer más clics o visitar más páginas.

Una tasa de rebote alto (79% u 80 %) en un determinado periodo de tiempo significa que el visitante no tiene ningún interés en el sitio o página web. Por el contrario una tasa de rebote baja (10 % ó 15 %) significa que al visitante le gustan los contenidos del sitio web. La página por la que abandona la visita el usuario se denomina *página de salida* y es un indicador muy efectivo, aunque no siempre es significativa. Por ejemplo, en los sistemas de comercio electrónico, la página de salida es muy relevante, ya que dependerá de si el usuario desea visitar el sitio, ver su contenido, navegar por opciones o, por el contrario, realizar una compra, para la cual hará clic en el producto seleccionado, o lista de la compra, y saldrá directamente.

El porcentaje de salidas de las diferentes páginas dependerá del contenido de esas páginas o las acciones que proponen en ellas. Para el director de marketing, será muy importante conocer los informes del porcentaje de salidas de las diferentes páginas del sitio web.

11.4.6 TASA DE CONVERSIÓN

La tasa de conversión (*conversion rate*) es el porcentaje de resultados u objetivos conseguidos en relación al número de visitas. Es una métrica muy significativa en la gestión empresarial, ya que suele mostrar el interés por el sitio web, las compras realizadas por visita, el rellenado de un formulario, la reserva de un viaje o una consulta, etcétera.

La tasa de conversión se suele medir como un porcentaje, normalmente, en tantos por ciento (número de resultados obtenidos en la visita por cada 100 visitas). Por ejemplo, en el caso de un comercio electrónico "x", una tasa de conversión aceptable puede ser el 2% o el 3%, que significa que por cada 100 visitas se efectuarán dos o tres ventas, para el comercio "a" ¿Cuál puede ser una buena tasa de conversión? En comercio electrónico, 2% ó 3% suele ser una buena tasa, aunque dependerá del tipo de negocio, y esta tasa puede no ser aceptada en muchos casos y puede no ser aceptable para el comercio "b". Otro ejemplo puede ser el caso de la entrada a un sitio web donde se muestra un artículo determinado, cuyo objetivo es que pueda ser leído por el mayor número posible de usuarios. Si en un mes entraron 100.000 visitas y 29.967 leyeron un artículo determinado (es decir, alcanzaron el objetivo previsto), la tasa de conversión será de 29.967/100.000, es decir, 29,967 (prácticamente, el 30%).

11.4.7 COMPROMISO

El compromiso (*engagement*) es una métrica difícil de medir, pero de gran importancia para la gestión empresarial si se puede llegar a cuantificar. Kaushik plantea que la métrica compromiso busca conseguir los motivos emocionales o de sentimientos por los cuales un usuario decide navegar por un determinado sitio web y no otro. En esta métrica aparecen conceptos tales como simpatía, confianza, orgullo, etcétera.

Se puede considerar que el número de veces que un usuario visita un sitio web, así como la frecuencia, permite deducir el nivel de compromiso. Por ejemplo, suelo visitar casi a diario (mañana y noche), e incluso a lo largo del día, determinados periódicos españoles y latinoamericanos y algunas revistas de tecnología y negocios de Estados Unidos e Inglaterra, siempre desde tres dispositivos distintos: la computadora de mi oficina, mi teléfono inteligente y la computadora de mi casa. Estas acciones implican un cierto compromiso con estos medios de comunicación. Cuantas más páginas vea un visitante, más alto

será su nivel de compromiso; por ejemplo, tengo un alto nivel de compromiso con la página inicial de cada uno de los medios que leo con detenimiento, y otro alto nivel de compromiso con la sección (pestaña) de tecnología donde leo las noticias que publican el día de la visita.

Existen otras métricas que pueden reflejar el nivel de compromiso: el tiempo en el sitio web, el registro o identificación en un sitio web, la suscripción a un servicio RSS o un boletín, la publicación de un comentario o la descarga de contenidos. Así, por ejemplo, si un usuario visita con mucha frecuencia el excelente sitio Slideshare, donde hay numerosa y excelente documentación de negocios, tecnología e innovación, y además está identificado y lo visita con frecuencia, cuando desea entrar o realizar una descarga desde la PC de su oficina u hogar, no necesita identificarse en cada entrada, pues Slideshare lo reconoce con el solo hecho de estar conectado.

En resumen, el compromiso es una métrica difícil de medir, porque existen muchos conceptos que es preciso analizar, pero una vez que se consiguen estos valores, los sitios web se ven muy favorecidos y, de igual modo, los usuarios del sitio.

11.4.8 OTRAS MÉTRICAS

Las métricas anteriores son consideradas por Kaushik como las métricas fundamentales, pero existen otras métricas de gran importancia en Analítica Web, que describimos a continuación.

Visitas provenientes de buscadores o directas

Algunos sistemas de estadísticas permiten desglosar el origen de las visitas: indirectas (el usuario escribe la dirección URL), directas (el usuario tiene seleccionado el sitio en su marcador de favoritos del navegador y sólo hace clic sobre la dirección correspondiente), de otras fuentes web (llegan redireccionados de otros sitios web) y de buscadores. Es muy importante conocer el lugar de procedencia de las visitas. Por ejemplo, si un sitio web está bien posicionado en los buscadores (Google, Firefox, Bing) registrará habitualmente un porcentaje alto proveniente de buscadores; ésta es una de las tareas importantes de los especialistas SEO y analistas web.

Una información muy interesante es la distribución de las visitas naturales por motor de búsqueda; otra información vital son las visitas provenientes de enlaces patrocinados. Por ejemplo, Google Analytics separa las visitas provenientes de resultados naturales o visitas de campañas de marketing de enlaces patrocinados en Google.

Ranking de páginas más vistas/páginas por visita

Un sitio web está compuesto de varias páginas, las cuales difieren entre sí por el número de veces que son visitadas por los usuarios. Muchas herramientas de Analítica publican un ranking de las páginas web más visitadas de un sitio. Esta métrica es muy valiosa para conocer los contenidos más atractivos para un usuario. Las campañas de posicionamiento en buscadores de éxito permitirán deducir cuáles son las páginas más atractivas para los usuarios y viceversa.

Las páginas por visita (*pages/visit*) es el promedio de páginas visualizadas por visita al sitio web. Un valor de cuatro páginas por visita significa que cada visita consulta un promedio de cuatro páginas. Otra métrica interesante que proporciona Google Analytics es el porcentaje de visitantes nuevos (*new visitors*) y visitantes que regresan al sitio o a la página (*returning visitors*).

Procedencia de las visitas

Esta métrica puede ser independiente o asociada a la métrica de visitas, dependiendo de la herramienta de Analítica. Se trata de conocer la procedencia de las visitas que llega al sitio web, y que se calcula por la dirección del protocolo IP (Internet Protocol) de la conexión. Esta métrica es muy importante para las organizaciones y empresas, aunque la dificultad para detectar la zona geográfica, el país y la ciudad dependerá del medio de comunicación al que se accede a Internet: línea fija (por cable), líneas ADSL, líneas de fibra óptica, redes móviles, redes inalámbricas, satélite, etcétera.

11.5 INDICADORES CLAVE DE RENDIMIENTO (KPI)

Los resultados del negocio deben permitir definir los objetivos del sitio web. Según el tipo de negocio y sus propósitos, se han de elegir las métricas adecuadas para obtener los resultados para los que se ha diseñado el sitio web.

Un indicador clave de rendimiento o desempeño, **KPI** (Key Performance Indicators), es una métrica que ayuda del modo más racional a conseguir los objetivos previstos del sitio web. **En consecuencia, todos los KPI son métricas, pero no todas las métricas son KPI**. Dependiendo de los objetivos estratégicos de la organización o empresa, se diseñan los objetivos del sitio web y se deberán deducir, dependiendo del modelo de negocio, cuáles son las métricas idóneas que deberán convertirse en KPI; es decir, se deben seleccionar y determinar cuáles son las métricas que ayudan a conseguir los objetivos del negocio. Los KPI son muy variados y dependerán de las métricas que pueden obtenerse del análisis del tráfico de la Web.

Un caso típico para ver los KPI, adecuado a un modelo de negocio, es el caso de una tienda tradicional o supermercado. Los indicadores clave de rendimiento típicos son valores totales de venta por hora, valores promedio de ventas por

cliente (o importe de la factura media), artículos por venta, ventas por vendedor y ventas por metro cuadrado. En el caso de una tienda de comercio electrónico, los indicadores KPI importantes pueden ser tasa de abandono (número de visitantes que entran en la página del sitio web y lo abandonan sin navegar por el sitio), tasa de conversión (porcentaje de visitas que entran en el sitio web, toman el carrito de la compra y realizan una compra efectiva), tiempo de permanencia en el sitio y páginas visitadas de las categorías de artículos ofertados, horario y día de la semana preferente de las visitas y lugar geográfico de acceso al sitio. Los mejores KPI, en general, son los propios de un negocio. Los expertos de marketing suelen considerar que de las métricas tradicionales (visitas, páginas vistas, tiempo en página, objetivos cumplidos, etcétera) utilizadas como KPI, se caracterizan porque cuando se modifican, implican una variación en la cuenta de resultados.



Figura 11.1. Determinación de un KPI
Fuente: (Acera, 2012:59)

Lovett (2012) hace una analogía muy interesante entre métricas y KPI: las métricas son el flujo de datos que actúa como parte vital de sus operaciones de social media y los KPI son los signos vitales. Los KPI son como si fueran la temperatura, el pulso, la respiración y la presión sanguínea de sus actividades de social media. Sin los KPI, insiste Lovett, es difícil decir si sus programas están vivos y, mucho menos, si funcionan adecuadamente. De esta forma, los KPI que funcionan correctamente dan una idea muy ajustada de una buena planificación de los social media y si un KPI falla puede, ser indicio de que algo va mal en sus programas o medios sociales y deberá realizar inmediatamente un diagnóstico para descubrir las razones y evitar un fracaso total. Según Mortensen5, un KPI debe tener las siguientes características:

1. Deberá mostrar el objetivo de la organización.
2. Ser definido por la dirección de la empresa.
3. Proporcionar contexto (todos los KPI han de tener contexto, ya que sin él los KPI sólo son números que no significan nada para los no iniciados).
4. Tener significados en distintos niveles (significado y contexto están estrechamente enlazados, como se estudió anteriormente en la definición).

5. Estar basados en datos reales.
6. Ser fácilmente entendibles (establecer expectativas).
7. Conducir la acción (autorizar acciones).

Ejemplo práctico simple de métrica *versus* KPI

Supongamos un colegio público o privado de estudiantes de bachiller. Algunas métricas y KPI pueden ser:

- **Métricas**: número de alumnos, número de profesores, horas de clase, horas de descanso, etcétera.
- **KPI**: porcentaje de aprobados, tasa de ingreso en la universidad, premios nacionales de bachiller conseguidos, etcétera.

11.6 INFORMES (GOOGLE ANALYTICS)

Una vez que se han definido las métricas y los indicadores clave de desempeño (KPI), es preciso considerar dos conceptos importantes en el análisis de un sitio web: los informes y la segmentación.

Los informes, como su propio nombre señala, son los datos que proporciona la herramienta de Analítica que permitirá a los Web Master, Community Manager y directivos y empleados de la empresa el análisis significativo de dichos datos, con el objetivo de poder tomar decisiones del modo más eficiente.

Consideremos el caso de Google Analytics (`www.google.com/intl/es/analytics`). Ofrece tres tipos de informes: informes estándar, informes personalizados en pestañas con idéntico nombre y una tercera opción en la pantalla de inicio, de informes sociales.

Google proporciona una vasta información al usuario con informes de todo tipo y muy extensos, donde es posible disponer de datos fiables de casi todos los indicadores fundamentales en el análisis del tráfico del sitio web.

Informes estándar

Seleccionando la pestaña “informes estándar”, aparece un menú de opciones en la pantalla donde se muestran los temas sobre los que la herramienta proporciona información:

- **Público**. Datos sobre los visitantes.
- **Publicidad**. Datos sobre las campañas en la aplicación de publicidad Google Adwords.
- **Fuentes de tráfico**. Datos sobre el origen o fuente de las visitas.

- **Conversiones**. Datos sobre las conversiones (conseguir los objetivos del sitio web).

Cada una de estas opciones presenta, a su vez, un determinado número de informaciones, entre las que se pueden destacar, dependiendo de la opción seleccionada, las siguientes:

- Visitantes.
- Ubicación (país, ciudad o continente de donde procede la visita).
- Idioma (datos en función del idioma del navegador).
- Frecuencia y visitas recientes.
- Visitantes nuevos contra recurrentes.
- Tráfico.
- Visión general del contenido.
- Páginas, páginas de destino y páginas de salida.
- Otros informes.

Informes personalizados

Los informes personalizados son informes específicos creados expresamente por el usuario. En Google Analytics, cuando se selecciona la pestaña del mismo nombre, aparece una ventana de "informes personalizados" que, a su vez, tiene las siguientes opciones: "Visión general", "Contenidos consumidos" y "Conversiones". La creación de un nuevo informe se realiza pulsando la opción "Nuevo informe personalizado" y, a continuación, se configura el nuevo informe deseado.

De esta forma, las empresas o los usuarios propietarios de un sitio web pueden configurar informes a medida, con el objetivo de rentabilizar los datos obtenidos por la herramienta analítica.

Informes sociales

La mayoría de las herramientas de Analítica actuales disponen de la posibilidad de proporcionar informes sociales, procedentes de medios sociales. Google tiene una opción de esta categoría. Los informes sociales le ayudan a medir el impacto de las redes sociales en los objetivos de la empresa y en las conversiones realizadas. De igual modo, los datos sociales y web integrados aportan una visión integral de su contenido y comunidad.

Google Analytics facilita medir el impacto de las redes sociales con las métricas que puedan interesar a la propia empresa; por ejemplo, el informe de conversiones permite cuantificar el valor que aportan las redes sociales. Otro

indicador importante es el descubrimiento de fuentes sociales que remiten a los visitantes más implicados. Es importante también descubrir cuáles son los contenidos que comparten sus visitantes y dónde lo hacen.

Estos informes sociales forman parte de la Analítica Social. Google Analytics ofrece a sus clientes diversas soluciones de Analítica, que potencian la eficacia de la Analítica. Así, las soluciones ofertadas, en el momento de la escritura de este apartado, son:

- Analítica de Contenido.
- Analítica de Móviles.
- Analítica de Conversiones.
- Analítica Social.
- Analítica de Anuncios.

11.7 HERRAMIENTAS DE ANALÍTICA WEB

En el mercado, existen un gran número de proveedores de herramientas de analítica web, tanto gratis (normalmente de código abierto "*open source*") como de pago (propietarias). Google Analytics está considerada como la herramienta más popular y más utilizada; tiene la ventaja de ofrecer dos versiones: una gratuita Google Analytics y otra de pago Google Analytics Premium, con muchas opciones y funcionalidades, aunque la versión básica se puede utilizar para muchas aplicaciones y alcanza a profesionales y empresas. Dada la gran cantidad de herramientas de analítica web existentes en el mercado y también un gran número de estudios comparativos sobre estas herramientas, hemos seleccionado varios informes que suelen tener gran aceptación para identificar las herramientas de analítica web más populares y eficientes, entre los que destacamos

TrustRadius: TrustRadius es el sitio de revisión más confiable para tecnologías de negocios, tanto para compradores como para proveedores. Un sitio web de referencia como plataforma de evaluación de soluciones software de tecnologías de negocios y que se enfoca en usuarios (compradores y vendedores) que necesitan soluciones de software para negocios. Tiene un énfasis especial en analítica web y publica sendos informes de Analítica Web relativos a "listados completos de herramientas y proveedores" (`www.trustradius.com/web-analytics`) y también un ranking de herramientas Top Web Analytics donde ya realiza una selección en función de las puntuaciones recibidas por usuarios profesionales, organizaciones y empresas.

Forrester. La prestigiosa consultora Forrester publica anualmente numerosos informes sobre proveedores y herramientas de software conocidos como The

Forrester Wave (muy similares a los cuadrantes mágicos de Gartner ya citados varias veces a lo largo del libro). En el caso particular de proveedores de herramientas de analítica web, se llama The Forrester Wave: Web Analytics y la edición publicada a finales de 2017 con el propósito de que tenga impacto a lo largo de 2018, es *Web Analytics, Q4 2017.* (https://www.forrester.com/report/The+Forrester+Wave+Web+Analytics+Q4+2017/-/E-RES136199).

Listado de herramientas Web de TrustRadius

El catálogo de herramientas (productos y servicios) denominado TrustMap (www.trustradius.com/web-analytics) contempla tres categorías diferentes de herramientas (todas ellas con informes sobre las mismas y enlaces a sus sitios web).:

- Herramientas de analítica web incluidas en el Top Rated Web Analytics Tools: Google Analytics, StatCounter, Google Analytics Premium, Woopra, Mantomo (antigua Piwik, herramienta muy popular) y Mixpanel.
- Otras herramientas de analítica web destacadas: Adobe Analytics, AT Internet, Chartbeat, Kissmetric, Webtrends Analytics y IBM Digital Analytics.
- Listado completo de herramientas de analítica web (en junio de 2018, estaban publicadas 75 soluciones). Cada una de ellas contiene además de su enlace Web, revisiones sobre la herramienta, alternativas a ella y comparación con otras herramientas

The Forrester Wave: Web Analytics, Q4 2017

La consultora Forrester evalúa a proveedores de analítica web y los clasifica en su diagrama final en función de sus coordenadas: presencia en el mercado y oferta actual. En la edición Q4 de 2017, informe de expectativas para 2018, clasifica a los proveedores de acuerdo con 32 criterios. Los siete proveedores más significativos seleccionados por Forrester son: Líderes (Adobe y AT Internet), desempeño fuerte (IBM y Google), competidores (Webtrekk, Mixpanel y Cooldata)

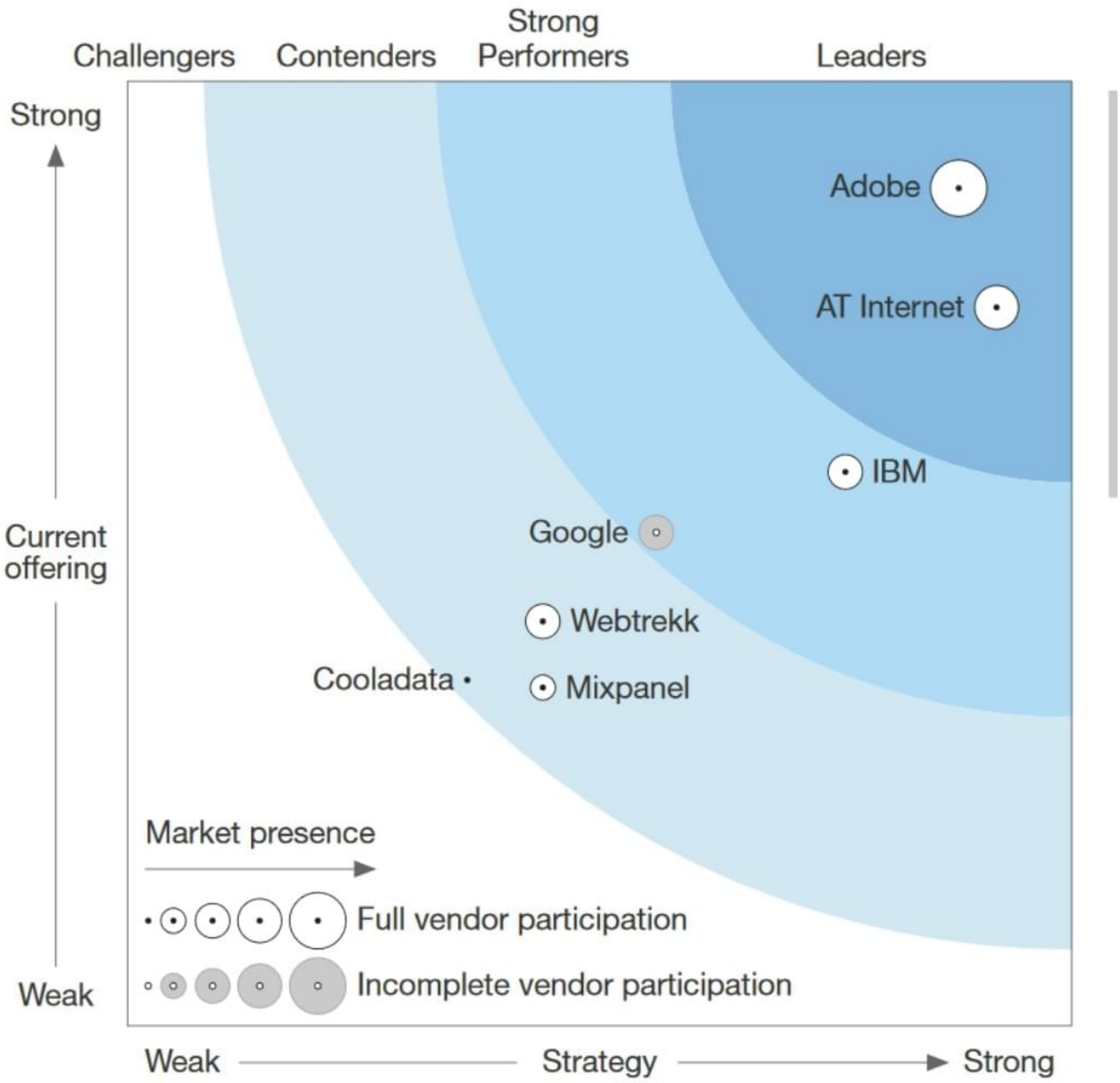


Figura 11.2. Informe Forrester Wave: Web Analytics Q4 2017
Fuente:
https://www.forrester.com/report/The+Forrester+Wave+Web+Analytics+Q4+2017/-/E-RES136199

Otras herramientas de analítica web populares y muy utilizadas (todas ellas se pueden consultar en el listado de herramientas de TrustRadius)

Las herramientas de analítica web gratuitas y más populares (algunas citadas anteriormente) son:

- Google Analytics
- Mantomo (antigua Piwik, herramienta muy popular)
- StatCounter
- Open Web Analytics (OWA). Herramienta de código fuente

Otras herramientas muy extendidas en su uso por profesionales y empresas son:

- CrazyEgg
- Flurry
- Clicky

11.8 ANALÍTICA WEB MÓVIL (MOBILE ANALYTICS)

Analítica Web para móviles o Analítica Móvil (Mobile Analytics) nace de la creciente necesidad de las empresas de conocer el retorno de la inversión de su canal móvil. A medida que aumenta la adopción de la telefonía móvil para el acceso a Internet, las empresas necesitan estar presentes en el canal móvil a través de su sitio web (web app) o aplicaciones nativas. La Analítica Móvil se refiere al campo específico de la Analítica Web, en el canal móvil de la organización, y comprende el conjunto de prácticas y tecnologías para coleccionar y analizar los datos de la presencia en Internet desde móviles (smartphones) y tabletas, con el objetivo de tomar decisiones de negocio. Una de las primeras decisiones que ha de tomar la empresa, antes de proceder a la fase de Analítica de datos, es seleccionar una aplicación web o nativa. El objetivo es conocer cuál es el modo en que el usuario interactúa con la aplicación o con la página web desde el móvil.

La Analítica Móvil debe poder responder a preguntas tales como:

- ¿Quién está utilizando las aplicaciones móviles de mi empresa o del profesional individual?
- ¿Cuál es la mejor aplicación: aplicación web o aplicación nativa?
- ¿Qué productos y servicios demandan los usuarios de teléfonos móviles o tabletas?
- ¿Cómo funciona mi sitio web desde mi dispositivo iPhone, Android, BlackBerry o Windows Phone?

El servicio de Analítica Web Móvil7 debe incluir al menos:

- Recolección de datos.
- Análisis (exploración y recomendaciones de negocio).
- Entrega y gestión de informes (reporting).

En la recolección de datos existen dos tipos de dimensiones: tipos de dispositivos por medir y plataformas sobre las que se medirán. En la etapa de análisis será preciso contemplar cuáles son los KPI (indicadores clave de rendimiento) para medir los objetivos de atención al cliente, ventas, contenidos de las páginas, etcétera. Y en la etapa de informes (reporting), entrega y gestión

de informes deberá estudiarse la integración de los datos con todos los canales de la empresa.

Información de las herramientas de Analítica Móvil

Las herramientas de Analítica Web Móvil, al igual que las herramientas de escritorio, deberán proporcionar información que pueda ser de utilidad para la empresa o el profesional. Algunas de las informaciones de interés que ha de proporcionar la herramienta de Analítica Móvil son:

Relativas a la actividad de los usuarios

- Usuarios activos.
- Duración de la sesión.
- Participación del usuario (duración promedio de sesión, screens/session).
- Visitantes nuevos y recurrentes.
- Fidelización de los usuarios.
- Datos demográficos (idioma, país/territorio, ciudad).
- Flujo de interacción.

Relativas al dispositivo y a la aplicación

- Nombre y versión de la app.
- Número de instalaciones, bloqueos y excepciones.
- Ingresos de la app.
- Proveedor de servicios.
- Marca del dispositivo móvil.
- Resolución de pantalla.
- Versión del SO (sistema operativo).

Herramientas de Analítica Móvil

El sistema normal de medición de estadísticas en los sitios web de escritorio no se puede aplicar en su totalidad a las páginas de móviles y es necesario tener en cuenta otras consideraciones. En primer lugar, el acceso a Internet en el caso de dispositivos móviles no suele ser tan seguro como es el caso de aplicaciones de escritorio. El acceso a las aplicaciones no está garantizado, ya sea por falta de cobertura 3G o 4G/5G, inexistencia de red Wi-Fi, bajas velocidades de acceso, ausencia de sincronización, etcétera.

Existen numerosas herramientas de Analítica Web Móvil, tanto gratuitas como de pago, así como de modalidad *freemium* (servicios básicos gratuitos y servicios avanzados o especiales de pago). Una selección de las herramientas más utilizadas en Analítica Web Móvil, y con los tres tipos de versiones (propietarias, códigos abiertos o gratuitos), es la siguiente:

- Google Mobile Analytics. Funcionalidades específicas para aplicaciones web (apps) y marcación (HTML5).
- FlurryAnalytics. Aplicación específica para Analítica Móvil.
- Adobe Omniture/Adobe SiteCatalyst. Permite trabajar con librerías para apps y marcación de HtML5.
- Countly. Es una aplicación de Analítica en tiempo real para la medición de aplicaciones móviles. Es una aplicación de código abierto (open source).
- Localytics. Plataforma para medición de apps.
- Piwik Mobile. Similar a Piwik de escritorio.
- Apsalar. Se ha fusionado con Singular para formar una "plataforma analítica unificada" para dispositivos móviles.
- Woopra Mobile (igual que Woopra de escritorio).

11.9 ANALÍTICA SOCIAL

La Analítica Web analiza la actividad básica de un sitio o página web específico, de un sitio de comercio electrónico, un blog o wiki. Sin embargo, a medida que los medios sociales han llegado a todas las organizaciones y empresas y han aparecido numerosas plataformas sociales, la Analítica Web sigue manteniendo su vigencia, pero ha aparecido la Analítica Social (ya comentada con anterioridad), que se encuentra en pleno desarrollo y con una creciente evolución, y que hace uso de todos los recursos de la Analítica Web, además —y sobre todo— de herramientas de visualización apoyándose fundamentalmente en cuadros de mando o tableros de control (dashboards) soportados en la gran variedad de gráficos de todo tipo: infografías, mapas o las modernas Narraciones de datos (Data Storytelling).

En el capítulo 10 ya hicimos una introducción a la Analítica Social (Analítica de Medios Sociales o Redes Sociales) y ahora en este apartado profundizaremos en este campo de la Analítica y, sobre todo, en el aspecto práctico de métricas/KPI sociales y herramientas específicas de Analítica Social.

La enorme proliferación de datos sociales requiere de técnicas especializadas de medición con el objeto de convertir los datos en información relevante que permita tomar decisiones acertadas en la organización. La Analítica Social

requiere la cuantificación y cualificación de los elementos ordinarios que se procesan en las redes sociales, tales como enlaces, fans, comentarios, emociones, influenciadores, etcétera. Es necesario medir las acciones, los sentimientos, las emociones, la relevancia, la influencia, etcétera, y convertirlas en medidas tangibles; es decir, además de la contabilización y medición de datos, se han de analizar las conversaciones de los diferentes usuarios y empresas en las redes sociales.

Diversos estudios sobre el exceso de información se publican con frecuencia, y casi todos ellos coinciden. La recopilación de datos por parte de los usuarios requiere saber lo que se está buscando y examinar los conjuntos de datos que están a su disposición. En mayor medida, las organizaciones y empresas requieren de un análisis de datos preciso, fiable y oportuno ante el aluvión de los grandes datos; en resumen, se necesita un enfoque metódico para evitar dicha sobrecarga.

Lovett (2012:69-72), uno de los grandes especialistas mundiales en Analítica Social y socio de Web Analytics Demystified, consultora muy respetada en Analítica de Datos, recomienda los siguientes pasos para evitar la sobrecarga:

1. Establecer expectativas sobre lo que espera aprender de los datos.
2. Clasificar sus iniciativas de Análisis de Social Media específicas. Averiguar métricas de éxito en social media.
3. Poner en práctica el código de requerimiento sabiamente.
4. Activar la recopilación de datos en piezas pequeñas.
5. Analizar los datos para separar la señal del ruido.
6. Mejorar sus expectativas, métodos de recopilación y análisis.
7. Informar de sus descubrimientos a los accionistas clave de la organización.
8. Evaluación constante de los planes de social media.

La sobrecarga de información, desde un punto de vista tecnológico y en el mundo conectado en que el vivimos, es inevitable; sin embargo, una buena gestión de social media con programas adecuados y un Análisis de los Datos conseguirá su uso racional y una relación señal/ruido aceptable en la mayoría de los casos.

Si bien a veces puede resultar que, pese a la buena gestión de los medios sociales, éstos son tan abiertos y libres, que pueden producir la sobrecarga de los datos, simplemente, con el incremento de las comunicaciones a través de dispositivos móviles (tabletas, teléfonos inteligentes, computadoras portátiles, *netbooks*, lectores de *e-books*, etcétera). Un caso de estudio, de impacto, se produjo con ocasión de los Juegos Olímpicos, celebrados en Londres (julio y agosto de 2012), donde los aficionados que asistían a la inauguración fueron

advertidos de que evitaran mandar mensajes de texto y publicar tuits que no fueran urgentes durante la celebración de las pruebas, porque podían provocar una sobrecarga de datos en las redes de comunicación, que afectaría a la cobertura televisiva. Efectivamente, así sucedió en varios eventos: los comentaristas de la prueba de ciclismo en ruta fueron incapaces de relatar en qué punto se encontraba el cabeza de carrera, por problemas en la red de datos que impedían su geolocalización a partir del sistema de navegación por satélite, GPS, que llevaban los ciclistas. Según un portavoz del Comité Olímpico Internacional, el problema se produjo por los mensajes que fueron enviados por los cientos de miles de fans que salieron a la calle para animar a su equipo británico.

Se requieren herramientas avanzadas y adecuadas para el procesamiento automático de datos sociales que normalmente se presentan como flujos continuos de datos en tiempo real (data streaming). En la actualidad, todos los planes de social media de las empresas requieren la necesidad de un plan de Analítica Social complementada y/o integrada con las técnicas de Analítica Web.

La Analítica Social requiere el uso eficiente de herramientas que han de soportar las métricas e indicadores clave de rendimiento o desempeño (KPI) específicos a los medios sociales.

11.9.1 Métricas sociales

Las redes sociales suelen tener cada una de ellas sus propias métricas, aunque existen una serie de métricas sociales (o indicadores KPI) comunes, que se pueden aplicar a la mayoría de las redes sociales y que están relacionadas con el volumen de usuarios y el grado de participación en las mismas. La información que proporcionan las métricas sociales permitirá averiguar una buena información relativa a la empresa, tales como el sentimiento que genera la marca (opiniones positivas, negativas o neutras), fiabilidad y aceptación de los contenidos mejores y peores en cada red social, cuál es la red social que dirige más tráfico a nuestro sitio web, etcétera.

Algunas métricas sociales comunes empleadas para la evaluación de campañas en redes sociales, dependiendo de los objetivos de las marcas, son las siguientes: reputación, popularidad, autoridad, alcance, *engagement* (interacción/compromiso), conversión y fidelización.

Reputación

La reputación o percepción de la marca (organización o empresa) exige practicar la escucha activa para tratar de averiguar cuáles son los temas que se mencionan de la marca y cuál es la imagen que reflejan de dicha marca.

Popularidad

Esta métrica mide el impacto que tiene la presencia de la marca en los medios sociales. Los indicadores que se han de medir son:

- Alcance de las publicaciones del sitio web (número de usuarios a los que llegan las publicaciones).
- Número de suscriptores.
- Número de enlaces entrantes (por la importancia que tiene en los algoritmos de posicionamientos en buscadores).
- Número de *likes* recibidos.
- Menciones y respuestas obtenidas.

Influencia

Analiza la difusión de los mensajes del sitio web y la medida en que sus seguidores transmiten nuestros mensajes a su propia red. Así, métricas por observar son:

- Número de clics que una página web de la empresa ha tenido o ha sido compartida en redes sociales.
- Menciones a los contenidos de una página web.
- Hora más idónea para publicar en redes sociales.
- Formato del contenido con mejores respuestas (texto, videos, imágenes, enlaces, infografías).
- Temática de los contenidos con mayor alcance: noticias, productos, publicidad, artículos, concursos, etcétera.

Esta métrica está directamente relacionada con el número de seguidores que se tienen en los distintos medios sociales en que está presente, así como las causas del aumento o disminución de seguidores.

Autoridad

Se refiere a la autoridad del contenido de nuestra marca en la Web, que se mide por las referencias y menciones sobre la presencia y contenido de nuestro sitio web en otros sitios web, redes sociales y medios sociales. Se tiene autoridad cuando se tiene visibilidad y presencia en canales externos a los propios de la empresa.

Engagement (interacción/compromiso)

El *engagement* o interacción es una métrica muy empleada por la mayoría de las redes sociales, que mide el grado de interacción de los usuarios con un perfil

social, es decir, el número de interacciones con nuestro perfil o con una publicación en particular. Cada red social considera unas métricas u otras para cuantificar el número de interacciones. Un indicador importante es la ratio de *engagement*, que es un coeficiente que se calcula dividiendo el número total de interacciones entre el número de impresiones (o alcance).

Total de interacciones / Tasa de engagement= Alcance de la publicación

El engagement es el grado en el que un consumidor interactúa con una marca: es un compromiso entre la marca y los usuarios; es decir, es el compromiso y la forma en que interactúa un seguidor con una marca. Básicamente, se trata de un término que mide de cierta forma el grado de interacción entre una marca y los usuarios.

Los expertos en marketing señalan que el engagement se basa en crear cierta fidelidad y motivación para que nuestros usuarios defiendan y se sientan parte de la marca y, de esta manera, nos refieran nuevos usuarios. Tener usuarios comprometidos (engaged) con su marca es algo invaluable, ya que al pasar del tiempo se convertirán en "evangelizadores" de la marca. El éxito de su negocio depende de los usuarios recurrentes.

Las métricas utilizadas por redes sociales para medir la ratio o tasa de engagement dependen de que sean publicaciones, videos, historias, etcétera. Así, algunas de las métricas utilizadas por algunas redes sociales son:

11.10 HERRAMIENTAS DE ANALITICA SOCIAL

La Analítica Social utiliza un número muy variado de herramientas software que se agrupan en diferentes categorías globales: Analítica propiamente dicha, estadísticas, investigación, monitorización, influencia (relevancia), gestión de contenidos y reputación.

Twitter Analytics (analytics.twitter.com)

Herramienta gratuita de Twitter que permite analizar la cuenta de la organización, empresa o particular. Proporciona información sobre el impacto que han causado los tuits, retuits, tasa de interacción, seguidores más relevantes, ubicación de los seguidores, número de veces que se ha visto un tuit, menciones, favoritos, clics, etcétera. Uno de los datos más útiles de análisis en esta herramienta es ver el detalle de las impresiones de cada tuit. Al hacer clic en cualquier tuit individual de su red, se mostrará un desglose completo de cada elemento de interacción en dicho tuit, incluyendo clics en la URL, clics en su nombre de usuario, clics en las imágenes y un gráfico de barras para observar el engagement, interacción o compromiso durante las 24 horas recientes.

Para acceder a la aplicación de Analítica de Twitter, basta con iniciar la sesión e ir directamente a la aplicación o mediante su dirección web. Al entrar en la aplicación, directamente desde la versión de escritorio del perfil propio (personal o empresa), se presenta un menú horizontal con las opciones: **Inicio**, ***Tweets***, *Audiencias*, ***Card*** **de Twitter**, **Videos**, **Eventos** y **Herramientas**.

Facebooks Insights (www.facebook.com/insights)

Facebook incorpora también una aplicación propia de Analítica conocida como Facebook Insights (o "Estadísticas", en español), en la cual se pueden analizar los diferentes indicadores seleccionados por Facebook de modo que se pueda evaluar el rendimiento del sitio web. Se accede a la sección de estadísticas directamente desde la página de Facebook o desde el perfil del usuario. En función de las necesidades de la empresa. Es muy importante que los responsables de los medios sociales (*Community Manager, Social Media, Manager*, SEO) se enfoquen en el seguimiento de las estadísticas de modo periódico o extraordinario. En la versión en español, en la opción o pestaña **Visión General**, se pueden ver las diferentes pestañas con las métricas destacadas de los siete últimos días. Los indicadores más importantes son (Figura 11.3):

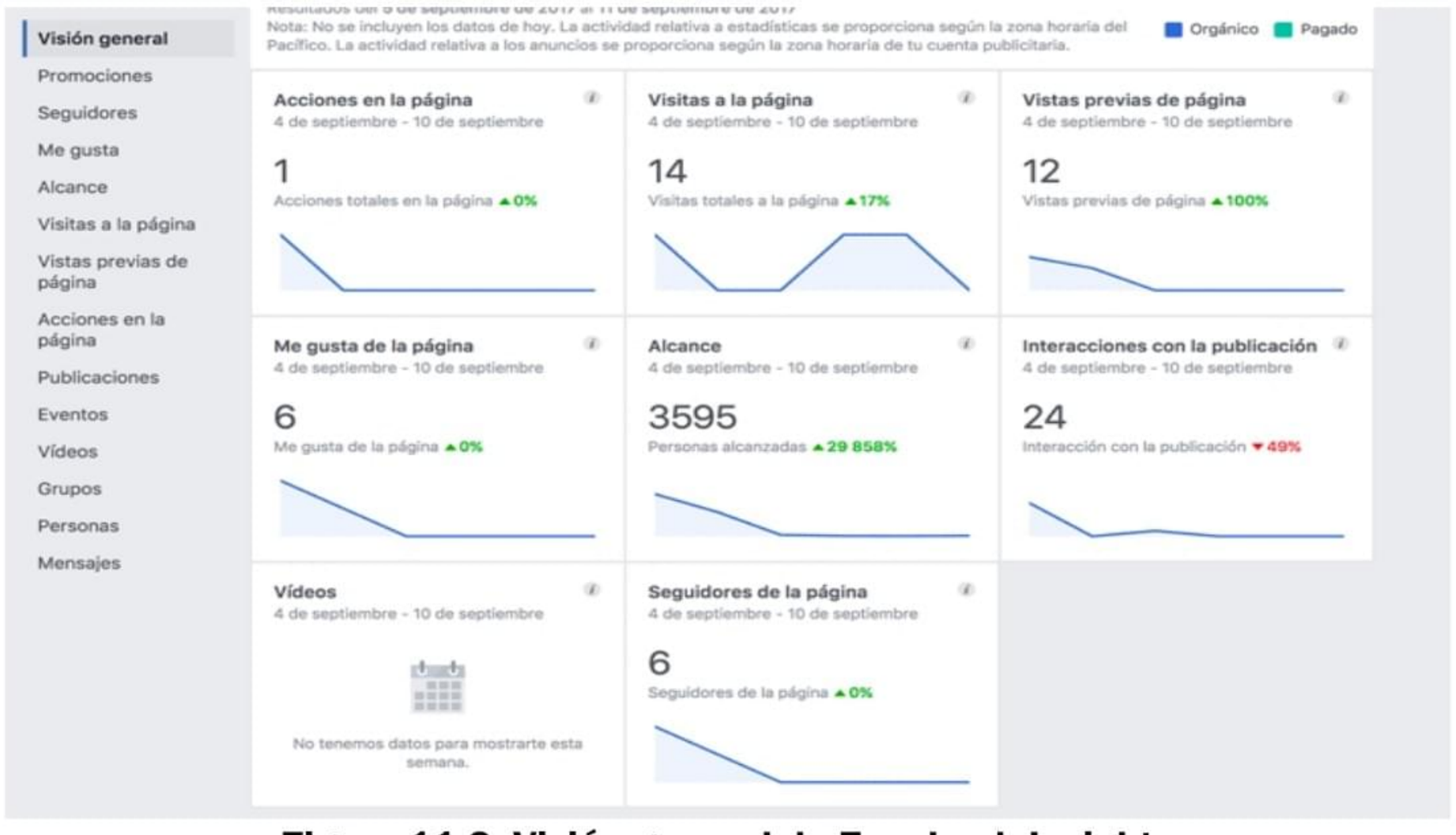


Figura 11.3. Visión general de Facebook Insights

- **"Me gusta".** Número de usuarios que siguen la página y los nuevos "Me gusta" recibidos en los últimos días, seguidores ganados y perdidos cada día, tipos de dispositivos desde el que se han realizado, etcétera.
- **Alcance de la publicación.** En esta sección se mide el número de usuarios a los que se ha alcanzado con nuestras publicaciones: el alcance total y el alcance de la publicación de la última semana junto con los porcentajes comparativos.

- **Visitas a la página.** Fuentes externas que traen tráfico a la página, así como las partes de la página que visitan.
- **Publicaciones.** Rendimiento de las publicaciones, alcance entre seguidores y no seguidores, alcance pagado u orgánico, número de seguidores que ha tenido conectados por día durante la última semana (acompañado de una gráfica que ofrece el número por horas).
- **Videos.** Ofrece un amplio análisis de los videos subidos a la plataforma: los videos más destacados en función de la duración de los mismos.
- **Personas.** Análisis del perfil demográfico de los seguidores y también de las personas a las que se alcanza con nuestras publicaciones o que interactúan con ellas.
- **Exportar.** Es una opción muy interesante, ya que permite la posibilidad de descargar todos los datos de la página en formato Excel o .CSV.

El *blog Publicidad en la nube*[5]tiene publicado un excelente tutorial y una infografía sobre la herramienta de Analítica de Facebook Insights.

Instagram Insights

Es la herramienta que Instagram ofrece una vez que se abre una cuenta empresarial en la aplicación y se maneja desde Facebook (Instagram es propiedad de Facebook), por lo que también se debe tener una página en esa plataforma. Instagram Insights ofrece los datos básicos de su cuenta, tales como crecimiento de seguidores, cantidad de comentarios e historias y las veces que un post se envía como mensaje directo, así como la interacción entre ellos. Además, le dice el número de cuentas empresariales que lo ven y la cantidad de veces que sus publicaciones son guardadas. En caso de publicidad de pago, se muestra el alcance del post publicitario desde el Facebook Ads Manager y comparará el rendimiento de ese post con las publicaciones orgánicas.

YouTube Analytics (www.youtube.con/analytics)

La aplicación de Analytics propia de YouTube permite al propietario de un canal estudiar al detalle el impacto de sus videos, descubrir el funcionamiento, fallos y otros indicadores de interés. Se pueden conocer el número de visitas, las visualizaciones y el tiempo de visualización de sus videos (una métrica indispensable), el dispositivo desde el que se ha visitado el video, cómo llegan los usuarios a sus videos (fuentes de tráfico), de dónde proceden (directamente o desde otras páginas de redes sociales). Otro indicador muy relevante es poder estudiar el perfil del espectador (edad, nacionalidad, profesión, aficiones, etcétera).

En lo relativo a los anuncios, en el caso de profesionales o empresas, es posible analizar el rendimiento de los anuncios, interacciones de la audiencia con los

videos (suscriptores conseguidos y perdidos, “me gusta”, “no me gusta”, favoritos, comentarios, comparticiones, etcétera.).

Conceptos básicos de YouTube Analytics[6]

En la página oficial de YouTube Analytics se pueden ver los conceptos básicos a los que se puede acceder:

Cómo establecer la configuración predeterminada de YouTube Analytics:

- Informe de visión general.
- Informe en tiempo real.
- Informes de ingresos.
- Informes de tiempo de visualización.
- Informes de interacciones.
- Informe de traducción.
- Grupos de YouTube Analytics.
- Listas de reproducción en YouTube Analytics.
- Datos de YouTube Red en YouTube Analytics.
- Cambios en los informes después de transferencias.

Una de las grandes funcionalidades de YouTube Analytics son las diferentes opciones gráficas[7] de visualización de los informes solicitados:

- Gráfico lineal.
- Gráfico multilíneas.
- Áreas apiladas.
- Gráfico circular.
- Gráficos de barras.
- Mapas interactivos.

LinkedIn

LinkedIn, al igual que otras redes sociales, ha desarrollado sus propias herramientas de Analítica para que sus usuarios puedan evaluar el rendimiento e impacto de sus acciones en la red social.

Los usuarios de perfiles básicos pueden acceder a datos limitados: quién ha visto su perfil, quiénes entre sus contactos son los más vistos, quiénes han interactuado con sus actualizaciones (“me gusta”, compartir, comentar, recomendar). Si se dispone de una cuenta *premium*, se puede obtener una

información más completa (número de visitas que recibe a la semana, número de apariciones de su perfil en búsquedas, palabras clave, sector profesional, origen demográfico de los accesos, etcétera).

En el caso de disponer de páginas de empresa de LinkedIn, se puede acceder a su propio recurso de Analítica y, en caso de ser el administrador de la página, se puede obtener información muy relevante sobre su rendimiento (interacción de los usuarios con las diferentes publicaciones, identificar tendencias, datos avanzados de demografía, conocer bien el tráfico y actividad de la empresa, etcétera).

El acceso a las estadísticas de la página de LinkedIn requiere el acceso a la página de empresa y seleccionar la opción "Análisis". Un panel de control multimedia muestra la página dividida en secciones: "Actualizaciones", "Seguidores" y "Visitantes".

Pinterest Analytics (analytics.pinterest.com)

Es una herramienta propia de Pinterest y a la que se puede acceder desde el perfil de la empresa. En caso de no disponer de una cuenta de empresa, se puede convertir una cuenta personal en una cuenta comercial de modo gratuito. Una vez que se dispone de la cuenta de empresa, se puede acceder a Pinterest Analytics desde el perfil propio haciendo clic en el menú (con forma de engranaje) y, a continuación, en Analytics.

El panel principal de Pinterest Analytics está dividido en tres categorías de análisis diferentes:

- Perfil de Pinterest. Impresiones, *repins*, clics, *all time*.
- Personas a las que llega (público). Analítica de las personas a las que se llega.
- Sitio web. Obtener análisis del sitio web, previas verificaciones.

11.11 HERRAMIENTAS DE MONITORIZACIÓN

La monitorización es la acción y efecto de monitorizar (definición del diccionario de la RAE). Aplicado al caso de los medios sociales, es una de las labores más complejas y el proceso de monitorización lleva consigo un importante seguimiento y control. Escuchar, monitorizar y obtener datos sobre las actitudes de los usuarios constituye la base del proceso de monitorización.

La investigación, monitorización y su posterior análisis son los pilares sobre los que se debe sustentar todo proyecto de social media. El seguimiento y la monitorización requieren de tecnologías y herramientas adecuadas para las plataformas de medios sociales.

La importancia de la monitorización reside en el análisis de los datos obtenidos, y sólo a través de este análisis se podrá crear conocimiento y tomar decisiones estratégicas. La monitorización del contenido de los medios sociales entraña obtener informes, cruzar datos y analizar estadísticas. Todas ellas son tareas difíciles, sobre todo por la gran variedad en la naturaleza de los datos.

Existen numerosas herramientas de monitorización de carácter global y específico para redes sociales concretas, que resultan adecuadas para administrar y obtener diferentes tipos de métricas de monitorización (internas, externas, de reputación, análisis de influencia, de alcance, de administración, etcétera) para las diferentes plataformas de medios sociales.

Audiense (audiense.com)

Herramienta de uso personal dirigida a la administración y monitorización de la red social Twitter. Es una herramienta de origen español, antes denominada SocialBro, que ayuda a conocer quiénes son sus seguidores. Los objetivos principales de Audiense, extraídos de su sitio web oficial, son:

- Descubrir los intereses de sus seguidores, su influencia, localización, idioma y mucho más.
- Averiguar cuándo es la mejor hora para tuitear y obtener el máximo resultado de cada tuit que se publica.
- Utilizar las herramientas de segmentación para encontrar empresas o personas a las que seguir.
- Localizar fácilmente usuarios de Twitter y crea listas con ellos.

Audiense sólo está disponible en versión web y se aplica en diversas áreas con múltiples funcionalidades, tales como segmentación y análisis de audiencias, monitorización de actividades y programación de publicaciones. Asimismo, cuenta con un gran número de filtros que se pueden aplicar en muchos indicadores: seguidores, cuentas seguidas, evolución de seguidores, seguidores comunes y análisis de la audiencia en métricas tales como localización (geolocalización), idioma, género, influencia, reputación, etcétera. Ofrece estadísticas detalladas, gráficos sobre las mejores horas para tuitear y difundir contenidos, buscar usuarios afines, enviar mensajes directos, crear listas o discriminar a los usuarios inactivos o los potenciales propagadores de spam (spammers).

Esta herramienta gratuita puede ser utilizada para uso personal por cualquier cuenta de Twitter con menos de 5.000 usuarios (sus seguidores más la gente a la que se sigue).

Addictomatic (addictomatic.com)

Es una herramienta gratuita que permite obtener una visión general de la reputación de una marca y de su sector.

Agora Pulse (agorapulse.com)

Es una herramienta de monitorización, publicación y programación de contenidos para redes sociales populares como Twitter, Facebook, LinkedIn e Instagram. Se pueden publicar contenidos y seguir el rastro de las publicaciones programadas.

Buffer (buffer.com)

Buffer es una herramienta que permite programar sus publicaciones en diferentes redes sociales (Twitter, Facebook, LinkedIn, Pinterest, Google+ e Instagram). Permite gestionar la carga de trabajo y planificar cuándo se desea publicar, sin necesidad de estar conectado en ese momento. Existe la posibilidad de descarga en el dispositivo móvil mediante las tiendas de apps de Android o de Apple.

Cyfe (cyfe.com)

Es una herramienta del tipo cuadro de mando (dashboard) que permite monitorizar todas las métricas que se requieren en una red social (Facebook, Twitter, LinkedIn, YouTube, etcétera) o página web. Se puede concentrar toda la información necesaria en una {única pantalla con diferentes presentaciones visuales.

Google Alerts (www.google.com/alerts)

Es una herramienta propia y gratuita de Google que se puede usar con una cuenta específica o sin ella, y que se utiliza para monitorizar la actividad que se produce en la Web en función de palabras clave. Google envía una notificación o alerta al correo electrónico del usuario donde se muestran los resultados de búsqueda que se han rastreado en la Web. Se pueden crear alertas para su marca (organización o empresa) y también para competidor o competidores, al estilo del funcionamiento del buscador de Google. Existen opciones relativas a idioma, región, cantidad, fuentes y frecuencia.

Las sugerencias de alertas en la versión en español (octubre 2017) de Google Alerts eran muy variadas: empresas, finanzas, música, películas, política, deportes, tecnología, televisión, salud, automóviles, programación y nuevas secciones. Dentro de cada una de ellas las opciones eran también muy variadas.

Hootsuite (hootsuite.com)

Es una de las herramientas más completas de monitorización de redes sociales para realizar análisis y seguimientos en ellas (Facebook, Twitter, Google+,

LinkedIn). Se utiliza para la gestión de perfiles con publicación de contenidos (rastrear menciones, personas que hablen de la marca, búsquedas relevantes, etcétera).

En esencia, es un gestor de contenidos, de planificación y de programación de publicaciones, estadísticas y análisis de la competencia. Ofrece un cuadro de mando personalizable que permite gestionar las principales redes sociales (perfiles y páginas) desde cualquier dispositivo (computadora, teléfono inteligente o tableta).

Es una herramienta dirigida a empresas y, por consiguiente, de pago, aunque ofrece tres versiones: gratuita, que permite administrar hasta cinco cuentas de redes sociales, Pro, con gestión de perfiles ilimitados e informes y otros servicios complementarios (cuota mensual), y Enterprise, versión más avanzada y profesional.

Mention (mention.com/es/)

Es una excelente herramienta de monitoreo de medios sociales. Diariamente, monitorea una gran cantidad de fuentes, que incluyen redes sociales, *blogs*, foros, entre otros, proporcionando ideas muy notables (*insights*) de negocio que ayudan en la toma de decisiones. Es una buena herramienta de Inteligencia de Negocios, ya que permite crear reportes automáticos que se pueden compartir con clientes, socios y empleados.

Social Mention (socialmention.com)

Herramienta gratuita de monitorización general muy potente. Permite monitorizar la información que se transmite por las redes sociales sobre su marca o perfil. Analiza casi un centenar de sitios web diversos, incluyendo redes sociales, para conocer el estado de la Web, cuáles son las palabras clave más relevantes, número de menciones que recibe, etcétera.

Social Report (socialreport.com)

Es una herramienta de pago que proporciona informes de rendimiento (por *e-mail*) de los perfiles de la marca en redes sociales. Realiza la monitorización de un gran número de medios sociales y gestiona además su presencia en línea. Permite realizar el seguimiento del rendimiento de plataformas sociales, incluyendo todo tipo de datos y de métricas específicas de cada una de las plataformas. Es una herramienta de gran potencia para la creación automatizada y personalizada de informes analíticos con una característica muy notable: permite la exportación de informes en formato Excel y PDF.

Sprout Social (sproutsocial.com)

Se trata de una herramienta web dirigida a empresas de mediano o gran tamaño, así como agencias y profesionales de los medios sociales tales como *Community Managers*, *Social Media Manager*, SEO, etcétera. Facilita la publicación (manual, automática, en tiempo real o periódica) en perfiles de las redes sociales más populares: Twitter, LinkedIn, Instagram y Facebook. Ofrece métricas cuantitativas y cualitativas de las principales plataformas con informes personalizados. Una de sus funcionalidades más notables es la posibilidad de monitorización y de exportación de datos.

TweekDeck (tweekdeck.twitter.com)

Es una herramienta gratuita de estadísticas de Twitter que ayuda a conocer la comunidad, el rendimiento, alcance y evolución de los tuits (cuáles son eficiente y cuáles no). Permite ver otros indicadores tales como número de veces que se ha viso el tuit, personas que han interactuado con él, ubicación de los seguidores, retuits recibidos, evolución por periodos de tiempo, etcétera; también permite exportar datos.

Radian6 (radian6.com)[8]

Según numerosos informes de consultoras y empresas especializadas, es una de las herramientas de pago más completa para análisis social de redes. Permite escuchar, medir e interactuar con los usuarios de la marca. El gran problema económico es su alto coste mensual, aunque la cantidad de funcionalidades que ofrece suele compensar dicho gasto.

Radian6 fue comprada por la empresa líder en soluciones de software de gestión empresarial en la nube, Salesforce.com; ha sido integrada con otras herramientas del proveedor y su nombre oficial es Social Studio (`www.salesforce.com/products/marketing-cloud/social-media-marketing/`). En la actualidad, es una herramienta de social media marketing de Salesforce, pensada en sus miles de clientes de soluciones de gestión empresarial (CRM, ERP, etcétera), aunque se puede utilizar como herramienta independiente. Por estas razones, Social Studio (Radian6) analiza conversaciones de infinidad de fuentes que permiten ver con facilidad, mediante cuadros de mando muy visuales, gran cantidad de información de lo que las empresas y la industria piensan y opinan de su marca.

SocialBro (es.socialbro.es)

En la actualidad su nombre es **Audiense**[9]y ya ha sido comentada anteriormente, pero dada la resonancia del dominio, simplemente la recordamos. Es una herramienta muy útil para gestionar la reputación de una marca en Twitter. Permite saber el perfil de sus seguidores (o de abandonos), identificar los

seguidores más influyentes y de interés y analizar el comportamiento de la competencia. Dispone de versión gratuita y de pago.

11.12 HERRAMIENTAS DE REPUTACIÓN E INFLUENCIA SOCIAL

Es muy importante tener información sobre aquello que dice la gente sobre la empresa, marca, producto o servicio para conocer la influencia social de la compañía, así como su reputación digital. En la actualidad, la presencia de una organización en las redes sociales no es suficiente, sino que se necesita ser influyente y, al mismo tiempo, ser bien considerada por las personas influyentes o influenciadoras (influencers). La influencia no se mide sólo en número de seguidores, ni en el número de actualizaciones que se realicen, sino que se mide en un número grande de indicadores que normalmente se combinan mediante algoritmos complejos de Inteligencia Artificial (aprendizaje automático y aprendizaje profundo), que señalan los sitios web más reputados especializados en la medición de la influencia y reputación social.

Klout (klout.com)

Es una de las aplicaciones más reconocidas para analizar la influencia del usuario y de una marca en los medios sociales, y con el tiempo se ha convertido casi en una métrica estándar (Klout Score es el índice de influencia Klout). Es una herramienta muy completa que muestra infinidad de resultados tales como influenciadores, influenciados, evolución de parámetros de medida y clasificación del usuario según su actividad y comportamiento. La puntuación (score) que Klout asigna a los perfiles monitorizados se estima a partir de más de 25 variables, y calcula un valor en una escala de 1 a 100 (índice Klout Score). Los valores altos del índice indican que el perfil es muy influyente en las redes analizadas. Los factores que componen el índice son alcance real, amplificación e impacto en la Red.

Klout es una herramienta imprescindible para medir la monitorización de la reputación y también es válida para localizar influyentes con los que interactuar en acciones de comunicación en la Web. Los parámetros de medición son muy diversos, desde la propia actividad, la repercusión (mide la cantidad de personas a las que llegan nuestros mensajes), la redifusión en nuestra comunidad y la reacción de quienes nos siguen a través de nuestra capacidad de generar interacciones (engagement).

Es una herramienta que permite su utilización en marketing, en publicidad y en recursos humanos. Comienza a ser muy frecuente que muchas empresas de estas disciplinas anteriores exijan un determinado nivel mínimo de influencia

Klout a quienes aspiran a conseguir un puesto de trabajo, un concurso o una oferta comercial.

Nació como una herramienta de Twitter, pero al día de hoy Klout soporta numerosas herramientas de medios sociales: Twitter, Facebook, LinkedIn, Google+, YouTube, Pinterest, Four Square, Blogger, WordPress, Tumblr, Instagram, etcétera.

Peerindex (peerindex.com)

Es una aplicación similar y competencia de Klout. Funciona de modo parecido, calculando un índice de influencia de 0 a 100, donde a mayor índice, mayor influencia. Se integra en Twitter, Facebook, LinkedIn, Quora, Blogger y Word press. Su índice de influencia se basa en autoridad, audiencia y actividad.

Kred (kred.com)

Es una herramienta que permite valorar la influencia (en una puntuación de 0 a 1000 que ejerce sobre los seguidores el contenido que se publica en los medios sociales. Su fortaleza reside en su algoritmo —denominado— algoritmo de Kred cuya característica principal es su transparencia a la hora de otorgar valores para obtener su puntuación final.

Twitalyzer (twitter.com/twitalyzer)

Es una herramienta gratuita propia de Twitter. Al igual que Klout, proporciona un índice de influencia, un número entre 0 y 100, para medir el impacto general que se tiene en Twitter durante los últimos treinta días. Twitalyzer ofrece, además, los resultados de las puntuaciones que se tienen en Klout y Peerindex. De este modo, se pueden comparar los resultados entre las tres herramientas y obtener una visión más completa de la influencia social de una marca o de un usuario individual.

Herramientas complementarias de redes sociales de impacto en la Inteligencia de Negocios

Como ya se ha comentado, las herramientas de análisis social son muy numerosas. Sin embargo, existen herramientas que, sin centrarse directamente en el análisis social, contribuyen considerablemente a la buena gestión de las líneas de negocio fundamentales de una empresa y son herramientas de Inteligencia de Negocios que ayudan en las diferentes capas de un sistema de IN, tanto como herramientas colaborativas, de trabajo en grupo, publicación de informes, etcétera. Aquí hacemos un resumen de herramientas de interés general para soluciones de Inteligencia de Negocios.

Asana (www2.asana.com/es)

Es una excelente plataforma on line para gestión multitarea en la empresa. Es ideal para trabajo en equipo de modo colaborativo y facilita la gestión de tareas conjuntas en tiempo real. Tiene versiones gratuitas y de pago.

Bitly (bitly.com)

Se trata de una herramienta gratuita para acortar enlaces URL y análisis de su rendimiento. Es muy idónea cuando se requiere gestionar direcciones de páginas web del negocio y cuya longitud es grande para el envío por correo electrónico, servicios de mensajería, etcétera. Su gran inconveniente es que la nueva dirección acortada, normalmente, no guarda ningún nexo de unión con la página web original, por lo que se corre el riesgo de una mala gestión de la privacidad; en consecuencia, tanto emisor como receptor deben ser conscientes de ello.

Canva (canva.com)

Es una herramienta muy eficiente para diseñar y publicar imágenes en las redes sociales. Es gratuita, pero con opción de pago con algunos recursos especiales. Es una sencilla e intuitiva herramienta de creación de infografías y dispone además de una amplia colección de plantillas para realizar presentaciones.

Evernote

Es una herramienta muy popular para la gestión y almacenamiento de contenidos. Es una excelente herramienta para utilizar en cualquiera de las etapas de un ciclo de vida de *Big Data*, de Inteligencia de Negocios. La gran ventaja es su asentamiento en el mercado con versiones gratuitas y de pago, y su integración con un gran número de herramientas

Feedly (feedly.com)

Es un lector RSS (versión gratuita) para suscripción a sitios web y *blogs*, de forma que se pueden recibir y leer publicaciones actualizadas sin necesidad de una visita diaria u ocasional a estos sitios web, uno por uno. Feedly es un lector de *feeds* que permite suscribirse a las noticias de múltiples sitios web desde un mismo lugar (otros lectores de RSS son: Netvibes y Digg).

Mitto

Se trata de una herramienta para administrar las diferentes contraseñas de los múltiples sitios, propios o externos, utilizados en la marca. Almacena en la nube de forma segura todas las contraseñas o *passwords*. Es gratuita y fácil de usar.

CASO DE ESTUDIO: GOOGLE ANALYTICS

Google Analytics[10] es una herramienta gratuita de Analítica Web de Google, alojada en la nube, y reconocida como una de las herramientas más eficientes y de referencia en el área de Análisis Web, que ayuda a la creación, actualización y mantenimiento de los sitios web de una empresa o de un usuario particular. Es también una herramienta idónea para conseguir mejorar el retorno de inversión de su negocio o empresa (ROI).

En marzo de 2005 Google compró la empresa Urchin —por aquellos días, una de las herramientas más acreditadas del mercado en Análisis Web—. Google hizo una adaptación para su buscador y herramientas asociadas, así como una nueva interfaz y, en noviembre de 2005, lanzó la nueva herramienta con el nombre de Google Analytics.

La herramienta Google Analytics es, a su vez, un conjunto de herramientas de medición de Analítica Web para la empresa, que proporciona las siguientes funcionalidades descritas en la página oficial de Google (`www.google.es/intl/es/analytics/`):

- **Estadísticas de varios canales**[11]. "Sus clientes investigan, comparan y toman decisiones de compra en distintos momentos y lugares. Los embudos multicanal de Google Analytics le permiten ver las interacciones entre distintos medios digitales y muestran cómo se combinan estos canales para crear ventas y conversiones. Lo más importante es que esta herramienta le ayuda a tomar decisiones de marketing fundamentales sobre las inversiones en publicidad en el nivel superior (los canales concretos) y en el inferior (por ejemplo, las palabras clave y las ubicaciones de anuncio específicas)".

- **Soluciones para dispositivos móviles**[12]. "Analítica de Móviles (analizar sus anuncios, aplicaciones y sitios para móviles; evaluar los sitios web y las aplicaciones para móviles, así como las visitas desde dispositivos móviles con conexión a Internet, incluidos las tabletas y los teléfonos, tanto de gama alta como básicos). Ofrece la mejor experiencia de usuario en diferentes pantallas y dispositivos".

- **Informes sociales.** "Ayudan a medir el impacto de las redes sociales en los objetivos de su empresa y en las conversiones relacionadas. Mide las métricas que más interesan a la empresa".[13]

Google Analytics proporciona informes estándar e informes personalizados sobre cómo los visitantes utilizan su sitio, cómo han llegado a él y qué puede hacer para que sigan visitándolo. Para la obtención de los informes, la herramienta de Google utiliza:

- Herramientas de análisis.

- Analítica de Contenido.
- Analítica de Redes Sociales.
- Analítica de Móviles.
- Analítica de Conversiones.
- Analítica de Publicidad

LA PLATAFORMA DE GOOGLE ANALYTICS

Google Analytics proporciona soluciones API para recopilar y configurar las interacciones del usuario con el contenido *on line*, así como para crear informes de ellas. Los desarrolladores y programadores interactúan e influyen en el procesamiento a través de una interfaz de usuario completa, bibliotecas de cliente y las mencionadas soluciones API.

La plataforma[14] de Google Analytics realiza todas las actividades de Analítica Web mediante cuatro etapas o componentes: recopilación de datos, configuración, procesamiento e informes.

Recopilación de datos

Recopila o recolecta los datos (*collection*) de las interacciones del usuario que se reciben desde el sitio web, aplicación móvil o desde cualquier sitio conectado a Internet que se considere.

En esta etapa, los datos se organizan y almacenan en bases de datos para que podamos acceder a ellos cuando sea necesario y de modo rápido. Google, en su sitio web, señala que esta etapa permite administrar cómo se procesan los datos.

La confidencialidad es un aspecto muy importante y prioritario para Google. Las funcionalidades de privacidad de Google Analytics más notables son:

- Sólo se puede acceder a datos personales contando con el permiso del propietario de la cuenta.
- Selección de la compartición de información con Google y con sus proveedores.
- No almacena ni informa o reporta ningún tipo de información personal.

Configuración

Permite administrar cómo se procesan los datos. La información pasa por un proceso denominado “agregación”, en el que se prepara para que pueda ser analizada. En esta etapa se organizan y almacenan los datos en tablas de bases de datos para poder acceder a ellos cuando sea necesario y del modo más rápido posible. Una vez que se procesa la información, se almacena en sus bases de

datos y ya los datos permanecen inalterables. La configuración de los servicios de Analytics requiere una cuenta de Google y ser propietario o administrador de un sitio web.

Una configuración adecuada permite obtener información real, atendiendo a criterios importantes tales como la fecha, la periodicidad, tipos de gráficos y dimensiones primarias/secundarias.

Procesamiento

La etapa de procesamiento permite procesar los datos de interacción del usuario con los datos de configuración para convertir la información recolectada en datos útiles y clasificados por categorías, con el objetivo de proporcionar informes que puedan permitir tomar decisiones eficientes.

Informes

Esta etapa permite acceder a todos los datos procesados en forma de informes o reportes. Los informes que ofrece Google Analytics pueden ser estándar y personalizados, relativos a audiencia, adquisición, comportamiento de usuarios y pueden ser informes de los usuarios que se encuentran activos en el sitio web o en la aplicación móvil. Se puede acceder a los reportes a través de la interfaz específica de Google Analytics o bien desarrollar su propia aplicación, utilizando las herramientas de API que ofrece Google Analytics.

Google Analytics genera informes acerca de la audiencia del sitio web que se pueden exportar a Excel o a PDF. Es posible habilitar el acceso a la cuenta corporativa o personal para aquellos usuarios con los que interese compartir las vistas e informes. También es posible realizar informes personalizados con las métricas e indicadores que más interesen para los objetivos de medición. Otras características notables de Google Analytics son el filtrado de datos (para reportar a un cliente en función de sus necesidades) y la segmentación (para obtener información muy precisa de los públicos objetivo). La segmentación de los datos globales en grupos permite analizar las tendencias y los comportamientos.

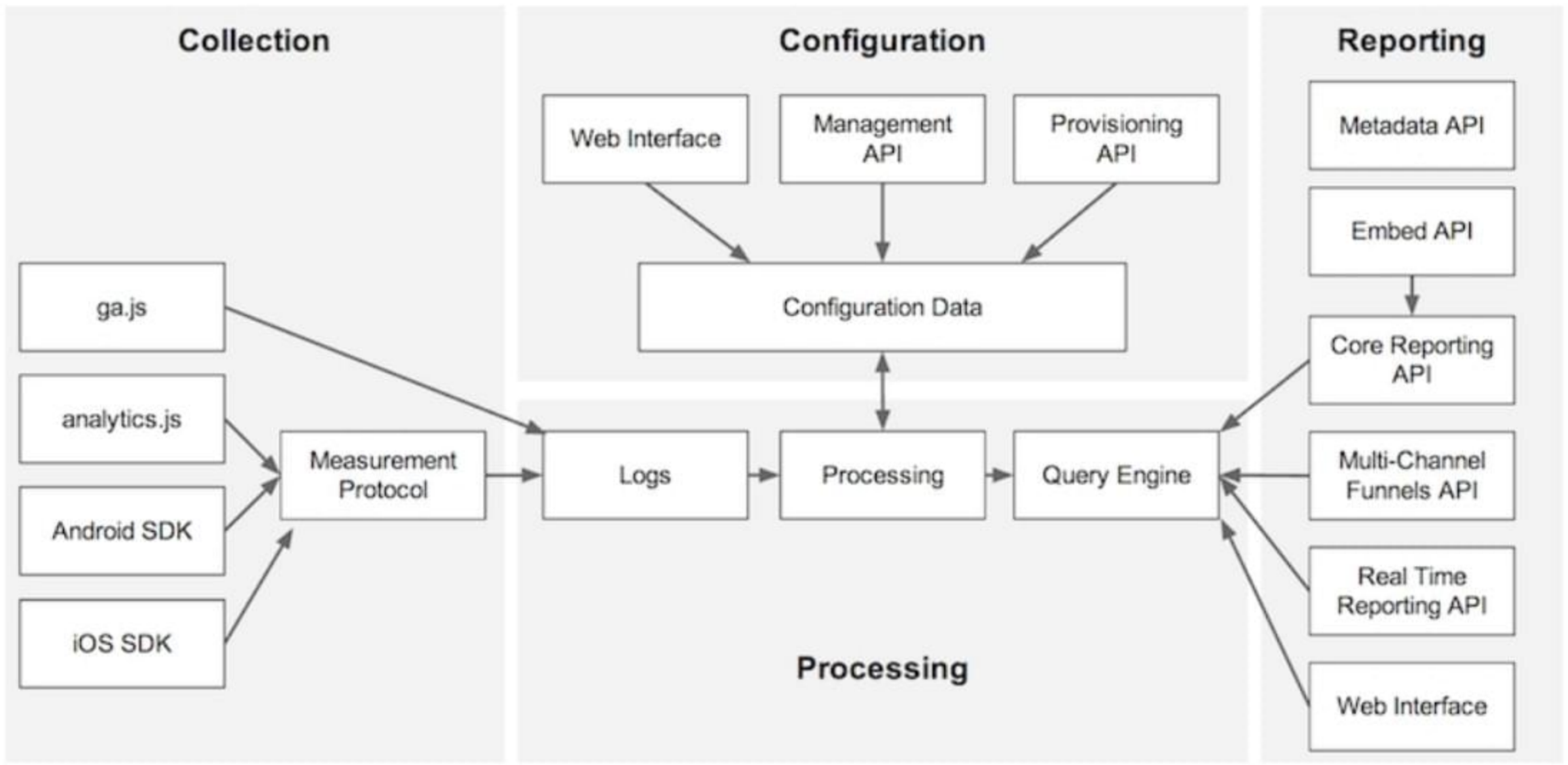


Figura 11.4. Componentes de la plataforma Google Analytics
Fuente:
`https://developers.google.com/analytics/devguides/platform/`

Google Analytics proporciona una colección de herramientas API para recopilar y configurar las interacciones del usuario con el contenido *on line*, así como para crear informes de ellas.

PAQUETE DE CONVERSIONES

El paquete de conversiones[15] de Google Analytics puede ir más allá de las páginas vistas y los recuentos de usuarios. Un ejemplo que publica Google en su página de inicio es una pauta de la aplicación del embudo de conversiones, sobre las tasas de éxito de las visitas al sitio web de la empresa (número de visitas que terminan en una compra o en el objetivo final del sitio web). Así, Google Analytics permite ver cómo funcionan todos sus canales de *marketing* digital en conjunto para atraer visitantes a su sitio; mide las ventas, las descargas, las reproducciones de video y otras acciones valiosas para usted. Asimismo, averigua las razones por las que algunos visitantes compran sus productos y otros no, para poder ajustar su sitio y sus programas de *marketing* a fin de alcanzar sus objetivos empresariales.



Figura 11.5. Modelo de embudo de conversión de Google

FORMACIÓN EN GOOGLE ANALYTICS

Otra de las grandes ventajas de la herramienta Google Analytics es la excelente oferta de formación, también gratuita. En el sitio web oficial de la herramienta[16], se ofrece a los usuarios un Centro de Ayuda de Google Analytics con artículos, documentación, videos, etcétera. Además, ofrece dos plataformas de formación: Analytics Academy y Analytics IQ, en las cuales se pueden seguir cursos gratuitos *on line* de Google Analytics y Análisis de Datos, así como preparación para un examen de certificación oficial muy reconocida en el mercado: Analytics Individual Qualification (IQ).

Google Analytics Móvil

Google Analytics anunció, a principios de julio de 2012, mejoras a su servicio de Analítica mediante una serie de reportes denominados Mobile App Analytics, diseñados para ayudar a los desarrolladores y distribuidores a medir indicadores y métricas de cómo emplean los clientes sus aplicaciones móviles. Al igual que sucede en el caso de las aplicaciones de escritorio, Google ofrece una de las mejores herramientas de Analítica Web Móvil8. Las cuatro funcionalidades más utilizadas son:

- Informe de dispositivos móviles:

— ¿Cuáles plataformas funcionan mejor?

— ¿Qué dispositivos usan los visitantes para encontrar el sitio web de la empresa?

Google Analytics muestra cuáles son los sistemas operativos móviles específicos y qué dispositivos móviles específicos envían visitantes a su sitio, aplicación para móviles o páginas de redes sociales. También, se señala la diferencia entre visitantes únicos absolutos y visitantes nuevos y recurrentes.

- Consultar la procedencia del tráfico de móviles:
- La visualización de estadísticas por ubicación permite conocer el origen actual del tráfico desde móviles, así como realizar predicciones acerca de dónde aumentará el tráfico.
- Medir el rendimiento de las aplicaciones para móviles. Se mide el uso de la aplicación como si se tratara de un sitio web. Los SDK de Google Analytics ofrecen una forma sencilla de medir el éxito de sus aplicaciones para móviles (especialmente en dispositivos IOS de Apple y Android).

11.13 RESUMEN

Según Avinash Kaushik, Analítica Web es "el análisis de datos cuantitativos y cualitativos de su sitio web y de la competencia, para impulsar una mejora continua de la experiencia on line que tienen tanto los clientes habituales como los potenciales y que se traduce en unos resultados esperados (on line y off line)". (Kaushik 2011:24)

- Se pueden considerar cuatro modelos en Analítica Web: Analítica Web tradicional, Analítica Social, Analítica Móvil y Analítica de Big Data.
- Una métrica es una valoración cuantitativa de estadísticas, que describen tanto los eventos como las tendencias de un determinado sitio web; en realidad, es una medida cuantitativa que permite conocer el estado de un sitio web, de una página web o un proceso que se realiza en un sitio para un atributo o parámetro determinado.
- Métricas cruciales de Kaushik: visitas, visitantes, visitantes únicos, tiempo en la página y en el sitio, tasa de rebote, tasa de salida, tasa de conversión y compromiso.
- Otras métricas por considerar y muy importantes: visitas provenientes de buscadores o directas, ranking de páginas más vistas/páginas por visita y procedencias de las visitas.

- Un indicador clave de rendimiento (KPI) es una métrica que ayuda del modo más eficiente posible a conseguir los objetivos previstos del sitio web.
- Todas los KPI son métricas, pero no todas las métricas son KPI.
- Analítica Web para Móviles se refiere al campo específico de la Analítica Web en el canal móvil de la organización, y comprende el conjunto de prácticas y tecnologías para coleccionar y analizar los datos de presencia en Internet desde móviles y tabletas, con el objetivo de tomar decisiones de negocio.
- La información de las herramientas de Analítica Web debe proporcionar resultados similares a las métricas ordinarias y, además, información relativa a la actividad de los usuarios, al dispositivo y a la aplicación.
- Analítica social es la parte de la Analítica que realiza el análisis de los datos sociales procedentes, a su vez, de medios sociales. Existe una abundancia de datos sociales que exige un tratamiento específico y unas herramientas especiales para este tratamiento.
- Analítica social se puede considerar la disciplina que ayuda a las empresas a analizar, calcular y explicar el rendimiento de las iniciativas de medios sociales. Es el proceso de medir, analizar e interpretar los datos sociales que se presentan a través de los diferentes canales, medios de comunicación y dispositivos.
- Las métricas sociales son medidas cuantitativas que permiten conocer el estado de un sitio web (medio social), de un componente de un sitio web (medio social) o del proceso que se realiza en un sitio web para un atributo o parámetro determinado (Acera, 2012). Las redes sociales obligan a la creación de nuevas métricas distintas de las tradicionales. Los KPI sociales miden los resultados más importantes y de impacto en los medios sociales. Herramientas populares son Twitter Analytics, Facebook Insights o Instagram Insights.
- Las herramientas de monitorización son uno de los elementos imprescindibles que se han de utilizar en los sitios web. Algunas herramientas muy utilizadas son Google Alerts, Audiense, Social Mention y Hootsuite.
- Las herramientas de reputación e influencia social son también de gran aplicación para medir el impacto de la marca comercial u organización. Herramientas muy utilizadas son Klout, Peerindex o Twitalyzer.

NOTAS

[1] *Avinash Kaushik. Analítica Web 2.0. 2ª Ed. Barcelona:Gestión 2000, 2011 pág. 20.*

[2] *Gorostiza y Barainca (2016), en su excelente libro Google Analytics: Mide y vencerás, hacen una historia de la evolución de la Analítica Web bastante completa y acompañada de un buen número de recursos.*

[3]*Iñaki Gorostiza y Asier Barainka. Google Analytics. Mide y vencerás. Madrid: Anaya multimedia. 2016*

[4] *Recordemos que un sitio web (web site o site) es el componente web de un dominio y un sitio web se compone o tiene varias páginas web (page o web page). El sitio web tiene la dirección URL de la organización y cada página tiene su propia dirección URL que se deriva de la página del sitio.*

[5]http://www.publicidadenlanube.es/facebook-insights-tutorial-espanol/#8-_Personas

[6]https://support.google.com/youtube/answer/1714323?hl=es

[7]https://support.google.com/youtube/answer/1714323?hl=es

[8]*Al entrar en radian6.com nos redirige al portal de Salesforce Marketing Cloud*
https://www.salesforce.com/products/marketing-cloud/social-media-marketing/?d=radian6&internal=true

[9]*Excelente artículo del periódico económico Expansión, donde se analizan las características principales de Audiense: Elena Arrieta. "SocialBro arranca una nueva etapa bajo el nombre de Audiense", 10/3/2016*

:http://www.expansion.com/economia-digital/companias/2016/03/10/56e1c25e46163fc7528b45b7.html

[10]https://www.google.es/intl/es/analytics

[11]https://www.google.es/intl/es/analytics/features/multichannel-funnels.html

[12]https://www.google.es/intl/es/analytics/features/mobile.html

[13]www.google.es/intl/es/analytics/features/social.html

[14]https://developers.google.com/analytics/devguides/platform/

[15]https://www.google.es/intl/es_ALL/analytics/features/conversion-suite.html

[16]https://analytics.google.com/analytics/academy/

CAPÍTULO 12
GESTIÓN DEL CONOCIMIENTO Y HERRAMIENTAS COLABORATIVAS

CONTENIDO

12.1 Introducción
12.2 Definición de conocimiento: un enfoque de gestión y toma de decisiones
12.3 ¿Qué es Gestión del Conocimiento?
12.4 Sistemas de Gestión del Conocimiento (SGC)
12.5 Tecnologías de descubrimiento y recuperación de información
12.6 Tecnologías de colaboración
12.7 Tecnologías de almacenamiento
12.8 Trabajo en grupo (*groupwork* y *groupware*)
12.9 Herramientas colaborativas
12.10 Integración de los SGC en sistemas de información
12.11 Soluciones y plataformas colaborativas y de comunicaciones unificadas
12.12 RESUMEN

OBJETIVOS

- Definir el conocimiento, los diferentes tipos de conocimiento y el concepto de capital intelectual.
- Definir el concepto de Gestión del Conocimiento, Sistema de Gestión del Conocimiento y cuáles son sus ventajas, además de describir el ciclo de vida de Gestión del Conocimiento.
- Describir los roles, procesos y tecnologías implicadas en los Sistemas de Gestión del Conocimiento.
- Conocer el concepto de trabajo en grupo (*groupwork* y *groupware*) y de flujo de trabajo (workflow).
- Describir el proceso de toma de decisiones en la empresa.
- Identificar las tecnologías y herramientas colaborativas de tipo social.
- Conocer plataformas colaborativas de gran uso en organizaciones y empresas.

Disponible online

PARTE III
ANALÍTICA DE NEGOCIOS Y ANALÍTICA DE DATOS

CAPÍTULO 13

INTELIGENCIA ARTIFICIAL APLICADA Y ALGORITMOS EN INTELIGENCIA DE NEGOCIOS

CONTENIDO

13.1 Introducción
13.2 Inteligencia Artificial: Definición, historia y evolución
13.3 Tecnologías de Inteligencia Artificial
13.4 Aprendizaje automático
13.5 Aprendizaje profundo (Deep Learning)
13.6 Computación cognitiva
13.7 Bots y chatbots
13.8 Chatbots de empresa: el caso de la atención al cliente
13.9 El algoritmo en inteligencia artificial como modelo de negocio en la economía digital
13.10 RESUMEN

OBJETIVOS

- Identificar y conocer los conceptos fundamentales de inteligencia artificial: su origen, historia y evolución.
- Conocer las tecnologías de soporte de la inteligencia artificial moderna y aplicada.
- Conocer las técnicas de aprendizaje automático y aprendizaje profundo así como su aplicación en los negocios.
- Aprender los conceptos básicos de computación cognitiva.
- Conocer los fundamentos y utilización de los asistentes virtuales de voz (bots y chatbots) en inteligencia de negocios.
- Conocer los conceptos básicos de algoritmos y su aplicación e impacto en los modelos de negocio de la economía digital.
- Identificar y conocer aplicaciones de algoritmos en inteligencia de negocios.

13.1 INTRODUCCIÓN

Desde el advenimiento de *Big Data,* la Inteligencia Artificial está llegando a numerosos sectores que, hasta hace unos años, era prácticamente impredecible que llegara y que, en la actualidad, están impactando en los sistemas de inteligencia de negocios de las organizaciones y empresas.

El *Deep Learning* (aprendizaje profundo) es el campo de mayor crecimiento en la Inteligencia Artificial. Ayuda a las computadoras a dar sentido a ingentes cantidades de datos en forma de imágenes, sonido y texto. Mediante el uso de distintos niveles de redes neuronales, las computadoras pueden ver, aprender y reaccionar ante situaciones complejas, igual o, incluso, mejor que los humanos. Esta tendencia está transformando la manera de considerar los datos, la tecnología y los productos y servicios.

Los *bots* y los *chatbots* (*bots* conversacionales) son una de las aplicaciones más populares y de mayor futuro en la industria 4.0 y, por ende, en organizaciones y empresas. Gracias a ellos, la Inteligencia Artificial aplicada está llegando a infinidad de sectores en las empresas y en las industrias. Ya en la Feria Mundial de la Electrónica CES 2017 (Las Vegas), celebrada en enero de 2017, los *bots* fueron las aplicaciones más populares, dado que han llegado no sólo incorporadas a teléfonos inteligentes, sino a una gran cantidad de dispositivos electrónicos. Alexa de Amazon fue la revelación y en los años 2017 y 2018 se ha convertido en uno de los *bots* (asistentes virtuales y de voz) más populares y comercializados.

Unas nuevas generaciones de plataformas de negocio están surgiendo en la convergencia del **aprendizaje automático** (*machine learning*), recientemente el

aprendizaje profundo (*Deep Learning*) y ***Big Data***, que generarán un gran cambio en todo el sector de Inteligencia de Negocios y de Analítica.

13.2 INTELIGENCIA ARTIFICIAL: DEFINICIÓN, HISTORIA Y EVOLUCIÓN

La Inteligencia Artificial (IA) es una disciplina científica y de ingeniería que ha alcanzado una gran resonancia en los últimos años, y cuyas tecnologías y herramientas tienen impacto real en un gran número de sectores de la industria, los negocios, la educación, la salud, etcétera. La Inteligencia Artificial apareció como disciplina en la década de los 50. Yoshua Bengio (2016)[1], uno de los padres del aprendizaje profundo (la rama de mayor futuro en la Inteligencia Artificial), destaca que en aquel entonces los investigadores creyeron que podrían emular la inteligencia humana en menos tiempo de lo que duraría su carrera. Las esperanzas se desvanecieron cuando quedó claro que los algoritmos y la potencia de cómputo no bastaban para culminar la tarea; sin embargo, en los últimos años el desarrollo de nuevas técnicas computacionales, inspiradas en las redes de neuronas del cerebro humano, ha resucitado la esperanza de materializar las promesas originales de la IA. Bengio plantea que el aprendizaje profundo, una técnica que se vale de redes neuronales complejas, permite que una máquina aprenda conceptos abstractos. En algunas tareas ya se aproxima a lo que pueden lograr los seres humanos.

¿Cuál es la definición de Inteligencia Artificial? Como sucede con las grandes disciplinas de impacto global, existen numerosas definiciones de IA, por lo que comenzaremos describiendo la definición dada por la RAE en su diccionario, donde aparece como acepción de la palabra inteligencia en el campo de la informática: “Disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico”. Otra definición muy reciente, ya más técnica y de la mano de un gran científico español especialista en IA, Ramón López de Mántaras (2017), considera que la Inteligencia Artificial tiene como objetivo diseñar algoritmos que, una vez programados, doten de comportamiento inteligente a las máquinas. John McCarthy acuña el término “Inteligencia Artificial” en una famosa conferencia en el Dartmouth College, la primera conferencia dedicada a la IA[2]. En un documento publicado en 2007 por la Universidad de Stanford, donde John McCarty —uno de los padres de la Inteligencia Artificial con su famosa presentación de 1956— enseñó desde 1962 hasta su jubilación en 2000, explicó que él entendía por Inteligencia Artificial (IA) "la ciencia y la ingeniería de crear máquinas inteligentes, especialmente programas de computación inteligentes. Está relacionada con la tarea similar de utilizar ordenadores para comprender la inteligencia humana, pero la IA no se limita a métodos que sean observables biológicamente".

Asimismo, **Alan Turing es considerado el padre de la Inteligencia Artificial.** En 1936, este visionario diseñó una máquina capaz de implementar cualquier cálculo que hubiera sido formalmente definido, pilar esencial para que un dispositivo pueda adaptarse a distintos escenarios y "razonamientos". En 1950, Turing publicó su famoso artículo *Computing Machinery and Intelligence*[3], donde propuso su famoso ***Test* de Turing**, uno de los grandes soportes de la actual Inteligencia Artificial. En el año 1956, como ya se ha comentado, John McCarty y otros científicos (Marvin Minsky y Claude Shannon) acuñan el término de Inteligencia Artificial.

Hasta la década de los noventa y primeros años del siglo XXI, la Inteligencia Artificial estuvo reducida a los laboratorios de investigación y al campo de la educación universitaria y de posgrado.

La consagración definitiva de la Inteligencia Artificial llegó en 1997, cuando IBM demostró que una computadora dotada de Inteligencia Artificial era capaz de vencer en el juego de ajedrez a un humano, el campeón del mundo **Gari Kaspárov.** Se llamaba **Deep Blue** y sirvió de base para que la industria tecnológica y la sociedad en general cobrara conciencia de la relevancia y las posibilidades de la IA.

En el año 2000, se presenta el robot **ASIMO**, fabricado por Honda, un robot humanoide totalmente operativo.

La llegada de Watson

Si hay un ejemplo de Inteligencia Artificial por defecto, ese es IBM Watson. Un sistema que hizo su aparición estelar, al estilo de Deep Blue, ganando una competición de alto nivel, aunque en este caso más compleja que en 1997. En 2011, Watson ganó el popular concurso televisivo Jeopardy! frente a los dos máximos campeones de este programa, en el que se realizan preguntas sobre cultura y conocimiento de todo tipo. En primer lugar, Watson tuvo que ser capaz de entender las preguntas y las respuestas que daba, para lo cual ayudaron sus 200 millones de páginas de contenido almacenadas en su sistema. También tuvo que realizar jugadas inteligentes, a la hora de sopesar la elección de las categorías, y tuvo que apostar una cantidad en la ronda final.

Desde entonces, IBM Watson se ha convertido en el estandarte de los sistemas cognitivos, procesamiento de lenguajes naturales y el razonamiento y el aprendizaje automático. Esta tecnología se está utilizando actualmente para ayudar en los tratamientos contra el cáncer, el comercio electrónico, la lucha contra el cibercrimen o la banca internacional.

En octubre del año 2011 Apple presentó Siri —su asistente de voz inteligente— integrado a su teléfono inteligente iPhone 4S y que se distinguía por responder a preguntas y realizar tareas sin utilizar el teclado.

Descargado en: eybooks.com

En 2012 y años sucesivos se fueron presentando los asistentes de voz Google Now de Google, Cortana de Microsoft, Alexa de Amazon, Bixby de Samgsung, etc.

Además, Google ha logrado seguir los pasos de IBM y conquistar un juego tradicionalmente de humanos: **Go**. Se trata de un juego oriental tan antiguo como complejo; se dice que su tablero, de 19×19 cuadrados, permite más posiciones durante una partida que átomos hay en el universo.

Google usa el aprendizaje automático para completar las expresiones de búsqueda y, a menudo, logra predecir acertadamente lo que busca el usuario. Facebook y Amazon usan algoritmos predictivos para ofrecer al usuario recomendaciones basadas en su historial de lecturas o compras. La IA es el componente central de los autos sin conductor (que ya son capaces de evitar choques y atascos de tráfico) y de sistemas de juego, tales como AlphaGo de Google DeepMind, la computadora que en marzo de 2016 derrotó al maestro de Go, el surcoreano Lee Sedol, en un torneo a cinco partidos.

Han pasado veinte años desde que Google aplicó el algoritmo *Page Rank*, que intentaba averiguar lo que cada usuario estaba buscando. Google, el rey de la innovación, no inventó *machine learning*, pero sí sentó las bases de su uso masivo. *Machine learning* es el sustento de los sistemas de recomendaciones (Amazon, eBay, LinkedIn y Twitter), de detección de fraude, de reconocimiento de voz, de algoritmos para la predicción de enfermedades, delitos, averías en máquinas o tendencias de consumo.

Dada la amplitud de aplicaciones de la IA, todas las empresas necesitan imperiosamente integrarla en sus productos y servicios; de lo contrario, no podrán competir con otras que usen redes de recolección de datos para mejorar las experiencias de los clientes y guiar las decisiones empresariales. La próxima generación de consumidores crecerá con tecnologías digitales de IA a su alrededor, y esperará que las empresas se anticipen a sus necesidades y entreguen respuestas instantáneas y personalizadas a cada consulta[4].

Hasta ahora, la IA ha sido demasiado cara o compleja como para permitir un uso óptimo en la mayoría de las empresas. Su integración con las operaciones habituales puede ser difícil y, generalmente, demanda emplear a expertos en inteligencia artificial y en Ciencia de Datos. Por eso, muchas empresas siguen tomando decisiones importantes guiadas por el instinto en vez de por la información.

En enero de 2017, Libratus, un programa realizado por la Universidad Carnegie Mellon, vence en una partida de póker a cuatro expertos jugadores.

Por último, también en 2017, el Parlamento Europeo debate sobre la necesidad de considerar el estatus de los robots y si ha de ser similar al estatus de los residentes europeos.

Esta situación cambiará en los próximos años, conforme el uso de la IA se extienda, con el potencial de hacer a cada empresa y a cada empleado más inteligentes, veloces y productivos. Los algoritmos de aprendizaje automático pueden analizar miles de millones de señales para redirigir llamadas de clientes al agente más adecuado o identificar clientes interesados en la compra de un producto.

Tras muchos despegues fallidos, la Inteligencia Artificial ha progresado espectacularmente durante los últimos años, gracias en gran parte a la versatilidad de la técnica denominada *Deep Learning* (aprendizaje profundo). Partiendo de un volumen de datos lo suficientemente grande como para poder extraer información valiosa, redes neuronales artificiales profundas, basadas en el funcionamiento del pensamiento humano, son susceptibles de ser enseñadas a hacer todo tipo de cosas.

El futuro de la Inteligencia Artificial: el aprendizaje profundo

El aprendizaje profundo o *Deep Learning* parte de un volumen de datos lo suficientemente grande como para poder extraer información valiosa mediante redes neuronales artificiales profundas, basadas en el funcionamiento del pensamiento humano, y que, como ya se dijo, son susceptibles de ser enseñadas a hacer todo tipo de cosas.

Los avances en aprendizaje profundo —una rama de la IA que toma como modelo la red neuronal (o neural) del cerebro— tal vez permitan que asistentes digitales inteligentes ayuden a planear vacaciones tan bien como un asistente humano, o que las empresas evalúen las actitudes de los consumidores respecto de una marca según millones de señales originadas en redes sociales y otras fuentes de datos. En el ámbito de la salud, los algoritmos de aprendizaje profundo pueden ayudar a los médicos a identificar células cancerosas o anomalías intracraneales desde cualquier lugar del mundo en tiempo real.

Noticia de impacto en inteligencia artificial

En 2017 se produjo una gran noticia para la Inteligencia Artificial y el aprendizaje profundo: el premio Fundación BBVA Fronteras del Conocimiento[5], en la categoría de Tecnologías de la Información y la Comunicación (TIC) —el premio español más reconocido en el mundo tecnológico y uno de los más reconocidos a nivel internacional—, fue concedido en su novena edición (enero 2017) al investigador en Inteligencia Artificial Geoffrey Hinton “por su trabajo pionero y profundamente influyente” a la hora de lograr que las máquinas sean capaces de aprender, según señala el acta del jurado. El científico galardonado “se ha inspirado en cómo funciona el cerebro humano y en cómo ese conocimiento puede ser aplicado para dotar a las máquinas de la capacidad para desempeñar tareas complejas como lo hacen los humanos”. Hinton está considerado el mayor experto mundial en aprendizaje profundo.

13.3 TECNOLOGÍAS DE INTELIGENCIA ARTIFICIAL

El mercado de tecnologías de Inteligencia Artificial (IA), como ya venimos comentando, está en pleno auge. Las empresas han comenzado a invertir en Inteligencia Artificial —desde las grandes de Internet, tales como Google, Amazon, Facebook, hasta las tradicionales grandes de informática, tales como IBM, Oracle, HP, SAP, etcétera, pasando por las pequeñas y medianas empresas—. La Inteligencia de Negocio como disciplina global y transversal requiere la incorporación de las tecnologías más relevantes en su arquitectura.

Un estudio realizado por la consultora Forrester Research, publicado a principios de 2017, calculaba un incremento de más del 300% en la inversión que se realizaría en Inteligencia Artificial en 2017, comparado con 2016. Gartner predice que los negocios de inteligencia artificial en 2018 se proyectan a un total de 1,2 billones de dólares (1,2 trillones: 1.200.000.000.000 de dólares, 1€ aprox. 1,4$), un incremento del 70 por ciento respecto a 2017 y se prevé alcanzar 3,9 billones (3,9 trillones) en 2.022
(`https://www.gartner.com/newsroom/id/3872933`)

El mencionado estudio de Forrester, TechRadar[6], analizó y evaluó cuáles son las principales tecnologías de Inteligencia Artificial que dominarían el año 2017. Gil Press, en Forbes[7], ha hecho una síntesis del estudio, donde se recogen y comentan las 10 tecnologías seleccionadas por Forrester como tecnologías de impacto de la Inteligencia Artificial en los negocios. Gil Press comienza describiendo el nacimiento de la IA, en 1955, como una subdisciplina de la informática. Hoy en día, la IA incluye una variedad de tecnologías y herramientas, algunas ya contrastadas y consolidadas, otras relativamente nuevas. Las tecnologías que Forrester considera que deben adoptar las empresas para soporte en la toma de decisiones son:

1. Generación de lenguaje natural

Es un subcampo de la Inteligencia Artificial que consiste en **crear texto a partir de datos obtenidos**. Esto permite que las computadoras puedan comunicar ideas con gran precisión y exactitud. Se utiliza actualmente en servicio al cliente, generación de reportes y en el resumen de hallazgos (*insights*) de Inteligencia de Negocios.

Algunos proveedores de este servicio son Attivio, Automated Insights, Cambridge Semantics, Digital Reasoning, Lucidworks, Narrative Science, SAS y Yseop.

2. Reconocimiento de voz

Siri no es el único agente que te entiende. Cada vez más sistemas incorporan la **transcripción y transformación del lenguaje humano a formatos útiles para las aplicaciones de computadoras**, muy utilizados en la actualidad en sistemas interactivos de reconocimiento de voz y aplicaciones móviles.

Algunos ejemplos de proveedores son NICE, Nuance Communications, OpenText y Verint Systems.

3. *Hardware* optimizado con IA

El *hardware* tiene que comenzar a ser más amigable con las tecnologías de Inteligencia Artificial, y ello entraña creación de unidades procesadoras de gráficos (GPU) y dispositivos específicamente diseñados y estructurados para ejecutar tareas orientadas a la IA. Los vendedores seleccionados de estas tecnologías son Alluviate, Cray, Google, IBM, Intel y Nvidia.

4. Agentes virtuales

Según Forrester, es la tecnología de IA de mayor impacto. Un agente virtual es una **computadora o programa capaz de interactuar con humanos**. El ejemplo más común de esta tecnología son los *chatbots*. Actualmente, se utiliza en servicio y atención al cliente y para la administración de las *smart homes* (casas inteligentes). Los proveedores destacados son Amazon, Apple, Artificial Solutions, Assist AI, Creative Virtual, Google, IBM, IPsoft, Microsoft y Satisfi.

5. Plataformas *machine learning*

El aprendizaje automático o aprendizaje máquina (*machine learning*) es una disciplina de las ciencias de la computación y una rama de la Inteligencia Artificial, cuyo objetivo es **desarrollar técnicas que permitan a las computadoras aprender**. El desarrollo de algoritmos, programación de API (interfaces de programación de aplicaciones), herramientas de desarrollo y *training* (entrenamiento), así como el procesamiento y análisis en tiempo real de *Big Data* y muchas otras aplicaciones en numerosas máquinas, han dado una gran relevancia a las plataformas de aprendizaje automático.

Algunos proveedores son Amazon, Fractal Analytics, Google, H2O.ai, Microsoft, SAS, Skytree y ADEXT.

6. Manejo de decisiones

Las máquinas inteligentes ya proporcionan **normas y lógica** a los sistemas de Inteligencia Artificial y se utilizan para la instalación inicial, mantenimiento y despliegue de sistemas y operaciones empresariales. Es utilizada en una amplia variedad de aplicaciones empresariales, **asistiendo o ejecutando la toma de decisiones automatizadas.** Los proveedores más importantes son Advanced Systems Concepts, Informatica, Maana, Pegasystems y UiPath.

7. Plataformas de aprendizaje profundo

Se trata de tipos especiales de *machine learning*, que consisten en utilizar redes neuronales artificiales con múltiples capas de abstracción. Esta tecnología imita las funciones del cerebro humano para procesar datos y crea patrones que utilizará en la toma de decisiones, como también ocurre

en el ejemplo antes mencionado. Actualmente, se utiliza principalmente para reconocer patrones y clasificar aplicaciones únicamente compatibles con conjuntos de datos a gran escala. Proveedores de servicios de plataformas de aprendizaje profundo son: Deep Instinct, Ersatz Labs, Fluid AI, MathWorks, Peltarion, Saffron Technology y Sentient Technologies.

8. Biométricas

Las tecnologías biométricas (**biometría**) contemplan la identificación, medición y análisis de las características físicas (forma o composición del cuerpo) y de comportamiento de las personas. Esto permite interacciones más naturales entre humanos y máquinas, incluyendo —pero no limitado a— reconocimiento de tacto, imagen, habla y lenguaje corporal. Se utiliza mucho en investigación de mercados. Los proveedores son 3VR, Affectiva, Agnitio, FaceFirst, Sensory, Synqera y Tahzoo.

9. Automatización de procesos robóticos

Esta automatización se da al **integrar *scripts* y otros métodos para imitar y automatizar tareas humanas** que apoyen en los procesos empresariales. Actualmente, se utilizan en casos donde resulta muy costoso o ineficiente que los humanos ejecuten una determinada tarea o proceso. Eso sí, recordemos que **la Inteligencia Artificial no viene a reemplazar la capacidad humana**, sino a complementarla y a potencializar todo el talento que tienen las personas. La automatización de procesos, por ejemplo en *marketing* y ventas, es una de sus grandes aplicaciones. Los proveedores destacados son Advanced Systems Concepts, Automation Anywhere, Blue Prism, UiPath, WorkFusion.

10. Analíticas de texto y NLP (Procesamiento de Lenguaje Natural)

El Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés) **utiliza analíticas de texto para facilitar el entendimiento estructural de los enunciados**, así como su **significado, entonación y lo que pretenden decir a través de métodos estadísticos** y de *machine learning.* Actualmente, se utiliza en sistemas de seguridad y detección de fraudes y en una amplia gama de asistentes automatizados y aplicaciones para extraer datos no estructurados. Algunos proveedores de estas tecnologías son: Basis Technology, Coveo, Expert System, Indico, Knime, Lexalytics, Linguamatics, Mindbreeze, Sinequa, Stratifyd y Synapsify.

13.3.1 Panorama actual de la Inteligencia Artificial

Sin duda, las 10 tecnologías y técnicas analizadas por Forrester en el estudio citado anteriormente son de gran impacto. Por estas razones dedicaremos una atención especial a las dos grandes técnicas de aprendizaje (automático y profundo) que, mediante algoritmos, están llegando a casi todos los sectores estratégicos de la empresa. Los asistentes virtuales (*chatbots*, principalmente)

son la expresión más práctica de la llegada de la Inteligencia Artificial a la empresa y a los usuarios domésticos en general. El reconocimiento de voz, el procesamiento de lenguaje natural y las redes neuronales artificiales —soporte del aprendizaje profundo— son otras tecnologías que trataremos en los próximos apartados, dedicando especial atención a la biometría y a los GPU, unidades de procesamiento gráfico, creadas por NVIDIA como soporte para los videojuegos, que se han convertido en una de las herramientas clave del aprendizaje profundo.

13.4 APRENDIZAJE AUTOMÁTICO

Machine Learning (ML) se traduce al español como "**aprendizaje automático**", aunque también se traduce como "**aprendizaje máquina**". Se suele considerar una rama de la Inteligencia Artificial que busca construir algoritmos que permitan a las computadoras "aprender" a partir de conjuntos de datos y obtener como resultado un modelo que permita realizar predicciones, basándose en dichos datos y no en instrucciones estáticas.

El aprendizaje automático es una disciplina que toma experiencias de otras disciplinas, tales como la estadística, la complejidad computacional, ciencias de la computación e ingeniería. La expansión del aprendizaje automático como disciplina complementaria o autónoma de la Inteligencia Artificial se debe, esencialmente, al diluvio de los datos (*Big Data*) que se han producido estos últimos.

Hoy en día, el aprendizaje automático está más que nunca al alcance de cualquier programador. Para experimentar con estos servicios tenemos plataformas como IBM Watson Developer Cloud, Amazon Machine Learning, Azure Machine Learning, TensorFlow o BigML.

En esencia, la máquina puede aprender a partir de la experiencia. Se trata de alimentar la experiencia de la máquina mediante objetos con los que entrenarse, para posteriormente aplicar patrones de entrenamiento de los datos. Existen diferentes tipos de algoritmos que dan diferentes categorías de aprendizaje.

Aprendizaje supervisado (*Supervised Learning*)

Requiere de la intervención de los humanos para indicar qué está bien y qué está mal (es decir, para proporcionar el refuerzo). En muchas otras aplicaciones de la computación cognitiva los humanos, aparte del refuerzo, también proporcionan parte de la semántica necesaria para que los algoritmos aprendan. Por ejemplo, en el caso de un *software* que debe aprender a diferenciar los diversos tipos de documentos que recibe una oficina, son los humanos los que inicialmente han de etiquetar un conjunto significativo de ejemplos para que posteriormente la máquina pueda aprender.

Aprendizaje no supervisado (*Unsupervised Learning*)

Consiste en entrenar una red exponiéndola a un gran número de ejemplos, pero sin "decirle" qué buscar. Por el contrario, la red aprende a reconocer características y a agruparlas con ejemplos similares, detectando así grupos ocultos, vínculos, o patrones dentro de los datos.

El aprendizaje no supervisado se emplea para buscar cosas de las que se desconoce su apariencia, por ejemplo el rastreo de patrones de tráfico en busca de anomalías que pudieran corresponderse a un ciberataque, el análisis de grandes cantidades de reclamaciones de seguros para detectar fraudes o agrupaciones de caras peludas que resultan ser gatos en YouTube.

Aprendizaje reforzado (*Reinforced Learning*)

Es un híbrido entre el aprendizaje supervisado y el aprendizaje no supervisado. Se basa en la psicología conductista y consiste en entrenar a una red neuronal para que interactúe con su entorno, retroalimentándola ocasionalmente con una recompensa. Su entrenamiento consiste en ajustar los pesos de la red para buscar la estrategia que genere mayores recompensas de manera más consistente.

DeepMind (empresa especializada en aprendizaje automático y profundo, comprada por Google) es el mejor ejemplo del éxito de este enfoque. En febrero de 2015 publicó un estudio en *Nature*, en el que describe un sistema de aprendizaje reforzado capaz de aprender a jugar 49 juegos clásicos de Atari, asistiéndose únicamente de los píxeles de la pantalla y la puntuación. El sistema aprendió a jugar a todos y cada uno de ellos desde cero y alcanzó un nivel similar o superior al humano en 29 de ellos. En marzo de 2016, su programa AlphaGo derrotó a Lee Sedol, el segundo mejor jugador del mundo de Go.

13.5 APRENDIZAJE PROFUNDO (*DEEP LEARNING*)

Es una subcategoría del aprendizaje automático; el **aprendizaje profundo** trata del uso de **redes neuronales** para mejorar cosas tales como el reconocimiento de voz, la visión por computadora y el procesamiento del lenguaje natural. Rápidamente se está convirtiendo en uno de los campos más solicitados en **informática** (ingeniería de sistemas o computación). En los últimos años, el aprendizaje profundo ha ayudado a lograr avances en áreas tan diversas como la percepción de objetos, la traducción automática y el reconocimiento de voz (todas ellas áreas especialmente complejas para los investigadores en IA).

A principios de los años noventa, la utilidad de las redes neuronales artificiales se reducía a tareas tan simples como reconocer números escritos a mano. En 1999, Geoffrey Hinton[8] y su grupo de investigación de la Universidad de Toronto publicaron sus primeras investigaciones sobre aprendizaje profundo. Éste

investigador y su grupo han transformado la investigación en IA y han recuperado investigaciones "abandonadas" sobre **visión computarizada, reconocimiento de voz, procesamiento del lenguaje natural y robótica.** En 2012 se lanzaron los primeros productos que entendían el habla —el asistente Google Now— y poco después han ido apareciendo aplicaciones que permitían identificar el contenido de una imagen, una característica que incorpora la aplicación Google Now.

Dos décadas después, varios grupos de investigadores descubrieron que las unidades de procesamiento gráfico (GPU) se adecuaban exponencialmente bien para ejecutar algoritmos de aprendizaje profundo, dotándolos de una velocidad 100 veces superior. Los mismos chips que se emplean para recrear mundos imaginarios resultan excelentes para ayudar a las computadoras a comprender el mundo real a través del aprendizaje profundo.

El aprendizaje profundo es una técnica incluida como parte del aprendizaje automático, una disciplina más amplia de Inteligencia Artificial que se basa en los principios empleados para entrenar sistemas de computación, con el objetivo final de que las máquinas se instruyan a sí mismas.

Bengio (2016[9] define **aprendizaje profundo** como "una manera de referirse a la simulación de redes de neuronas que **aprenden** gradualmente a reconocer imágenes, a comprender el lenguaje o incluso a tomar decisiones. Esta técnica descansa en las redes neuronales artificiales, un elemento básico de la investigación actual en IA. Dichas redes no imitan exactamente el funcionamiento del cerebro, si no que se basan en principios matemáticos generales que, a partir de ejemplos, les permiten aprender a detectar personas u objetos en una fotografía, o a traducir los principales idiomas del mundo".

La arquitectura del aprendizaje profundo consiste en redes neuronales artificiales que tratan de reproducir el proceso de solución de problemas del cerebro humano. Una red neuronal está formada por capas. La información entra por la capa de entrada y una serie de neuronas artificiales organizadas en capas "escondidas u ocultas" procesan la información, aplicándole distintos valores numéricos aleatorios o "pesos" y enviando el resultado a la capa de salida. Así, por ejemplo, en el caso de un reconocimiento facial, la entrada puede corresponder a varios rostros que se analizan en cada capa de la red antes de discernir su identidad, empezando por la capa de entrada, luego cada capa oculta identifica rasgos cada vez más complejos y, finalmente, produce una salida con el resultado de la imagen correcta y descartando las restantes imágenes.

Una red neuronal es un sistema de programas y estructuras de datos que se aproxima al funcionamiento del cerebro humano. Una red neuronal suele implicar un gran número de procesadores que funcionan en paralelo, teniendo cada uno de ellos su propia pequeña esfera de conocimiento y acceso a datos en su memoria local. Habitualmente, una red neuronal se "adiestra" o se alimenta con grandes cantidades de datos y reglas acerca de las relaciones (por ejemplo, "un abuelo es más viejo que el padre de una persona"). Luego, un programa puede

indicar a la red cómo comportarse en respuesta a un estímulo externo (por ejemplo, a un dato que introduce un usuario de computadora que está interactuando con la red) o puede iniciar la actividad por sí misma (dentro de los límites de su acceso al mundo externo).

Una red profunda, con muchas capas escondidas, es capaz de distinguir con gran detalle las propiedades de los datos de entrada. Entrenar una red supone ajustar los pesos internos de las neuronas, de manera que sea capaz de responder del modo deseado cuando se introduzca una entrada concreta.

Google emplea *Deep Learning* en sus algoritmos de reconocimiento de voz e imagen, Netflix y Amazon lo utilizan para adelantarse a tus gustos y los investigadores del MIT (Instituto Tecnológico de Massachusetts), para predecir el futuro.

El *Deep Learning* toma conceptos básicos de la **IA** y los enfoca en la resolución de problemas del mundo real a partir de redes neuronales profundas que imitan la forma en que nuestro cerebro toma decisiones. Es decir, emplea los datos que conoce para tomar decisiones sobre datos nuevos. Por eso es la tecnología más similar al funcionamiento cerebral humano.

En el **enfoque *Deep Learning*** se usan estructuras lógicas que se asemejan en mayor medida a la organización del sistema nervioso de los mamíferos, teniendo capas de unidades de proceso (neuronas artificiales) que se especializan en detectar determinadas características existentes en los objetos percibidos. La visión artificial es una de las áreas donde el *Deep Learning* proporciona una mejora considerable en comparación con algoritmos más tradicionales. Existen varios entornos y bibliotecas de código de *Deep Learning* que se ejecutan en las potentes GPU modernas tipo CUDA, como NVIDIA cuDNN.

El *Deep Learning* representa un acercamiento más íntimo al modo de funcionamiento del sistema nervioso humano. Nuestro encéfalo tiene una microarquitectura de gran complejidad, en la que se han descubierto núcleos y áreas diferenciados, cuyas redes de neuronas están especializadas para realizar tareas específicas.

Deep learning (aprendizaje profundo). A partir de un volumen de datos lo suficientemente grande como para poder extraer información valiosa, las redes neuronales profundas, basadas en el funcionamiento del pensamiento humano, son susceptibles de "enseñarles" a hacer todo tipo de cosas.

Desde el algoritmo que emplea Google para ordenar los resultados de su buscador, hasta el complejo (y todavía imperfecto) sistema de conducción autónoma de Tesla, pasando por el motor de recomendaciones de Amazon y los asistentes virtuales de compañías como Apple y Microsoft, la Inteligencia Artificial es el motor de todos ellos.

Origen de *Deep Learning*

El origen de la investigación de las capas profundas se remonta a 2005, cuando el Instituto Canadiense de Investigaciones Avanzadas (CIFAR) —una organización financiada por el Gobierno de Canadá y donantes privados— financiarán un proyecto de investigación dirigido por Geoffrey Hinton de la Universidad de Toronto (en la actualidad director de investigación de IA en Google), y en el que participaba Yann LeCun de la Universidad de Nueva York (en la actualidad director del laboratorio de investigación de IA en Facebook), Andrew Ng de la Universidad de Standford, Bruno Olshausen de la Universidad de California en Berkeley y Yoshua Bengio de la Universidad de Montreal. Sus investigaciones avanzaron con gran éxito y hoy constituyen la base de todos los grandes desarrollos de Inteligencia Artificial, desde los *chatbots* hasta la computación cognitiva de Watson de IBM o DeepMind de Google.

13.5.1 UNIDADES DE PROCESAMIENTO GRÁFICO (GPU)

Uno de los factores clave que ha contribuido al éxito del aprendizaje profundo ha sido que la velocidad de cómputo se duplicó de repente gracias a las Unidades de Procesamiento Gráfico (GPU, *Graphical Processing Unit*)[10], diseñadas inicialmente por NVIDIA para los videojuegos, y que permitieron entrenar a redes de gran tamaño en tiempos razonables. GPU es un nuevo modelo de computación acelerada, que utiliza procesadores gráficos masivamente paralelos, para acelerar aplicaciones también paralelas por naturaleza.

Modelo de computación acelerada GPU

El aprendizaje profundo es un nuevo modelo de *software* que necesita un nuevo modelo de computación, y numerosos investigadores de IA han adoptado la computación acelerada que ofrece GPU. Por esta razón, la empresa de productos de videojuegos NVIDA ha realizado esfuerzos considerables en la investigación de GPU en la Inteligencia Artificial para acelerar el desarrollo de redes neuronales profundas DNN (*Deep Neural Networks).*

NVIDIA ha creado un Instituto de Investigación de Aprendizaje Profundo. Tal es la inversión en investigación que ha hecho NVIDIA en estas técnicas de IA, que durante 2017 tenía previsto formar a 100.000 profesionales en el uso de IA con procesadores gráficos. IDC estima que para 2020 el 80 % de las aplicaciones tendrá un componente de Inteligencia Artificial. El Instituto de Aprendizaje Profundo de NVIDIA proporciona a los desarrolladores, científicos de datos e investigadores capacitación práctica en el uso de las herramientas y tecnología de Inteligencia Artificial más recientes. En España ha brindado formación al Centro de Supercomputación de Barcelona, uno de los centros de investigación de referencia mundial en computación avanzada.

Aceleración de IA con GPU: Un nuevo modelo de computación[11]

Jensen Huang, director del Instituto de Investigación de Aprendizaje Profundo de NVIDIA, fue invitado por Yann LeCun —uno de los padres del aprendizaje profundo— a dar una conferencia en el simposium celebrado en la Universidad de Nueva York sobre "The Future of AI" at NYU, cuyo título fue *Accelerating AI with GPUs: A New Computing Model*. Un resumen extraído de su conferencia es el siguiente:

> *"Los avances en el aprendizaje profundo han provocado la revolución de la IA. Las máquinas alimentadas por redes neurales profundas de IA solucionan problemas demasiado complejos para los codificadores humanos. Ellos aprenden de los datos y mejoran con su uso. La misma DNN puede ser entrenada, incluso por no programadores para resolver nuevos problemas. El aprendizaje profundo es fundamentalmente un nuevo modelo de software, de forma que necesitamos una nueva plataforma de computadora para ejecutarse en ella, una arquitectura que pueda ejecutar eficientemente comandos codificados por programadores así como entrenamiento masivamente paralelo de redes neuronales profundas"*[12].

13.5.2 EL FUTURO DEL APRENDIZAJE PRODUNDO

El aprendizaje profundo promete muchos avances, haciendo que la construcción de **automóviles sin conductor** y la creación de mayordomos robóticos constituyan posibilidades reales. Aún están limitados, pero lo que pueden lograr resultaba impensable hace sólo algunos años. Y el ritmo al que avanzan no tiene precedentes. La capacidad de analizar conjuntos de datos masivos y de utilizar el aprendizaje profundo en sistemas informáticos que puedan adaptarse a la experiencia, en lugar de depender de un programador humano, conducirá a avances tecnológicos de gran impacto. Abarcan desde descubrimientos de medicamentos o el desarrollo de nuevos materiales hasta la **creación de robots con un mayor nivel de consciencia acerca del mundo que los rodea.**

En España también existen empresas que aplican los conocimientos de *Machine Learning* en beneficio de sus clientes. Una de las más importantes es Inbenta, dedicada al desarrollo de *software* de procesamiento de lenguaje natural. Su tecnología permite que una máquina entienda y recuerde la conversación con una persona, gracias a la incorporación de la retención cognitiva, memoria y detección de contexto en las interacciones de sus máquinas y los usuarios.

El aprendizaje tiene numerosas aplicaciones en el campo de los asistentes virtuales en los departamentos de atención al cliente de las grandes empresas y en sus comunicaciones en general (como *e-mails, chats*, etcétera) y en sectores como la banca, los seguros, los transportes, el *retail* o las telecomunicaciones.

Otro de los exponentes españoles del *Deep Learning* es Sherpa, una empresa que ha diseñado un sistema que combina funciones de buscador, asistente personal y modelo predictivo, pensado para dispositivos móviles. Es una de las competencias internacionales de los cuatro grandes asistentes del mercado: Siri de Apple para dispositivos iOS, Google New y Google Assistant para móviles Android, Cortana de Microsoft y Alexa de Amazon.

Aplicaciones reales de aprendizaje profundo
Traducción automática, como es el caso de Google Translate.
Asistentes personales como Siri, Cortana, Alexa y Google Assistant en teléfonos inteligentes.
Automóviles autónomos sin conductor.
Asistencia médica en casos tales como interpretación de radiografías, tomografías o resonancias magnéticas, y los datos médicos van a poder organizarse sobre la base de patrones muchos más eficientes.

El éxito del aprendizaje profundo se basa en muchas características, pero destacamos que la velocidad de cómputo se ha duplicado gracias a "Unidades de Procesamiento Gráfico (GPU) diseñadas inicialmente para los videojuegos, lo que permitió entrenar a redes de gran tamaño en tiempos razonables, y los algoritmos se han beneficiado de la disponibilidad de grandes volúmenes de datos (*Big Data*) etiquetados para los cuales era posible encontrar la respuesta correcta (por ejemplo, gato) al inspeccionar una fotografía en la que un gato es sólo uno de los componentes". (Bengio 2016: 53)[13]

13.6 COMPUTACIÓN COGNITIVA

Computación cognitiva (*cognitive computing*) no está incorporada como término completo en el DRAE (Diccionario de la Real Academia Española). Sin embargo, buscando los términos por separado nos encontramos con las siguientes definiciones:

- **Computación**: "conjunto de conocimientos científicos y técnicas que hacen posible el tratamiento automático de la información por medio de computadoras".
- **Cognitiva**: "perteneciente o relativo al conocimiento".
- **Cognición**: sinónimo de "conocimiento", que tiene varias acepciones; las más próximas a computación son: "1. Acción y efecto de conocer; 2. Entendimiento, inteligencia, razón natural".

En resumen, **computación cognitiva** se puede considerar de la manera más simple como un sistema para conseguir que las computadoras piensen de forma humana o que se acerquen de la manera más fiel posible al funcionamiento del cerebro humano.

La computación cognitiva siempre ha estado ligada a la evolución de la Inteligencia Artificial, pero han sido los avances en *Big Data* y *Big Data Analytics,* en aprendizaje automático y aprendizaje profundo y procesamiento de lenguaje natural los grandes habilitadores de la disciplina. En la actualidad, y en la era de la industria 4.0, se trabaja con dispositivos como sensores, cámaras, micrófonos, drones y otros, que capturan información del medio y se puede considerar que son los sentidos de la computadora que tratan de asemejarse a los del humano: vista, oído, tacto, gusto y olfato, además del habla.

La computación cognitiva está muy relacionada con el tipo de problemas que sólo los humanos pueden resolver con su conocimiento y el uso del lenguaje natural, asociándolos con los sistemas cognitivos artificiales. Se trata de dotar a las máquinas de procesos similares a los procesos mentales de los humanos, de tal manera que las propias computadoras puedan interactuar y relacionarse con las personas. Su objetivo final es conseguir que la actividad de las máquinas se acerque lo máximo posible a la actividad que sucede en el cerebro humano.

Los algoritmos de aprendizaje automático y, más aún, de aprendizaje profundo (por su característica principal de incorporación de técnicas de redes neuronales profundas), busca desarrollar sistemas artificiales capaces de imitar a los humanos en aspectos tales como el aprendizaje de la propia experiencia, la constante adaptación del contexto y la interactividad entre distintos sujetos. De este modo, los sistemas de computación cognitiva utilizan los algoritmos específicos para el reconocimiento de voz, reconocimiento facial, etcétera, utilizando técnicas de procesamiento de lenguaje natural.

13.6.1 LA COMPUTACIÓN COGNITIVA EN LA ERA DE BIG DATA

En la actualidad, la computación cognitiva pretende extraer conocimiento de los grandes volúmenes de datos que se generan a una velocidad creciente y de una gran variedad, y hacer un uso inteligente de esos datos que se encuentran disponibles en diferentes formatos. El conocimiento se puede obtener del análisis de la información procedente de múltiples fuentes de datos —sensores, textos, videos, redes sociales, fotografías—. El análisis de todos estos datos nos proporciona conocimiento para tomar decisiones, pero también las técnicas y herramientas de Analítica Predictiva, que permiten una gran capacidad para conocer las tendencias de futuro para las organizaciones y empresas.

El gran cambio de la computación cognitiva es que sus tecnologías se entrenan y aprenden a partir de su propio conocimiento, y no de programación como la computación tradicional. La computación cognitiva permite trabajar con el

sistema como si tuviéramos un experto asesor (asistente virtual) que aprende rápidamente todo aquello que pueda necesitar el usuario. Cada vez es más posible entrenar a un sistema cognitivo para que asesore a los empleados de una empresa (banca, administración, etcétera) como si fuera un experto humano con años de experiencia en un determinado tema. Ejemplo de ello es **Watson** de **IBM**.

Los empleados podrán utilizar la computación cognitiva para aumentar su productividad o solucionar problemas del modo más rápido y eficiente. Aunque los riesgos de su utilización son altos, la computación cognitiva se ha de ver como una gran oportunidad para el empleado y no como una amenaza.

Actualmente, la computación cognitiva está liderada por IBM con su supercomputadora Watson, aunque los *bots* y *chatbots* (robots virtuales) forman parte de esta disciplina y los proveedores de Inteligencia Artificial van desde grandes hasta pequeñas y medianas empresas dirigidas por jóvenes emprendedores.

IBM Watson

IBM decidió hace varios años cambiar su modelo de negocio principal para enfocarse en las tecnologías cognitivas a través de su supercomputadora Watson, la referencia central del fabricante.

Watson se hizo famoso al competir en 2011 en un famoso concurso de TV estadounidense de preguntas, *Jeopardy!*, y derrotar a sus dos oponentes humanos, tal como lo hizo en 1997 la supercomputadora *DeepBlue*, que venció a Gary Kasparov (campeón mundial) jugando al ajedrez.

IBM Watson es una plataforma tecnológica que utiliza procesamiento del lenguaje natural y aprendizaje automático para revelar información clave de las grandes cantidades de datos no estructurados.

Watson es el primer sistema cognitivo diseñado de modo que las computadoras no se programen, sino que sean capaces de entender el lenguaje natural de las personas y aprender. Desde 2011 se ha convertido en una tecnología comercial accesible a través de la nube y que tiene clientes en numerosos sectores y países del mundo, entre ellos España, donde gracias a la colaboración con CaixaBank ha aprendido, además de técnicas financieras, el lenguaje español.

IBM Watson es un sistema informático de Inteligencia Artificial diseñado para realizar labores de computación cognitiva (entre ellas el procesamiento de lenguajes naturales y el razonamiento y el aprendizaje automático), desarrollado sobre la tecnología DeepQA de IBM. Es una plataforma tecnológica que utiliza procesamiento del lenguaje natural y aprendizaje automático para analizar y revelar información clave de las grandes cantidades de datos no estructurados.

En la práctica, Watson analiza datos no estructurados (artículos, reportes de investigación, datos empresariales, de redes sociales, de sensores, etcétera),

utilizando procesos del lenguaje natural para entender la gramática y el contexto; entiende preguntas complejas evaluando los posibles significados y determinando qué es lo que se está preguntando, y contesta a las preguntas más exigentes de sus clientes; extrae "casi en tiempo real" información clave de documentos, descubre y presenta información, patrones y relaciones entre datos. Watson aprende sobre un nuevo tema antes de contestar preguntas relacionadas. Utiliza algoritmos de aprendizaje automático, investigando en grandes volúmenes de datos *(Big Data*) para encontrar las muchas respuestas posibles.

Watson se utiliza hoy día en un gran número de aplicaciones en todo tipo de sectores, desde empresariales e industriales hasta administrativos, universitarios y de investigación. Uno de los proyectos relacionadas con IBM Watson es el Centro de IBM Watson para Internet de las cosas[14] —un centro de referencia mundial con sede en Múnich, Alemania, inaugurado el 15 de diciembre de 2015— y la plataforma expresa para ciberseguridad inaugurada a finales de mayo de 2016 y disponible comercialmente.

IBM Watson[15] es la primera plataforma de tecnología de computación cognitiva abierta. Representa una nueva era en la computación, en la cual los sistemas comprenden el mundo de la manera en que lo hacen los humanos: a través de los sentidos, el aprendizaje y la experiencia. Watson aprende continuamente, acumulando valor y conocimiento con el tiempo, a partir de interacciones previas. Con la ayuda de Watson, las organizaciones aprovechan el poder de la computación cognitiva para transformar industrias, ayudar a los profesionales a hacer mejor su trabajo y resolver desafíos importantes.

Definición de IBM Watson de Cognitiva[16]
"Una plataforma tecnológica que utiliza procesamiento del lenguaje natural y aprendizaje automático para revelar información clave de las grandes cantidades de datos no estructurados".

13.7 BOTS Y CHATBOTS

¿Qué es un *bot*? Un ***bot*** es un *software* de Inteligencia Artificial diseñado para realizar una serie de tareas por su cuenta y sin la ayuda del ser humano, como hacer una reserva en un restaurante, marcar una fecha en el calendario o recoger y mostrar información a los usuarios. El modelo más frecuente de *bot* es el *chatbot*, un robot capaz de simular una conversación con una persona, por lo que cada vez están más presentes en las aplicaciones de mensajería.

Numerosas plataformas de mensajería han ayudado a popularizar el uso y desarrollo de *chatbots*. La plataforma **Kik Messenger**, que cuenta con 275 millones de usuarios, lanzó una tienda de *bots*, entre los que se destacan un *bot*

para enviar a los internautas videos Vine de Twitter y otro que manda notificaciones a los usuarios sobre la compañía de maquillaje Sephora. Twitter también ha incorporado *bots* en su plataforma, como es el caso de uno que tuitea en el momento en que se registra un terremoto. Otra de las plataformas sociales que también ha incorporado esta tecnología es **Telegram**.

Pero, sin lugar a dudas, es Facebook el principal exponente de los *chatbots*, ya que, en la conferencia anual de desarrolladores, estas herramientas para operar en su canal de mensajería Messenger son parte central. La función que llevarán a cabo será facilitar la relación entre clientes, marcando así cómo responder a los usuarios con los denominados "mensajes estructurados", los cuales incluyen un título, una imagen, una descripción y una URL.

¿Quién los fabrica? Además de las plataformas citadas anteriormente, son numerosas las empresas fabricantes de *software* que desarrollan sus propios *bots* o bien facilitan las API necesarias para que los desarrolladores puedan construir sus propios *bots*. La multinacional Microsoft y la plataforma Slack son dos de las compañías que más están apostando en la creación de estos *softwares*.

13.7.1 LOS ASISTENTES VIRTUALES

Un ***bot*** (también llamado asistente virtual personal) es un programa informático basado en Inteligencia Artificial que imita el comportamiento humano para realizar diferentes tareas o funciones por su cuenta y sin la ayuda de un ser humano. El *bot* es capaz de comunicarse con los seres humanos (a través de texto, voz, emociones, etcétera), manteniendo una conversación con una persona utilizando el lenguaje natural en dicha conversación. Un ***chatbot*** *(chatter bot) o bot* conversacional es el modelo de *bot* más popular, capaz de simular una conversación con una persona, y se ha integrado en las aplicaciones de mensajería tipo *chat.* El *chatbot* ofrece un servicio a través de una conversación con el usuario y fundamentalmente por esta razón comienzan a estar muy presentes en aplicaciones de mensajería.

Los asistentes virtuales más populares son:

- Siri de Apple (para sistemas operativos iOS y Mac): es el más conocido, dado que también es uno de los más populares desde su incorporación a los teléfonos móviles iPhone y tabletas iPad.
- Cortana de Microsoft: incorporado al sistema operativo Windows 10.
- Google Now y Google Assistant: para el sistema operativo móvil Android.
- Alexa de Amazon.
- Bixby de Samsung, incorporado a sus teléfonos y tabletas inteligentes.

Sin embargo, muchos de los grandes fabricantes de computación, tanto de *hardware* como de *software*, así como empresas pequeñas innovadoras, están creando plataformas comerciales de *bots* a las que empresas clientes o particulares pueden conectarse y también desarrollar sus propios asistentes virtuales. Estos son los casos del ya mencionado Watson de IBM; Amazon con su plataforma Echo y asistente virtual Alexa; Google, que ha convertido su Google Now en una nueva plataforma denominada Google Assistant; Facebook Messenger, con su plataforma M para *bots* presentada en abril de 2016, que facilitará las API (interfaz de programas de aplicación) para el desarrollo de sus propios *chatbots* por empresas y particulares, que los podrán integrar en su propia red social.

Samsung ha comprado en 2016 el asistente virtual Viv y tiene el proyecto de incluirlo en todos sus productos (televisores, teléfonos, entre otros) a través de sus *bots*. Además de estas plataformas de las grandes empresas de computación, han nacido empresas que han creado sus propias plataformas de *bots* y que tienen también cientos de millones de usuarios, como es el caso de Kik y Slack, diseñadas especialmente para clientes empresariales y profesionales.

Sherpa es una plataforma española para *chatbots* que ha conseguido bastante popularidad al nivel de su competencia en los últimos años.

13.7.2 APLICACIONES DE LOS BOTS

Algunas aplicaciones de asistentes citados anteriormente son:

El asistente virtual **Cortana** de Microsoft, **Siri** de Apple o **Alexa** de Amazon, con su altavoz Echo (por citar algunos ejemplos) pueden actuar como agentes inteligentes y son capaces de interactuar con *bots* para la ejecución de aplicaciones de comercio electrónico o soluciones de atención al cliente, ya sea con programas de gestión de relaciones con los clientes (CRM) o con *call centers*.

WeChat. Es una aplicación de mensajería instantánea muy popular, que ha desarrollado sus propios *bots*, que permiten, por ejemplo, realizar reservas de hotel, confirmar citas médicas o comprar entradas para el cine a través de mensajes de texto que son interpretados por *bots*.

Google Assistant. Apoyado por el servicio de mensajería **Allo**, lanzado en septiembre de 2016, y basado en aprendizaje profundo (*Deep Learning*), pretende dar respuestas predictivas de modo instantáneo y tiene la capacidad de aprender sobre el usuario y las personas con las que se relaciona habitualmente para hacer sugerencias u ofrecer información. La revista *Hosteltur*[17], en su número de julio-agosto de 2016, describe numerosas aplicaciones de *bots* en la industria del turismo y, en particular, comenta diferentes aplicaciones presentadas en la feria Fiturtech 2016. Así, en el caso de Google Assistant cita

algunas aplicaciones de interés: Por ejemplo, si una persona habla con un contacto sobre una cita para cenar, el asistente virtual de Google le ofrecerá restaurantes y la posibilidad de realizar la reserva sin salir de la aplicación Allo (lanzada oficialmente en septiembre de 2016) y teniendo en cuenta sus preferencias culinarias permitirá acceder a recorridos virtuales de hoteles, restaurantes u otros recursos turísticos basados en tecnología StreetView (la aplicación de Google Maps) que estarán integrados.

Otros asistentes virtuales en uso en numerosas páginas web de organizaciones y empresas son:

- **Elvira.** Asistente de la Universidad de Granada, con una interfaz elaborada con apariencia de ser humano, que trata de dar respuestas acerca de la información de la Web: (`https://www.ugr.es; http://tueris.ugr.es/elvira`).
- **Irene.** Ayuda en la compra de billetes de tren en Renfe: (`http://consulta.renfe.com/renfeO/index.jsp`).
- **Anna.** Asistente al usuario en la compra de muebles en IKEA (cadena de almacenes): (`www.ikea.com/es/es`)

Los *bots*: ¿las nuevas aplicaciones móviles?

Existe una tendencia creciente de sustitución de aplicaciones móviles por *bots*. Esta es la opinión de Satya Nadella, presidente de Microsoft, quien en la presentación de su plataforma de *bots* planteó que, en lugar de tener que utilizar multitud de aplicaciones móviles, los usuarios podrán interactuar con los *bots* de una manera mucho más natural utilizando su propio lenguaje. Con este motivo, Microsoft —y muchos de los gigantes tecnológicos o empresas especializadas— facilitará el desarrollo de *bots* compatibles con su plataforma por parte de empresas y usuarios finales.

En Facebook, pocos meses después de la presentación de su plataforma M (Messenger) de *bots*, en abril de 2016, se han abierto más de 30.000 *chatbots* para ofrecer servicios personalizados. El cometido principal de estos *chatbots* es facilitar la relación entre el cliente y la marca, y desde la propia red social de la empresa. Facebook permite de esta manera que la red social de la empresa pueda ser el canal de comunicación con clientes, proveedores, y socios, utilizando un *chatbot* propio de la empresa como complemento a las funcionalidades características de su red social.

Voz

En los últimos años, prácticamente todos los gigantes tecnológicos, de Google a Apple (pasando por Microsoft), han presentado su propio asistente virtual de voz. Google se sumó a las otras grandes multinacionales del *software* y, en 2016,

lanzó Google Assistant, pensado para facilitar tareas cotidianas en el hogar a través del lenguaje.

Cada vez más, los usuarios prefieren dejar a un lado el teclado y comunicarse con los dispositivos a través de la voz. Aplicaciones de mensajería instantánea, tales como Facebook, Messenger y WhatsApp, ya hace tiempo que han incorporado esta funcionalidad. Según Google, alrededor del 20% de las búsquedas realizadas con Android en Estados Unidos se hacen por medio de voz.

13.7.3 PLATAFORMAS DE BOTS

En 2018, los *bots* y los *chatbots* seguirán siendo una de las tendencias más disruptivas de la Inteligencia Artificial y de las ciencias de la computación. Las plataformas de mensajería más utilizadas y populares, como Facebook y Skype, anunciaron ya en 2018 ofertas a los desarrolladores que desearan soportar aplicaciones de *chatbots* en sus propias plataformas. Al mismo tiempo, otras plataformas de mensajería, como Slack y Telegram, desarrollaron sus propios almacenes de *bots* (al estilo de los almacenes de aplicaciones), con el objetivo de potenciar sus líneas de negocios y atraer inversores.

Las plataformas de mensajería permiten construir sus propios *bots*. Esta posibilidad es una gran oportunidad para muchas empresas, que de no ser por ello no les quedaría otra opción que desarrollar de modo completo sus *bots* empresariales. De esta manera, bastará integrarse con las API que ofrecen estas plataformas y desarrollar sus propios *bots* al estilo de una *app* de teléfono inteligente. Las plataformas más populares para desarrollar *bots* son las siguientes:

Facebook Messenger. Desde el lanzamiento de su plataforma en abril de 2016, se han asociado a las páginas de Facebook más de 40.000 *bots* y, en consecuencia, se está convirtiendo en una plataforma ubicua al permitir que las comunicaciones de la empresa se realicen de modo omnicanal, pero a través de la propia red Facebook.

Kik. Es una de las plataformas de mensajería y de *bots* más populares, con cientos de millones de usuarios.

Telegram. Es una plataforma muy joven (se lanzó en 2013) pero con más de 100 millones de usuarios. Su característica más destacada es su propiedad de seguridad avanzada y encriptación.

Skype. El servicio de mensajería y llamadas por voz mediante videoconferencias, ofrece la posibilidad de desarrollar *bot*s incorporándolos a su propia cuenta de usuario.

Twitter. Esta red social también permite el desarrollo de *bots* en su plataforma.

WeChat. Ya lo comentamos con anterioridad. Su gran ventaja es ser el programa líder en mensajería en China.

Slack. Es una plataforma de mensajería de grupos diseñada para funcionar en ambientes grupales de trabajo. Proporciona uno de los primeros almacenes de *bots (bots stores)* de la industria, de modo que los equipos pueden desarrollar e instalar con gran facilidad sus propios *bots*.

Aplicaciones tradicionales de SMS y de correo electrónico, que permiten utilizar plataformas de API, como el caso de Twilio para construir *bots* a partir de SMS y *bots* de correo electrónico, usando proveedores como Lonelybots.

13.8 CHATBOTS DE EMPRESA: EL CASO DE LA ATENCIÓN AL CLIENTE

Las empresas comienzan a incorporar asistentes virtuales, por ahora preferentemente, en sus servicios de atención al cliente y de modo experimental. Los servicios ofertados, todavía con restricciones, son muy variados: servicios de atención, recomendadores virtuales, buscadores, centros de atención al cliente —sustituyendo o ampliando las funciones de los *call centers*— o *chat centers*, automatización de ventas, campañas de *marketing* en redes sociales, etcétera. Comienzan a utilizarse en servicios de hostelería y también como secretarios y asesores virtuales personales. En la actualidad, la mayoría de los *chatbots* están integrados en plataformas como las ya comentadas anteriormente: Facebook, Twitter, Slack, Line o Telegram, entre otras.

Algunos casos de éxito recogidos de la prensa especializada[18] son:

Skyscanner. Un comparador de vuelos integrado en Facebook Messenger, que ayuda a los viajeros a encontrar vuelos económicos o sugerir destinos en el caso de indecisión.

Destinia. Agencia de viajes española con un *chatbot* integrado en Facebook Messenger, que ofrece un asistente para búsqueda de hoteles en su plataforma. El programa simula una conversación de atención al cliente, peticiones, reservas, cancelaciones de hoteles y también de búsqueda de otros servicios como vuelos.

ImaginBank. Es el banco móvil de La Caixa —uno de los tres grandes bancos españoles— y ha sido el primer *chatbot* financiero de España y desarrollado por el propio Banco La Caixa. Funciona también a través de Facebook Messenger. Ofrece servicios de atención a los clientes sobre servicios financieros del banco; está muy dirigido a los clientes "millennials". Da buena información sobre preguntas concretas y promociones disponibles.

Mahou. La empresa cervecera española Mahou ha lanzado una iniciativa pionera en gran consumo. Permite a los usuarios de Facebook informarse en tiempo real

sobre propuestas de ocio en Madrid, especialmente relacionadas con la cerveza, comidas, restaurantes, etcétera.

AirEuropa. La línea área española ha elegido Twitter para la puesta en marcha de su primer *chatbot* de atención al cliente. Se utiliza a través de su sección Mensajes, de igual forma que si se fuese a enviar un mensaje privado a la compañía. El *chatbot* se ejecuta y pregunta sobre opciones preferidas por el cliente. Es bueno para conocer el estado de un vuelo, datos como la puerta de embarque, retrasos, etcétera. En caso de no poder responder adecuadamente, el *chatbot* redirige a un operador.

Ticketmaster. La compañía pionera en puntos de venta de entrada y centros de llamada en todo el mundo, se ha especializado en la búsqueda semántica y en procesamiento de lenguaje natural. Resuelve gran número de preguntas y dudas de sus clientes.

Moovit. Es un *chatbot* especializado en transporte. El *chatbot* de la aplicación está integrado en tiempo real con Facebook Messenger. Dado que la aplicación tiene gran número de usuarios y cuenta con muy buena información de transporte público en más de 800 ciudades de todo el mundo, es un buen asistente para la planificación de desplazamientos y sugiere trayectos, utilizando, preferentemente, la red de transporte público.

Correyvuela. Es una aplicación de economía colaborativa que está conectada a la red social Facebook y, además de sus servicios usuales, ofrece un servicio de venta de vuelos a través de su *chatbot*.

CASO DE ESTUDIO: ALEXA DE AMAZON[19]

Alexa es el asistente inteligente virtual de Amazon y fue presentado en la mayor feria electrónica de consumo del mundo, CES 2017 de las Vegas, para su integración a múltiples dispositivos: teléfonos móviles inteligentes, automóviles y electrodomésticos. Otro objetivo de Amazon es potenciar sus negocios y sus tiendas de comercio electrónico.

Algunas de las aplicaciones presentadas para Alexa en CES 2017 fueron: sincronización con el teléfono inteligente para llamar, poner música o incluso solicitar un taxi; un acuerdo con el fabricante de automóviles Ford, donde Alexa actuará como copiloto, permitiendo al conductor buscar direcciones o conocer información del vehículo mediante órdenes en voz alta. De igual forma Alexa, se integrará en electrodomésticos como en el caso del fabricante Whirlpool, que presentó una lavadora controlada con la voz, o el fabricante de teléfonos Huawei, que tiene previsto traer preinstalado Alexa en sus nuevos modelos.

Amazon presentó también su altavoz para el hogar, Echo, que integra el asistente virtual Alexa y que en la actualidad se comercializa en numerosos países del mundo. El usuario puede interactuar con su entorno en forma natural,

formulando preguntas o dando órdenes del tipo: “Alexa, prende la lavadora” o “Alexa, ¿está apagado el horno?”.

La interconexión del asistente virtual Alexa y su altavoz Echo para el hogar potenciará también la explosión de Internet de las cosas y el fomento del comercio electrónico, ya que se podrá hacer la compra de artículos de todo tipo por voz y a través del teléfono inteligente.

El gran rival de Alexa y Echo de Amazon, es el altavoz inteligente Google Home y Google Home Mini que también se comercializa en un gran número de países del mundo (entre ellos España que desde junio de 2018 se comercializa por 149 euros y 59 euros respectivamente)

CASO DE ESTUDIO: SLACK

Slack es una plataforma de mensajería que ha sido creada para trabajar en equipo. Es un sistema avanzado de mensajería instantánea pensada para la oficina y grupos de trabajo, muy utilizada por las empresas para que sus trabajadores se comuniquen entre sí. Dispone de varias herramientas incorporadas con la finalidad de facilitar las tareas de comunicación entre ellos, principalmente mediante *chats* de grupo y, últimamente, mediante *chatbots*.

Una de las últimas grandes propuestas de Slack es que ha abierto su API (interfaz de programación de aplicaciones), lo que permite comunicarse con ella, ya sea desarrollando sus propios *chatbots* o bien seleccionando uno de su oferta. Slack se ha convertido en una tienda de *bots* al estilo de las tiendas de *apps* (*bots store*), de manera que existen *bots* que los directivos de la empresa pueden utilizar para mejorar su productividad, ya que es cuestión de seleccionar la oferta que mejor se adapte a sus necesidades. En esencia, Slack es una plataforma colaborativa para trabajo en grupo.

Aparte del gran número de usuarios, Slack ha alcanzado una gran notoriedad como tecnología emergente y disruptiva. *MIT Technology Review*[20], la revista de referencia del MIT de Massachussets, ha publicado en un número extraordinario —enero de 2017— las diez tecnologías más disruptivas de 2016, entre las que se incluyen tecnologías muy diferentes: desde robots, almacenes de DNA **app**, interfaces conversacionales, entre otras, y la gran sorpresa es que ha sido seleccionada Slack.

¿Cuáles son las razones por las que la revista del MIT ha seleccionado a Slack como tecnología disruptiva de 2016? Una de las razones es el sistema de mensajería intraoficina, que se describe como el *software* de espacio de trabajo de más rápido crecimiento del mundo que se haya visto nunca. Sobrepasa dos millones de usuarios diarios en menos de tres años desde su lanzamiento en 2013. Es un *software* de comunicación fácil de utilizar, que está sustituyendo al correo electrónico como un método para realizar trabajos de oficina. Lee Gomes, el autor del artículo que aporta las razones del éxito de Slack, dice que éstas

residen en ser un lugar centralizado para comunicarse con los colegas a través de mensajería instantánea y *chat* —recientemente con *chatbots*— y que puede reducir el tiempo gastado en el correo electrónico.

Slack, como plataforma colaborativa, facilita las conversaciones entre colegas del mismo grupo de trabajo, o diferentes, y mejora enormemente la productividad. La propia revista señala otros softwares de comunicación de la competencia, también relevantes: Quip, Hipchat y Microsoft. A finales de 2016, y para hacer frente a la competencia de Slack, Microsoft lanzó Microsoft Team integrado con Office 365. A partir de 2018 la competencia será muy fuerte, sobre todo con la publicidad que le ha otorgado a Slack su selección en la MIT Technology Review.

En su página web (`www.slack.com`), disponible en inglés y en español se puede ver la cantidad de ofertas de la plataforma colaborativa Slack, tales como mensajería, llamadas telefónicas, videoconferencias, compartición de todo tipo de archivos, imágenes, archivos PDF, documentos, hojas de cálculo, etcétera, e, incluso, puede sincronizar con documentos almacenados en Google Drive, Dropbox o Box.

En resumen, Slack es una herramienta para trabajo en grupo (*groupware* o *groupwork*), que, si se cumplen las predicciones de la revista del MIT, puede llegar a convertirse en una herramienta de trabajo de ofimática y de gestión empresarial, que integrará todas las herramientas de comunicación actuales y futuras utilizadas en el trabajo diario en la oficina, fuera de ella, en el teletrabajo y restantes modalidades colaborativas.

CASO DE ESTUDIO: BIXBY DE SAMSUNG

El 20 de marzo de 2017, la empresa Samsung presentó el chatbot Bixby, que vendrá instalado de serie en sus nuevos teléfonos inteligentes Galaxy S8, y versiones futuras (las nuevas versiones Galaxy S9 y las previstas presentar en 2019, Galaxy S10, incorporaran de fábrica -según noticias de la propia Samsung- la nueva versión 2.0 del asiste virtual Bixby).

El asistente virtual de Samsung es un servicio en la nube y tiene como novedad importante que se pretende conectar a todo tipo de productos de la compañía surcoreana, desde televisores hasta frigoríficos o sistemas de aire acondicionado, ya que está integrado en Tizen, el sistema operativo propio de Samsung para electrodomésticos y terminales de todo tipo.

Bixby tiene un botón de acceso directo en el lateral del terminal y podrá controlar con la voz algunas aplicaciones y realizar casi las mismas funciones que se realizan desde la pantalla. Samsung tiene previsto como novedad el lanzamiento de un kit de desarrollo de *software* (SDK) para que los desarrolladores de aplicaciones puedan integrar BixBy en sus programas. Algunas de las funciones destacadas el día de la presentación son:

- Ejecución de comandos táctiles.
- Reconocimiento de imágenes.
- Traducción de idiomas (la versión original soporta 52 idiomas).
- Más información sobre usuarios.
- Aplicaciones compatibles. Proporcionará un kit de desarrollo de *software* y los desarrolladores podrán trabajar con las API de Bixby; por ejemplo, las API pueden ser de gran utilidad para los minoristas que busquen integrar en sus tiendas el asistente virtual.

La posibilidad de competir con sus otros rivales fabricantes de *software*, como Apple, Microsoft, Amazon y Google, y con las plataformas de mensajería instantánea, como Slack, Kik o Telegram, proporciona a Bixby una gran versatilidad. El bienio 2018-2019 será decisivo para ver los progresos de los restantes asistentes virtuales y del propio Bixby.

13.9 EL ALGORITMO DE INTELIGENCIA ARTIFICIAL COMO MODELO DE NEGOCIO EN LA ECONOMÍA DIGITAL

La Inteligencia Artificial y sus dos técnicas más sobresalientes, aprendizaje automático y aprendizaje profundo, se apoyan en el desarrollo y posterior ejecución de algoritmos. Las técnicas de aprendizaje automático y aprendizaje profundo utilizan técnicas avanzadas de algoritmos desarrolladas en lenguajes de programación especializados y ejecutados en software de Inteligencia Artificial, ya sea como programas independientes o en plataformas de computación cognitiva, como es el caso de Watson de IBM, o las plataformas *Deep Learning* de Google o las específicas de Facebook, Microsoft o Amazon.

Los lenguajes de programación utilizados para escribir algoritmos para aprendizaje automático y aprendizaje profundo son numerosos y van desde los tradicionales de Inteligencia Artificial (Prolog o LISP) hasta los lenguajes de programación tradicionales como C/C++, Java, JavaScript, etcétera. Sin embargo, en los últimos años se han consolidado lenguajes de programación que disponen de bibliotecas de programas con algoritmos resueltos, que ayudan a la construcción de los algoritmos de aprendizaje automático y aprendizaje profundo. Entre estos cabe destacar a los lenguajes R y PHYTON, por lo que si el lector desea aprender a desarrollar algoritmos para desarrollo en Inteligencia Artificial, le recomendamos que, además de aprender los fundamentos teórico-prácticos de algoritmos que le permitan resolver problemas de mayor o menor complejidad, piense en la formación en estos dos lenguajes y en el aprendizaje de su biblioteca de algoritmos resueltos que podrá utilizar, dada su condición de bibliotecas de código abierto (*open source*), e integrarlos en el desarrollo de sus propios algoritmos.

Origen de los algoritmos

Hace 3.500 años los babilonios imaginaron los primeros algoritmos y, cuentan las historias, que lo hicieron mediantes el cálculo de raíces cuadradas. El término más moderno proviene del matemático árabe Al'Khwarizmi, que escribió un tratado sobre los números. Los algoritmos, como métodos para resolución de problemas, se han utilizado desde los tiempos más remotos y en las carreras de ciencias matemáticas, físicas e ingeniería han sido también uno de sus pilares académicos y científicos. En el siguiente apartado daremos definiciones prácticas de algoritmos, así como consideraciones prácticas sobre su desarrollo y evolución.

¿Cuál es, entonces, la razón de su popularización actual en los campos y sectores más diferentes? Sin lugar a dudas, la razón es que se han convertido en la pieza clave para definir los procesos de negocios en la época de los grandes volúmenes de datos, de manera que un buen algoritmo aplicado a la gestión de datos constituye la espina dorsal de la economía digital, así como de la transformación digital de organizaciones y empresas.

La transformación digital junto con la Inteligencia Artificial, Internet de las cosas y *Big Data* están cambiando la vida cotidiana de los seres humanos debido, precisamente, a los datos que los usuarios dejan almacenados en la nube.

Los algoritmos son una herramienta utilizada a diario en la empresa, en la industria, en los negocios y, sobre todo, cuando navegamos por Internet. Aunque más adelante describiremos en detalle algoritmos de impacto, adelantemos algunos de uso común y que utilizamos casi a diario:

- **Algoritmo de búsqueda de Google (*Page Rank).*** Permite rastrear millones de páginas web en milésimas de segundo para ofrecer al usuario las mejores soluciones en su búsqueda de datos o para responder a preguntas, ajustándose a su perfil de comportamiento en la navegación por Google.
- **Algoritmo de Netflix** —una de las compañías líderes a nivel mundial en reproducción de películas y videos—. Presenta al usuario las mejores series o nuevas películas que puedan ser de su agrado, en función de sus hábitos diarios ante la citada plataforma de videos.
- **Algoritmo de Waze**. Ofrece las mejores rutas en función del tráfico, tanto para los viajeros usuarios del servicio como para los próximos vehículos autónomos.
- **Algoritmos de recomendación para la compra en línea**[21].

Los algoritmos empiezan a ser la esencia de la economía digital y soporte de una nueva rama de la economía a la que se suele denominar **economía del algoritmo.** Se está creando un nuevo mercado digital donde se pueden alquilar, comprar y vender algoritmos. Al igual que las tiendas *on line* de aplicaciones para móviles (*apps*), se comienza a desarrollar un nuevo modelo de negocio basado en

algoritmos y con un objetivo y líneas de negocio similares, es decir, desarrollo, venta, compra y alquiler de éstos. La plataforma Algorithmia —que posteriormente comentaremos— es ejemplo de un modelo de negocio de éxito basado en la creación, desarrollo y comercialización de algoritmos.

La arquitectura de los negocios se está redefiniendo por los algoritmos. Su uso práctico lleva años en industrias y negocios diferentes. Sin embargo, el desarrollo de máquinas cada vez más inteligentes, con el soporte de algoritmos —preferentemente de aprendizaje automático— es una de las razones para el éxito y penetración de los algoritmos en el sector industrial y en otros sectores importantes, tales como la salud, la aeronáutica, la automoción, etcétera.

"Mientras que los negocios digitales están ya transformando las organizaciones, creemos firmemente que las compañías basadas en algoritmos crearán aun mayor nivel de disrupción". "Los mercados de algoritmos abiertos crearán rápidamente (además de incentivar) un nuevo y completo ecosistema de desarrolladores de algoritmos de la misma forma que las *apps stores* y los dispositivos móviles han cambiado el desarrollo *software*" (Gartner, 2016)[22].

13.9.1 ¿QUÉ ES UN ALGORITMO?

La Real Academia Española (RAE) dice que el término puede provenir del latín *algobarismus* y del arábigo, abreviatura de *hisabu lgubar* (cálculo mediante cifras arábigas). También se consideran sus orígenes en el matemático persa Al-Juarismi. Las dos acepciones de **algoritmo** dadas por la RAE son: "1. Conjunto finito y ordenado de operaciones que permite hallar la solución de un problema; 2. Método y notación en las distintas formas de cálculo."

"Un **algoritmo** es un método con un conjunto de instrucciones utilizadas para resolver un problema específico". Un problema puede ser resuelto mediante varios algoritmos. Un algoritmo dado correcto resuelve un problema definido y determinado (por ejemplo, calcular una función determinada) (Joyanes y Zahonero, 2014). Desde el punto de vista de programación de computadoras, los algoritmos son un conjunto de instrucciones de programación que, introducidas en un programa, permiten analizar un conjunto de datos —previamente seleccionados— y establecer una salida o resultado. Los algoritmos, pensando en computación, necesitarán herramientas de programación tales como pseudocódigo, diagramas de flujo o lenguajes de programación (C, C++, Java, JavaScript, R, Python, etcétera).

Harari también define algoritmo como "un conjunto metódico de pasos que pueden emplearse para hacer cálculos, resolver problemas y tomar decisiones. Un algoritmo no es un cálculo concreto, sino *el* **método** que se sigue cuando se hace el cálculo" (Harari, 2016).[23]

Algoritmos sencillos que se suelen describir en los libros de algoritmos y programación, y que también Harari describe, pueden ser:

- Calcular la media aritmética de dos números.
- Realizar una receta de cocina. Hacer un gazpacho andaluz, un salpicón de mariscos, una sopa azteca o un ceviche.
- Calcular la superficie de un triángulo.

Harari, afirma rotundamente que "algoritmo es el concepto más importante en nuestro mundo. Si queremos comprender nuestra vida y nuestro futuro, debemos hacer todos los esfuerzos posibles por entender qué son los algoritmos y cómo éstos están conectados con las emociones"[24]. Es decir que interactuamos día a día con uno o varios algoritmos.

Los algoritmos, unidos a *Big Data*, Internet de las cosas y la Inteligencia Artificial, están presentes en casi todos los aspectos de la vida actual. Citemos algunos casos. Los algoritmos deciden lo que se publica en nuestro muro de Facebook, las recomendaciones de Amazon, permiten divertirnos mejor con nuestro videojuego favorito y las compañías de seguro disminuyen sus riesgos. Google es el caso más significativo con su algoritmo de *Page Rank*, que decide el orden de aparición de las búsquedas en el buscador, afinando los resultados, ajustándolos a nuestro perfil de comportamiento cuando navegamos. En resumen, los algoritmos controlan gran parte de lo que vemos y oímos en Internet.

Si nos centramos en la economía digital y en el *marketing* digital, los algoritmos intervienen en la fijación de precios, ayudan a las marcas a utilizar el conocimiento específico del cliente (datos personales, perfil de su comportamiento en navegación, hábitos de compra, etcétera), sugieren nuevos productos para ofrecer a los clientes, ayudan a las tiendas a establecer y modificar precios. Las empresas de medios de comunicación los utilizan para hacer recomendaciones y ofrecer contenido y anuncios, utilizando la publicidad programática basada en el comportamiento del usuario cuando se conecta a cualquier página web, de tal manera que le aparecen anuncios de publicidad relacionados con sus hábitos de navegación, tales como viajes, compras, lecturas, alojamientos hoteleros, etcétera.

13.9.2 LOS ALGORITMOS EN LA EMPRESA

Los algoritmos se han ido introduciendo en la empresa. Steve Prentice, vicepresidente de Gartner en *Computerworld*, señala:

> *El sector de retail utiliza algoritmos de automatización de precios y de analíticas para mejorar los resultados de su negocio. En recursos humanos los algoritmos ya han empezado a transformar la adquisición de talento, ya que son capaces de evaluar rápidamente la idoneidad de cada candidato para cada rol, pero esta*

> *misma tecnología puede además utilizarse para repartir mejor la cantidad de trabajo y tareas entre las personas adecuadas. En sanidad, la accesibilidad de los algoritmos avanzados clínicos está transformando la eficiencia de la sanidad que se administra en las organizaciones, así como sus capacidades para llevarlo a cabo.*[25]

La práctica de compartir y codesarrollar algoritmos entre organizaciones con intereses similares puede ser muy relevante para todas las empresas, además de ser un modelo de desarrollo muy útil en las industrias verticales. Los algoritmos están siendo usados por las empresas, fundamentalmente, para detectar patrones o tendencias y, en función de ello, generar datos útiles que permitan adaptar mejor sus productos o servicios

La capacidad de las técnicas utilizadas en los algoritmos para identificar y extraer el valor real de los datos está provocando en las compañías la aceleración del despliegue de la economía digital. El algoritmo se ha convertido en la gran promesa de la economía digital y está dando lugar al nacimiento de una nueva rama de la economía, denominada **economía de (los) algoritmos** o **economía algorítmica**.

Los algoritmos han tenido el gran éxito actual debido al gran aumento de los datos y su facilidad para su gestión, especialmente, en la actualidad, donde la inmensa cantidad de datos existentes son datos no estructurados (esencialmente *Big Data*). Estadísticas fiables hablan de que sólo 10% de los datos son estructurados —los datos tradicionales de las bases de datos relacionales—, por lo cual se han originado grandes oportunidades para empresas y desarrolladores que pueden crear algoritmos para comercializar y desplegar los restantes; es decir, los algoritmos se han constituido en un gran negocio. Si a esto unimos el crecimiento exponencial de los datos, la oportunidad sigue creciendo. Deloitte —y muchas otras grandes consultoras— estima que desde el comienzo de la Humanidad hasta 2003 se generaron dos exabytes de información, mientras que en 2011 se creó ese mismo volumen en dos días y, en 2020, se prevé que se tardará menos de diez minutos.

13.9.2 LOS ALGORITMOS COMO SERVICIO

El desarrollo de algoritmos ha dado lugar a un nuevo modelo de servicios (**AaaS**, *Algoritmos como Servicio*) al estilo de los modelos de la nube. Las empresas en la actualidad tienen a su disposición grandes volúmenes de datos y necesitan un algoritmo (en realidad, un programa de *software*) para su análisis y su aplicación en la toma de decisiones sobre cómo mejorar un servicio. De esta forma, las empresas —al igual que sucede con los servicios de la nube— en lugar de dedicar medios y personal para el desarrollo del algoritmo necesario, compra o alquila uno e, incluso, cuando no se exigen grandes requisitos, también puede encontrarlo en forma gratuita. El algoritmo como servicio es útil para las empresas que toman sus decisiones basándose en datos.

La tarifa de pago del algoritmo depende de su uso, de las cantidades de ejecuciones (volumen de datos), la complejidad de algoritmos y los beneficios esperados. Existen empresas que ofrecen servicios sin costo alguno, pero hasta un cierto límite de volumen de datos. Las empresas no necesitan dedicar recursos propios a la programación, diseño y prueba de algoritmos. Así, los servicios especializados en su creación ponen a disposición de organizaciones y empresas las capacidades necesarias para resolver un problema específico que realiza un desarrollador en un *software* específico o en una plataforma para el desarrollo de otras aplicaciones. El servicio de algoritmos ofrece un *software* para una necesidad específica y está dirigido al usuario final.

CASOS DE ESTUDIO DE ALGORITMOS

Los algoritmos existentes en la actualidad son innumerables y, de hecho, cálculos fiables consideran que superarán en número a las actuales *apps* de las tiendas de aplicaciones. Así, podemos considerar algoritmos de productividad en la empresa como: análisis de texto y de lenguaje natural, detección de emociones en redes sociales, correos electrónicos, foros, mensajes o llamadas telefónicas, etcétera. De este modo, se pueden diseñar campañas de marketing efectivas y se mejora la atención y la experiencia del cliente.

Existen también algoritmos para su uso en la vida diaria: análisis de imágenes (detección de fraudes), de rostros (vigilancia), de clientes (índice de abandono en procesos de compra, segmentación, ofertas de productos, patrones de consumo), para sectores como la industria, las estadísticas o actividades financieras. A continuación, se describen algunos algoritmos populares[26] utilizados en la vida diaria de las personas y las empresas:

- **Edge Rank de Facebook**. Algoritmo que influye en la visibilidad de las publicaciones en la página de inicio o en el muro de Facebook y afecta tanto a usuarios como a empresas. Su función es optimizar el listado de noticias en el muro del usuario y determina cuáles publicaciones aparecen y en qué orden. Asimismo, determina también a qué seguidores le aparecen las publicaciones de un usuario. **Edge Rank**, entre otras cosas, se encarga de predecir y filtrar el tipo de contenidos que recibimos de nuestros "amigos". Si el lector se ha percatado de que nunca recibe ciertos contenidos de amigos que otros sí, es que **Edge Rank** provocó su ausencia por medio de un cálculo de patrón previo.
- **Page Rank de Google**. Es el algoritmo más conocido desde hace años. Permite rastrear millones de páginas web en milésimas de segundo para ofrecer al usuario las mejores soluciones en su pantalla en función de la búsqueda realizada. Un ejemplo lo ven a diario los usuarios de Google en un teléfono inteligente Android o iPhone; al conectarse a Google, y tras ver la ventana inicial y algunos datos (por ejemplo, de tiempo), le aparecen al

usuario un listado de noticias de medios nacionales o internacionales, la mayoría de ellas relacionadas con su perfil de comportamiento en búsquedas de los últimos días o semanas, o bien con experiencias de compras, viajes, etcétera. Google revisa su algoritmo continuamente y cada vez introduce nuevas mejoras que beneficien al usuario pero que, en muchas ocasiones, producen problemas a las empresas, ya que les exige actualizaciones de sus aplicaciones.

- **Recomendaciones de Netflix.** Ofrece, según las preferencias de los clientes, nuevas series o películas que puedan ser de su agrado; sugiere nuevas series, películas o contenidos acordes con su perfil de comportamiento.
- **Recomendación de Amazon.** Nos mantiene *engaged* (comprometido/ocupado) dentro de su sitio web y después nos incita a comprar, usando para ello sus recomendaciones.
- **Waze de Google.** La compañía israelí Waze —adquirida por Google en 2013—, posee un algoritmo que proporciona a cada automóvil la ruta de tráfico más idónea para ahorro de tiempo y congestiones. Cambia los patrones de tráfico en tiempo real, en forma dinámica, con el objetivo principal de proporcionar la mejor opción para evitar tráfico, mejora de rutas, ahorro económico, etcétera. Este algoritmo de Waze se ha comenzado a utilizar en grandes ciudades para diseñar sus propios sistemas de transporte.
- **Google AdWords.** Es la principal fuente de ingresos de la compañía. Se basa en el seguimiento del comportamiento del usuario, uso de las palabras y las consultas de búsqueda para ofrecer publicidad contextual. Se trata de un fantástico algoritmo con el que se consigue esa entrañable sensación de que la publicidad lo persigue.
- **Aplicaciones Fintech.** Las aplicaciones y empresas financieras conocidas como **Fintech** utilizan algoritmos en sus áreas de servicios financieras para conseguir las mejores recomendaciones para usuarios personales o empresas.
- **Un algoritmo consejero de administración.** En 2014, la empresa de capital riesgo Deep Knowledge Venture presentó, en su Consejo de Administración, un algoritmo[27] para que pueda decidir en la compra de participaciones o inversiones.
- **Chef Watson de IBM.** Crea platos únicos de gastronomía, basándose en todos los ingredientes que existen en el planeta.
- **Algoritmos para negociar en el mercado de divisas.** Según *Expansión/WSJ*[28], "el porcentaje de operaciones de divisas que los bancos ejecutan con algoritmos ha aumentado enormemente. En 2014, las órdenes con algoritmos representaban no más del 10% —15% del volumen total de negociación del mercado de divisas—; en 2015 el uso de los algoritmos ha

aumentado notablemente debido a la supervisión ordenada por los gobiernos. Donde más se usan los algoritmos es en las operaciones basadas en el tipo de cambio que se fija cada día y que usan las grandes gestoras de activos como tipo de referencia para las conversiones de divisas. Actualmente los algoritmos se usan para ejecutar aproximadamente el 90% de las órdenes realizadas al tipo de cambio fijo diario".

Otro algoritmo muy sobresaliente, y de impacto en la industria y economía de algoritmos de numerosas empresas del mundo de los negocios digitales y del comercio electrónico, es el de Amazon.

Algoritmo de Amazon. El **algoritmo original**[29]de Amazon asocia cada producto comprado (o visitado) por un usuario con una lista de productos similares, que se obtiene en función de los elementos que han sido adquiridos en un mismo pedido, añadidos a un carrito de compra o almacenados en una lista de deseos (*wish list*). El algoritmo se ha ido modificando y ahora incluye anuncios, tales como lista de libros (dispositivos) más vendidos, clientes que compraron este libro (producto) también compraron este o estos otros, qué otros productos compran estos clientes tras ver este producto, etcétera.

13.9.4 TIENDAS DE ALGORITMOS

Comienzan a aparecer tiendas de algoritmos al estilo de las tiendas o almacenes de *apps* de dispositivos móviles, como las de Apple y Google, al estilo de los servicios de aplicaciones en la nube y que ya hemos definido antes como AaaS (algoritmos como servicio).

Si una empresa tiene un gran volumen de datos y necesita un algoritmo (un programa de *software*) para realizar o mejorar un servicio, en lugar de desarrollarlo puede comprarlo o alquilarlo como un servicio en la nube. Los algoritmos se están volviendo cada vez más precisos y eficientes, y sus objetivos pueden ser muy variados. Ejemplos de aplicaciones usuales son: reconocimiento visual, procesamiento de audio y video, sistemas automáticos de recomendaciones, aprendizaje automático, traducciones automáticas, etcétera.

El precio de cada algoritmo dependerá de las llamadas a las API que el desarrollador realice a través de la plataforma de algoritmos situados en la nube. También, y al igual que sucede con las tiendas de *apps,* los desarrolladores que crean sus propios algoritmos los pueden poner a disposición de otros usuarios, ya sea en forma gratuita o mediante pago.

La tienda más reconocida y de mayor prestigio es Algorithmia.com, que tiene un repositorio grande de algoritmos a disposición de empresas y usuarios personales en su plataforma, y también pone a disposición de los desarrolladores profesionales o aficionados la posibilidad de subir sus algoritmos a la plataforma,

cumpliendo los requisitos establecidos. Esta empresa ha sido desarrollada por el uruguayo Diego Oppenheimer, fundador del proyecto junto con Kenneth Daniel. Uno de los objetivos fundacionales de Algorithmia es "permitir a desarrolladores combinar y comercializar sus algoritmos haciéndolos accesibles al público a través de su tienda de apps para algoritmos; desarrolladores en compañías, universidades y otras organizaciones pueden fácilmente encontrar, probar e implementar estos algoritmos accediendo a la API universal de Algorithmia"[30].

Existen muchas otras tiendas de algoritmos que han emergido en estos últimos años y que buscan, esencialmente, captar talento algorítmico, tanto académico de universidades y escuelas de negocios, como talento profesional en empresas especializadas en algoritmos de aprendizaje automático y aprendizaje profundo. Richard Waters, un columnista de *Financial Times*, ha publicado un artículo ("*Crowdsourced algorithms promise to next big thing*"[31]) donde analiza el fenómeno de las tiendas de algoritmos y donde, además de citar a Algorithmia —referencia obligada—, menciona diferentes empresas que han emergido para aplicar el mismo enfoque del comercio o negocio algorítmico. Las empresas que cita, y que hemos verificado, son Quantopian (`quantopian.com`), Numerai (`numer.ai`) y Quantiacs (`quantiacs.com`). Las líneas de negocio que más destacan en sus sitios web son algoritmos para los mercados financieros y aplicaciones Fintech.

13.10 RESUMEN

El creciente aumento de los grandes volúmenes de datos está originando que las tecnologías de inteligencia artificial estén llegando a todos los sectores de la sociedad y, en particular en el mundo del software, a los sistemas de inteligencia de negocios y de analítica de datos.

A las técnicas tradicionales de aprendizaje automático se han sumado las técnicas del aprendizaje profundo originando ambas junto con las redes neuronales artificiales, la aparición de la inteligencia artificial aplicada (IAI).

El año 2011 es considerada por muchos especialistas como el año de partida de la IAI debido a varios hitos de impacto y resonancia mundial que se produjeron a partir de ese año. El supercomputador IBM Watson ganó en Estados Unidos el popular concurso televisivo Jeopardy! en el que ganó a los dos máximos campeones de este programa. Ese mismo año se presentó el asistente virtual Siri de Apple (un robot virtual o chatbot) que permitía al iPhone realizar tareas diversas sin utilizar el teclado y solo con la voz del usuario. A partir de esas fechas las noticias relacionadas con la inteligencia artificial aplicada continuaron y con resonancia debido a que las tecnologías que facilitaban la IAI iban también llegando al gran público.

El aprendizaje automático (aprendizaje máquina) se puede clasificar en tres grandes categorías: aprendizaje supervisado, no supervisado y reforzado.

El aprendizaje profundo con apoyo de redes neuronales artificiales está produciendo un gran impacto en la inteligencia de negocios y en la analítica de negocios. Las unidades GPU de procesamiento gráfico son uno de los factores clave en el éxito del aprendizaje profundo.

Los robots virtuales en su modalidad de asistentes de voz (*bots* y *chatbots*) son una de las herramientas tecnológicas que mayor impacto están produciendo y producirán en el sector de la inteligencia de negocios y la analítica.

Los algoritmos de inteligencia artificial se han convertido en modelos de negocio de la economía digital. Los algoritmos como servicio son un nuevo modelo de servicios de aplicación inmediata en los negocios.

Las tiendas de algoritmos constituyen un elemento clave en el desarrollo de los sistemas de inteligencia de negocios de las organizaciones y empresas, Algorithmia.com, es una de las tiendas de algoritmos más populares de la Web.

Las principales tecnologías de IA que dominarán el sector en el bienio 2017-2018 según la consultora Forrester serán:

- Generación del lenguaje natural.
- Reconocimiento de voz.
- Hardware optimizado con inteligencia artificial.
- Agentes virtuales (*bots* y *chatbots*).
- Gestión de decisiones.
- Plataformas de aprendizaje profundo.
- Biométricas.
- Automatización de procesos robóticos.
- Analítica de texto y procesamiento de lenguaje natural.

NOTAS

[1] Yosua Bengio, "Aprendizaje profundo". Investigación y Ciencia (edición española de Scientific American), agosto 2016, pp. 48-53.

[2] Sin embargo, el término fue acuñado en 1956 por los científicos John McCarthy, Marvin Minsky y Claude Shannon durante la conferencia de Darthmouth para referirse a "la ciencia e ingenio de hacer máquinas inteligentes", especialmente programas de cálculo inteligentes.

[3] Computing machinery and intelligence. Revista Mind, nº 59, 443-460, Alan Mathison

Turing, 1950. En: Loebner Prize [en línea] http://www.loebner.net/Prizef/TuringArticle.html

[4] *Marc Beniof, presidente y CEO de Salesforce, en un artículo publicado en WEF:[en línea]* https://www.weforum.org/es/agenda/2016/09/la-revolucion-de-la-inteligencia-artificial-se-aproxima-rapidamente-pero-sin-una-revolucion-de-la-confianza-sera-un-fracaso. *16 de septiembre, 2016*

[5] *Noticia del periódico Expansión, 17 de enero, 2017.* http://www.expansion.com/sociedad/2017/01/17/587dfb20e5fdea20048b456d.html

[6] *TechRadar™. Artificial Intelligence Technologies, Q1 2017. AI Technologies Will Augment Your Enterprise Applications, Amplify Your Intelligence, And Unburden Your Employees. January 18, 2017.* https://www.forrester.com/report/TechRadar+Artificial+Intelligence+Technologies+Q1+2017/-/E-RES129161.

[7] *Gil Press. Forbes, 23 enero, 2017. Top 10 Hot Artificial Intelligence (AI) Technologies.* https://www.forbes.com/sites/gilpress/2017/01/23/top-10-hot-artificial-intelligence-ai-technologies/#66f163e91928

[8] *El sitio web de Geoffrey Hinton, uno de los creadores de la teoría de aprendizaje profundo, es excelente y muy recomendable para conocer la historia y evolución del Deep Learning y constituye la investigación más sobresaliente sobre esta disciplina:* http://www.cs.toronto.edu/~hinton/

[9] *Yoshua Bengio. Aprendizaje profundo. Investigación y ciencia. Agosto, 2016. pp 48-53*

[10] https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/

[11] *Jensen Huang. Accelerating AI with GPUs: A New Computing Model. January 12, 2016.*https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/

[12] *Ibid, Huang, 2016.*

[13] *Op.Cit. Bengio (2016)*

[14] *Según IBM, este centro es su mayor inversión en Europa en las dos últimas décadas. Contará con 1.000 expertos en Internet de las cosas y en Industria 4.0. Integrará las tecnologías cognitivas de IBM con IoT, Industria 4.0 en la nube de IBM a través de la plataforma IBM Watson IoT Cloud. [en línea:* https://www-03.ibm.com/press/es/es/pressrelease/48491.wss].

[15] http://cognitiva.la/que-es-ibm-watson/

[16] *Cognitiva es una empresa de soluciones cognitivas especializada en desarrollar las capacidades de Watson, con presencia en 23 países de habla hispana en Latinoamérica. [en línea]* www.cognitva.Ia/que-es-ibm-watson.

[17] *Ibid, pp. 11-12.*

[18] *M. Prieto. Chatbots: los asistentes digitales llegan a las empresas. Expansión, 26/03/2017.*http://www.expansion.com/economia-

digital/innovacion/2017/03/26/58d126e8ca4741cc3b8b45d4.html *Artículo de M. Prieto con excelente y amplia información sobre asistentes digitales. Se extraen datos sobresalientes y prácticos. Recomendable su lectura.*

[19] *Javier G. Fernández, 14 de enero, 2017: [en línea]* http://www.expansion.com/economia-digital/companias/2017/01/14/587940f2e5fdeae1758b461a. html. *14 enero, 2017.*

[20] *MIT Technology Review. Special Edition (enero 2017). 10 Breakthrough Technologies 2016. Lee Gomes. Slack, pp. 34-35.*

[21] *Un grupo de investigadores del Instituto de Investigación en Inteligencia Artificial (IIIA) de la Universidad de Barcelona han desarrollado y puesto en marcha la plataforma Wwowww, que utiliza un algoritmo de recomendación para la compra en línea como una nueva experiencia de cliente. En la actualidad es una plataforma de comercio electrónico. Cuando se presentó en enero de 2015 fue una experiencia de compra de moda en línea (social shopping).*

[22] *Steve Prentice, vicepresidente de Gartner. Los algoritmos están redefiniendo la arquitectura de los negocios.* http://www.computerworld.es/tendencias/los-algoritmos-estan-redefiniendo-la-arquitectura-de-los-negocios

[23] *Yuval Noah Harari. Homo Deus. Breve historia del mañana. Editorial Debate 2016. Profesor y uno de los grandes pensadores de la era actual. Otro libro suyo de gran éxito: Sapiens. De animales a dioses.*

[24] *Ibid. Homo Deus. Libro de lectura recomendada si se debe introducir en el mundo de los algoritmos.*

[25] *Ibid. Steve Prentice, Gartner.*

[26] *George Dvorsk publicó en 2014 un artículo que se ha hecho viral: The 10 Algorithms That Dominate Our World, donde destacaba las características esenciales de diez algoritmos de impacto que había seleccionado en ese momento sobre los algoritmos más populares en diversos sectores. [en línea]* http://io9.gizmodo.com/the-10-algorithms-that-dominate-our-world-1580110464

[27] *Portal Tecnológico Ticbeat:* http://www.ticbeat.com/cyborgcultura/primer-algoritmo-puesto-junta-directiva/

[28] *Chiara Albanese. "Algoritmos para negociar en el mercado de divisas". Expansión/WSJ.* http://www.expansion.com/mercados/2015/10/11/561a7613268e3e787b8b464c.html

[29] *Greg Linden, Bren Smith y Jeremy York. Amazon.com Recommendations. Item-to-Item Collaborative Filtering. IEEE Internet Computing. IEEE Computer Society.Enero-febrero, 2003. pp. 76-80. [en línea]* https://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf

[30] algorithmia.com/about

[31] *Richard Waters. "Crowdsourced algorithms promise to be next big thing". Finantial Times, Friday 24 March 2017. p. 16.*

CAPÍTULO 14

CIENCIA DE DATOS Y CIENTÍFICOS DE DATOS EN INTELIGENCIA DE NEGOCIOS

CONTENIDO

14.1 Introducción
14.2 Definición de Ciencia de Datos
14.3 Disciplinas de Ciencias de Datos
14.4 El proceso de Ciencia de Datos
14.5 El científico de datos
14.6 El perfil del científico de datos
14.7 Herramientas de programación para Ciencia de Datos
14.8 Roles profesionales relacionados con datos
14.9 La Ciencia de Datos en la Inteligencia de Negocios
14.10 RESUMEN

OBJETIVOS

- Iniciar al lector en los principios básicos de la nueva disciplina de Ciencia de Datos (Data Science).
- Conocer las disciplinas relacionadas y necesarias para el aprendizaje en Ciencia de Datos.
- Aprender el proceso básico empleado en la Ciencia de Datos.
- Conocer el nuevo rol del científico de datos, una de las profesiones más demandadas en la actualidad y en el futuro.
- Identificar las herramientas de programación más populares utilizadas en Ciencia de Datos.
- Familiarizarse con las profesiones más relacionadas con la Ciencia de Datos.
- Conocer el impacto de Ciencia de Datos en la Inteligencia de Negocios y cómo obtener ventajas competitivas.

14.1 INTRODUCCIÓN

La **Ciencia de Datos** es una disciplina que se encarga de la extracción de conocimiento a partir de los datos y que se encuentra en plena expansión. El **científico de datos** es aquella persona que practica la Ciencia de Datos y cuya misión es extraer conocimiento de los datos disponibles y transmitir los resultados a los directivos de las organizaciones y empresas, con el objetivo de ayudar en la toma de decisiones para que ésta se eficaz y eficiente.

El científico de datos es un profesional que debe dominar las matemáticas y la estadística, junto con altos conocimientos de algoritmos y programación (lenguajes tradicionales como C/C++, Java, JavaScript, SQL o los más especializados en estadística y aprendizaje automático como R, Python, Scala o Julia), analítica y técnicas avanzadas de ciencias de la computación; es recomendable también tener conocimientos de otras disciplinas, tales como negocios, finanzas o salud.

El científico de datos agrega valor a la toma de decisiones mediante sus habilidades analíticas, y por esta razón, es esencial para numerosas organizaciones tener en su plantilla un científico de datos o un equipo de ciencias de datos. Se describirán las competencias y el perfil del científico de datos.

14.2 DEFINICIÓN DE CIENCIA DE DATOS

La evolución de la **Ciencia de Datos** (*Data Science*) tiene sus orígenes al final de la primera década del siglo XXI, y aunque existen muchas teorías sobre su nacimiento, parece que hay cierto acuerdo en que primero se popularizó el término **científico de datos** (*data scientist*) como un rol profesional experto en

Ciencia de Datos (aunque lógicamente el término **Ciencia de Datos** se lleva utilizando desde años atrás).

Ciencia de Datos es el ámbito del conocimiento que engloba las habilidades asociadas a la extracción del conocimiento de datos, esencialmente *Big Data*, que pueda ser comprendido por los expertos del área del dominio o sector. Incorpora diferentes componentes y se basa en métodos, técnicas y herramientas de numerosos campos, que van desde las matemáticas y estadística, informática (ciencias de la computación), almacenamiento y procesamiento de datos, visualización, reconocimiento de patrones, aprendizaje automático y profundo y algoritmos, entre otros sectores. El objetivo central de la Ciencia de Datos es extraer el significado de datos y la creación de productos de datos.

El término Ciencia de Datos ha adquirido gran notoriedad en los últimos años y se comienza a utilizar de modo intercambiable con Analítica de Negocios e Inteligencia de Negocios moderna, como comienzan a definir la consultora Gartner y Tableau, el proveedor de soluciones de *software* de Inteligencia de Negocios.

Ciencia de Datos es una ciencia multidisciplinar, y existe cierta unanimidad en considerar el diagrama de Venn, creado por Drew Conway[1] —*The Data Science Venn Diagram* (figura 14.1) — como la definición más empleada de *Data Science* y que, afortunadamente, su autor dejó de libre uso con licencia Creative Commons. Drew Conway es un importante científico de datos estadounidense que en 2010 hizo una representación gráfica de las disciplinas o áreas que comprenden la Ciencia de Datos, utilizando un diagrama de Venn. Las tres disciplinas esenciales que conforman la Ciencia de Datos son: **estadística y matemáticas** (*Math & Statistics Knowledge*), **informática y computación** (*Hacking Skills*) y **conocimiento del dominio o experiencia en el entorno** (*Substantive Expertise*), como puede ser el conocimiento que se debe tener del entorno; por ejemplo, sobre la gerencia, la publicidad o los recursos humanos. Por lo tanto, para que una persona desempeñe el perfil de **científico de datos** (experto en Ciencia de Datos), Conway considera que debe ser capaz de desempeñar estas tres actividades y competencias:

- Habilidades informáticas de hacking (experto programador hacker).
- Dominio de las matemáticas y estadística.
- Experiencia comprobada en el entorno (conocimiento del dominio–experiencia sustantiva).

Estadística y matemáticas (*Math and Statistics Knowledge*). Una vez extraídos los datos, el científico de datos deberá tener los conocimientos matemáticos necesarios para poder interpretarlos y procesarlos mediante las herramientas más adecuadas. La formación matemática es de gran importancia en la Ciencia de Datos.

Habilidades y competencias informáticas (*Hacking Skills*). La mayoría de los datos procederán de fuentes heterogéneas de datos; en consecuencia, deberá tener las habilidades necesarias para poder extraer, ordenar, procesar, analizar y visualizar estos datos. Debe crear los algoritmos necesarios, utilizando distintos lenguajes de programación para cada caso concreto.

Experiencia del entorno (*Substantive Expertise*) **o conocimiento del dominio del sector**. Para poder diseñar y desarrollar el análisis masivo de datos en diferentes casos de uso y aplicación, es necesario conocer el contexto. El científico de datos debe tener un alto conocimiento del entorno, que lo motive a plantear nuevos escenarios y crear nuevas hipótesis en las que trabajar, siempre cuidando la calidad de los datos, todo ello con la intención de obtener resultados que terminen incrementando el conocimiento del área de trabajo. Es de gran importancia el conocimiento del dominio de la Ciencia de Datos.

Además del conocimiento de las disciplinas anteriores, Conway también plantea la confluencia de otras áreas. En la Ciencia de Datos confluyen también el aprendizaje automático, la investigación tradicional y una tercera materia relacionada con las habilidades y conocimientos de *hacking*, que ya el mismo autor consideraba que era una zona importante, pero de peligro o de incertidumbre.

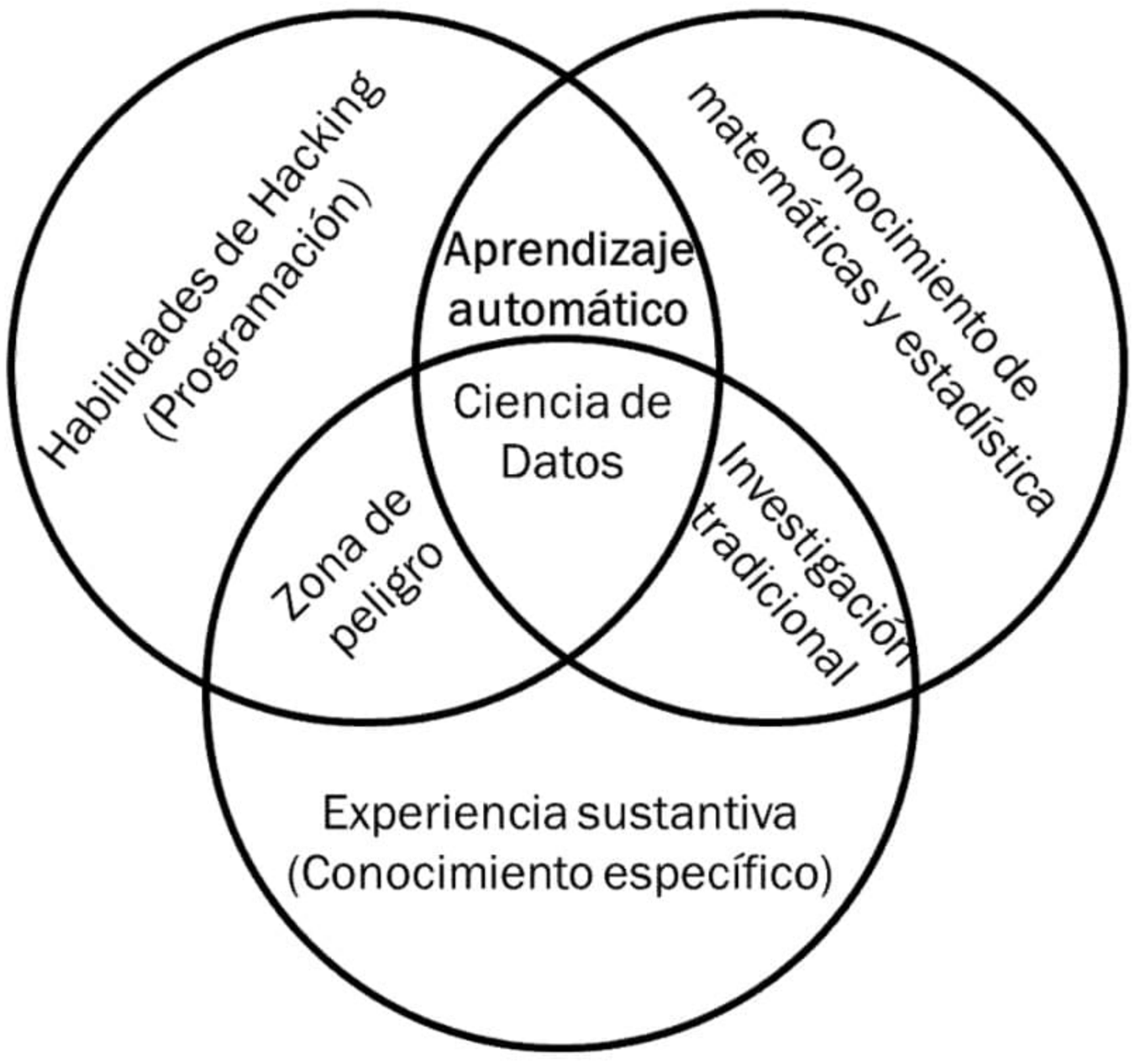


Figura 14.1. Disciplinas de la Ciencia de Datos (licencia Creative Commons) 2010. Fuente: Drew Conway [Traducida y adaptada].
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Como se ilustra en el diagrama de Venn, el científico de datos debe ser competente en las tres áreas básicas descritas anteriormente. Si no se tiene habilidad en alguna de estas áreas, entonces no se considera que estemos hablando de Ciencia de Datos.

El prestigioso portal de software KDnuggets, Data Science, Big Data, Analytics, AI publicó un estudio, sobre la evolución de los diagramas de Venn de Data Science desde el original creado por Conway en 2010 hasta octubre del año 2016. David Taylor (2016).

`https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html`

- **Aprendizaje automático.** El conocimiento de algoritmos de aprendizaje automático permitirá obtener resultados adecuados a los objetivos previstos.
- **Investigación tradicional.** La diferencia entre el científico tradicional y el científico de datos radica, en gran parte, en las habilidades informáticas y conocimientos de lenguajes de programación que debe tener el científico de datos, que le permitirán poder manejar mucha más información y procesarla más rápidamente.
- **Zona comprometida, zona de incertidumbre o zona de peligro.** Un científico de datos que, por ejemplo, no tenga destreza en los campos de la estadística y de las matemáticas, aunque tenga conocimiento del entorno y las habilidades informáticas así como del aprendizaje automático, es probable que procese los datos incorrectamente o los interprete de forma inadecuada y los resultados de la investigación pueden no tener validez, lo que implicará obtener unas conclusiones erróneas, que incluso podrían perjudicar a futuros proyectos que se pudieran basar en estos resultados incorrectos.

En resumen, la Ciencia de Datos, según el diagrama de Venn, es multidisciplinar (matemáticas y estadística, programación *hacking*, experiencia de dominio) y debe tener presentes los principios de investigación tradicional, experiencia comprobada y un área de peligro o incertidumbre.

En la popular enciclopedia *Quora*, donde prestigiosos profesionales e investigadores suelen escribir a modo de preguntas y respuestas (*What is?*), se encuentran un montón de definiciones y la evolución del término *Data Science*[2] de diferentes autores, incluyendo del propio Conway. Se recomienda su lectura si se desea conocer el origen del término, incluso con nuevas definiciones además de las más antiguas de 2010.

Definición de *Data Science* de IBM

IBM define *Data Science*[3] o Ciencia de Datos —ya en fechas más recientes— como el proceso de describir (extraer) conocimiento (*insights*) oculto a partir de cantidades masivas de datos estructurados y no estructurados, utilizando

métodos como estadística, aprendizaje automático, minería de datos y analítica predictiva. Es un área multidisciplinar que está cambiando el modo en que las organizaciones resuelven problemas y ganan ventaja competitiva, y que lo concentra en las tres grandes disciplinas: *Computer Science* (informática), matemáticas / estadística y dominio del conocimiento.

El objetivo de la Ciencia de Datos es la extracción de información útil de un conjunto de grandes volúmenes de datos (*big data*). Las compañías han reconocido el valor de los datos como un activo durante mucho tiempo; sin embargo, las enormes cantidades que ahora están disponibles necesitan nuevos medios para darles sentido y gestionarlos eficientemente. Por esta razón, han comenzado a proliferar ingenieros y científicos de datos que están construyendo sistemas para aplicar la Ciencia de Datos a grandes volúmenes de ellos.

Según IBM, la Ciencia de Datos[4] es una ciencia multidisciplinar que combina las matemáticas y la estadística, la informática o ingeniería de sistemas (*computer science*) y el dominio del conocimiento (figura 14.2). Además de estas áreas de conocimiento, se necesitan otras técnicas de computación avanzada e Inteligencia Artificial, como aprendizaje automático, Minería de Datos, reconocimiento de patrones, almacenamiento de datos, procesamiento avanzado de bases de datos y técnicas y herramientas de visualización de datos.

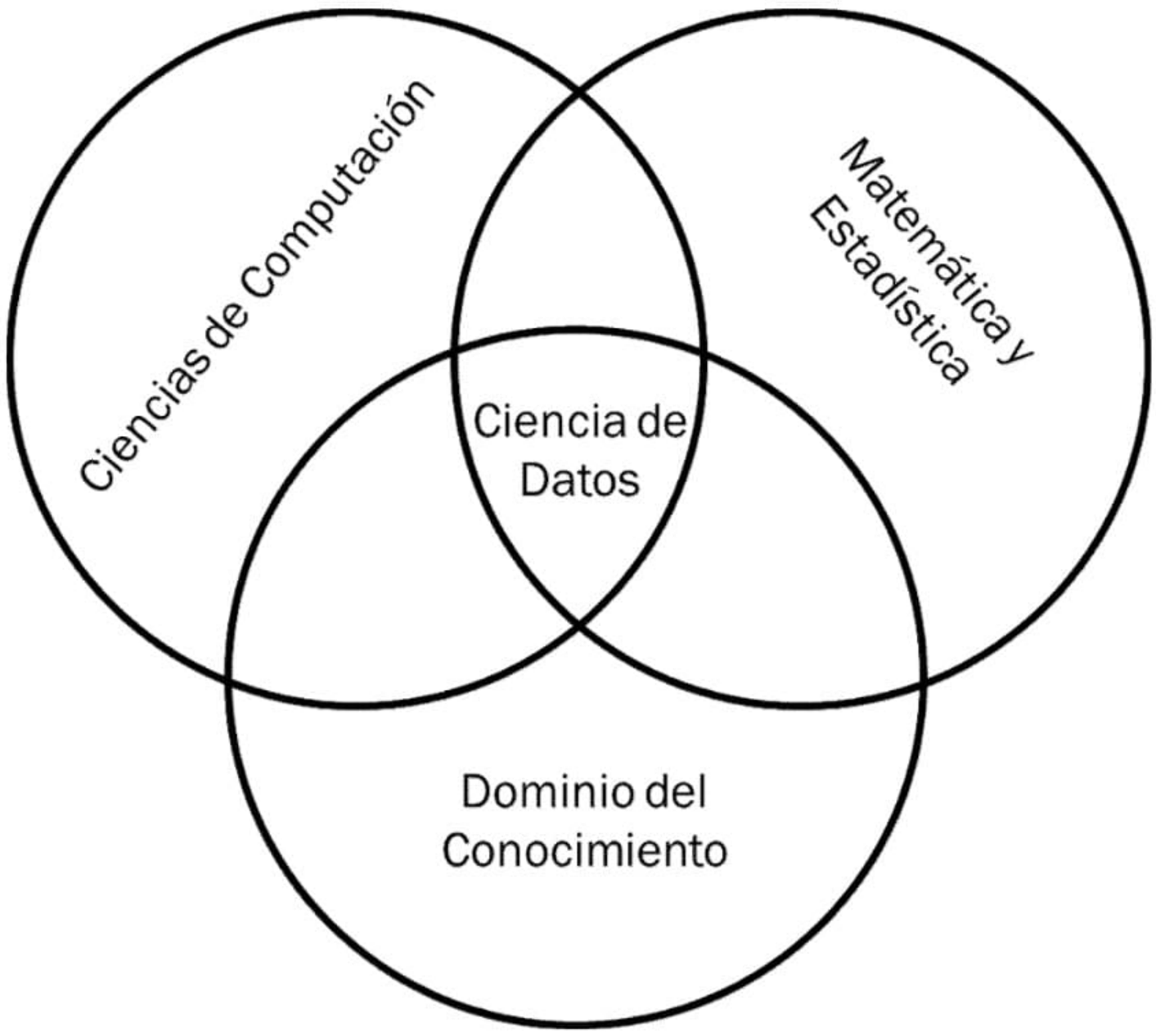


Figura 14.2. Disciplinas de Ciencia de Datos. Fuente: IBM [Traducida y adaptada]:
http://www.ibm.com/developerworks/opensource/library/os-datascience/figure1.png

14.3 DISCIPLINAS DE CIENCIA DE DATOS

Las disciplinas originales de Ciencia de Datos, definidas en el año 2010 por Conway (contempladas en su popular diagrama de Venn), han ido creciendo en cantidad y calidad a medida que las nuevas disciplinas iban llegado a las organizaciones y empresas, como ha sido el caso de *Big Data, Cloud Computing*, Analítica de Datos, etcétera. Las definiciones de Ciencias de Datos y los roles de los científicos de datos han ido evolucionando y, en el año 2017, la Ciencia de Datos se ha convertido en una disciplina autónoma que, como prueba de ello, han aparecido y se imparten cursos en numerosas maestrías, especializaciones y diplomados en *Data Science*, tanto en universidades como en institutos tecnológicos y politécnicos y en escuelas de negocios, con programas adaptados a las diferentes áreas de conocimiento reflejadas en las definiciones más rigurosas y populares. La investigación en Ciencia de Datos también ha aumentado considerablemente en numerosos centros de investigación.

El portal **KDnuggetts** (`www.kdnuggetts.com`), una referencia mundial en *Analytics, Big Data, Data Mining, Data Science* y *Business Intelligence*, con una gran cantidad de documentación rigurosa y de prestigio científico, publicó en octubre de 2016 un excelente y muy completo artículo sobre la evolución de las definiciones más populares de *Data Science* desde la primera publicación de Drew Conway en 2010. El artículo *Battle of the Data Science Venn Diagrams*[5], de David Taylor, describe la evolución de los diferentes diagramas de Venn que Conway y muchos otros autores han ido publicando desde el 2010. Hemos extraído algunos de los modelos más referenciados en la literatura de *Data Science,* aunque existen otras fuentes que recordamos, tales como Quora, Search Data Center, Wikipedia, etcétera.

En 2012, Brendan Tierney[6] publicó una infografía a modo de diagrama de Venn (figura 14.3), donde definió la Ciencia de Datos como una nueva ciencia multidisciplinar apoyada en un conjunto de competencias profesionales exigibles al científico de datos, tales como dominio del conocimiento, comunicaciones, presentación, curiosidad (inquisición), resolución de problemas y análisis del negocio, Las nuevas disciplinas que propone Tierney como espina dorsal de Ciencia de Datos son:

- Estadística.
- Visualizaciones.
- Reconocimiento de patrones.
- Neurocomputación.
- Aprendizaje automático.
- Inteligencia Artificial (AI).

- Procesamiento de datos y bases de datos.
- KDD (descubrimiento del conocimiento en bases de datos).
- Minería de Datos.

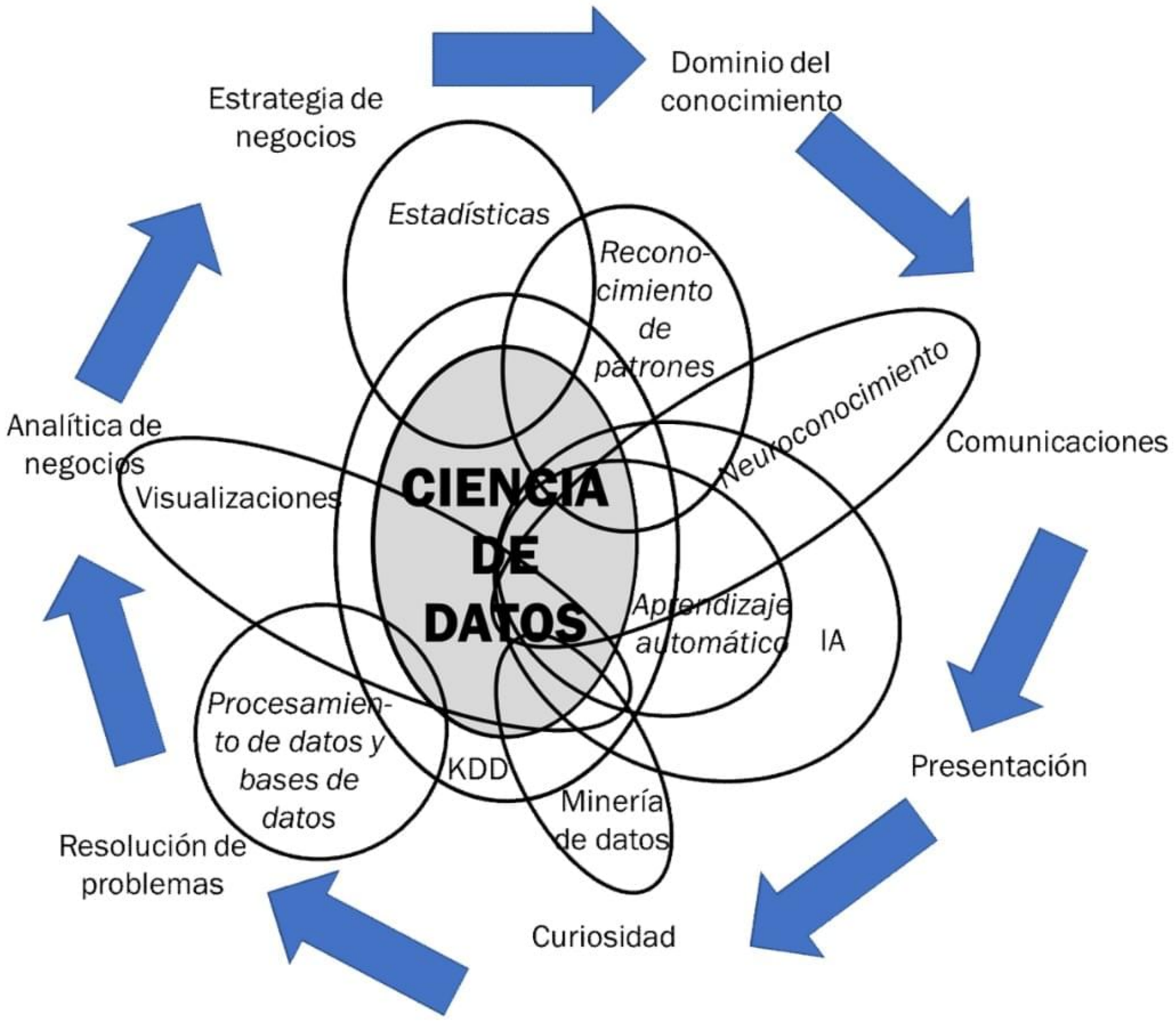


Figura 14.3. Diagrama de Venn de *Data Science* multidisciplinar (2012)
Fuente: Brendan Tierney- Oralytics [Traducida y adaptada]

En enero de 2014, Steven Geringer[7] diseñó un nuevo diagrama de Venn al que denominó **Diagrama de Venn de *Data Science* 2.0**, en el que ya considera la dificultad de definir *Data Science*, por lo que quita el término de la intersección, lo incluye como componente fundamental del diagrama y, en la intersección, inserta un unicornio (el mítico animal) para significar la complejidad de definir el término. Sustituye la habilidad de *hacker* por una disciplina más universal, *Computer Science* (informática, ciencia de computadoras), y sustituye la zona de peligro o de incertidumbre por *software* tradicional.

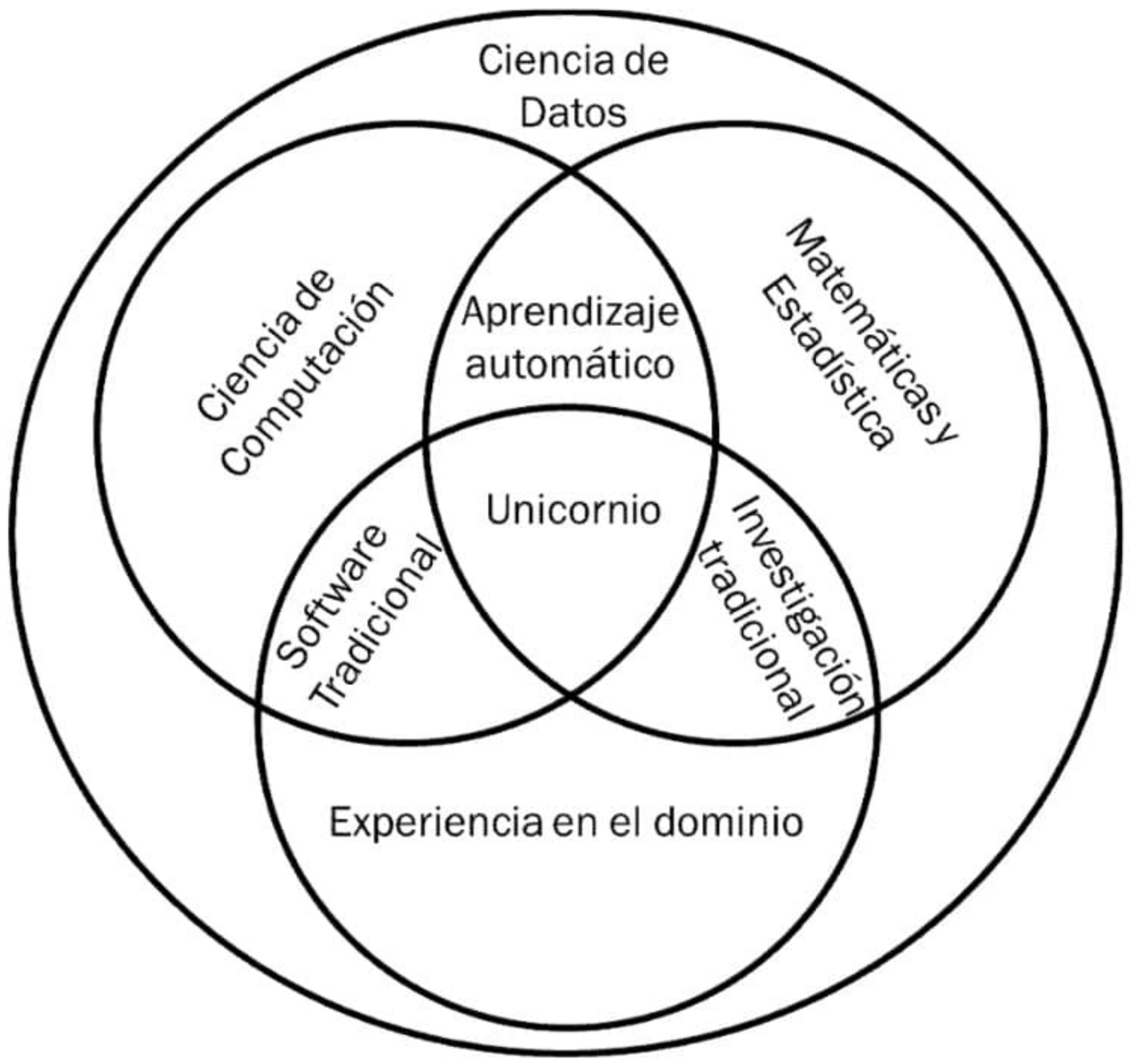


Figura 14.4. Diagrama de Venn de *Data Science* 2.0 (2014).
Fuente: Steven Geringer Raleigh. [Traducida y adaptada]

Un diagrama de Venn de Data Science muy ajustado a la realidad de 2016 es la definición implícita de Matthew Mayo[8] en el propio portal KDnuggetts.

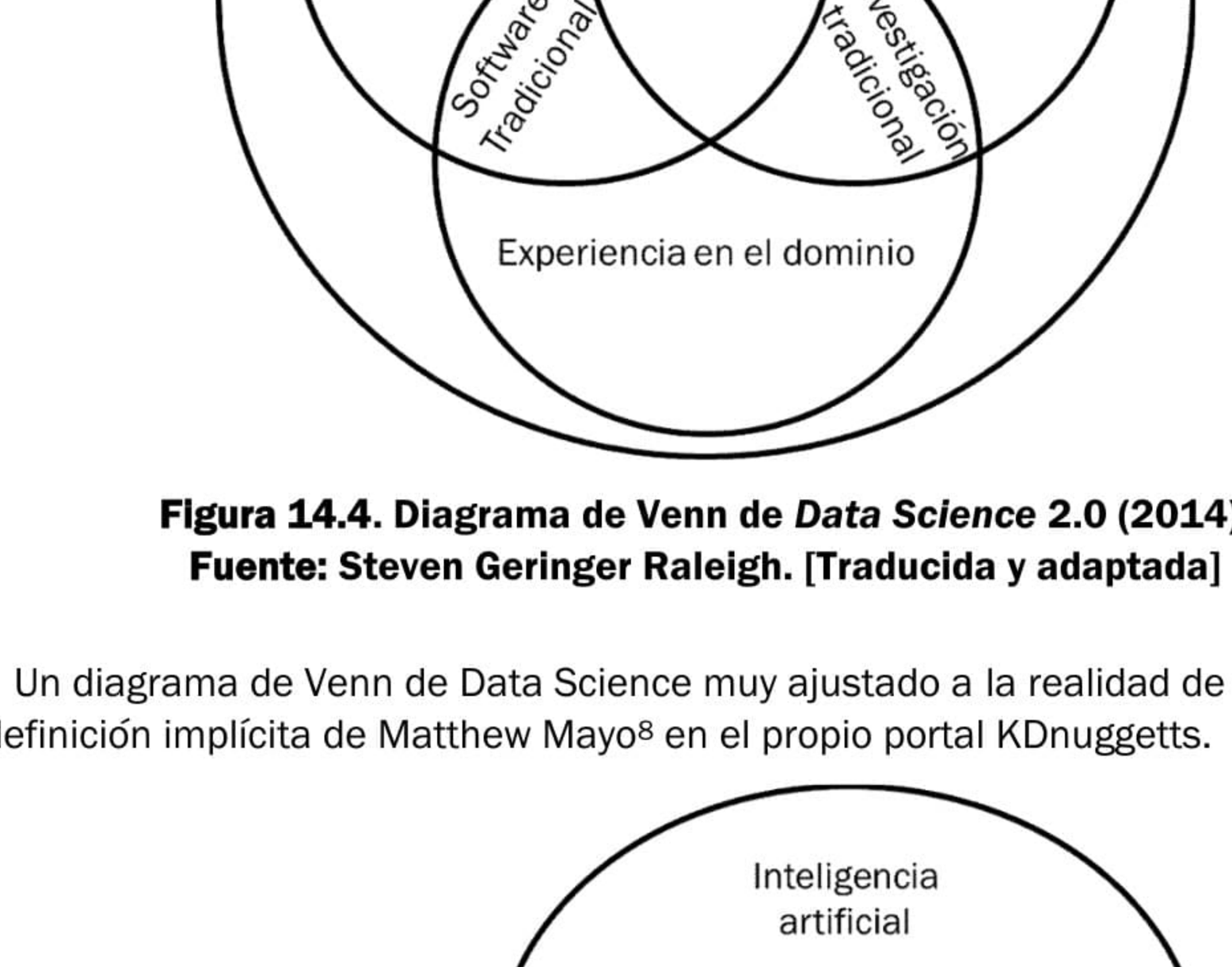


Figura 14.5. Diagrama de Venn de *Data Science* (2016)

Fuente: Mattew Mayo. KDnuggets. [Traducida y adaptada]

Por último, traemos el informe publicado en 2016 por la consultora Gartner Data Solutions[9], donde hizo una definición muy amplia de Data Science actualizando las materias y competencias que comprenden esta disciplina, desde la perspectiva de organizaciones y empresas. Gartner define Ciencia de Datos con un diagrama de Venn muy similar al original de Conway, donde aparecen áreas de gestión empresarial actuales, tales como Analítica Predictiva, gobierno de datos, experiencia de cliente, marketing (asociado a la experiencia de dominio), ingeniería de datos, etcétera.

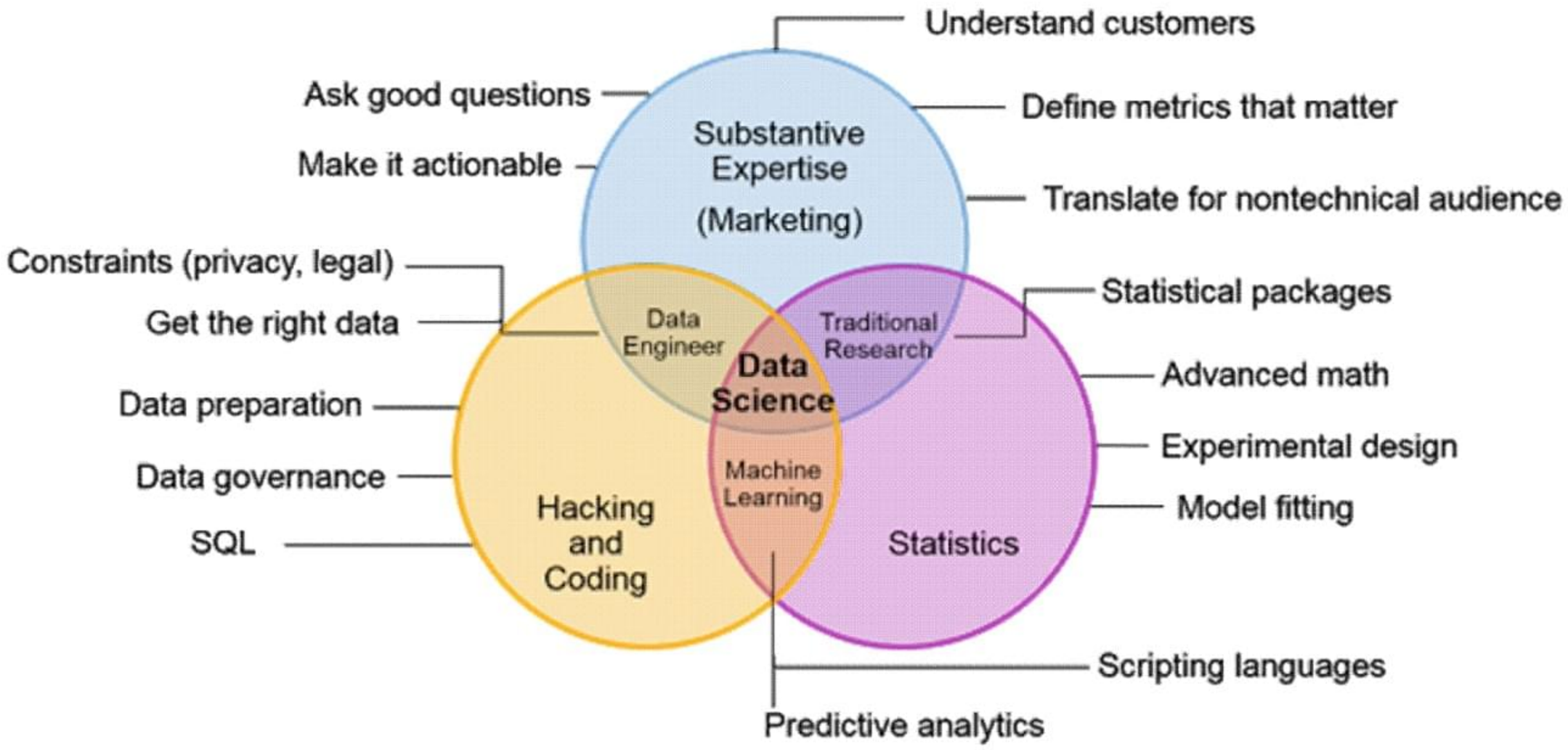


Figura 14.6. Diagrama de Venn de *Data Science* de Gartner (2016)
Fuente: Christi Eubanks. Gartner (original en inglés)

14.4 EL PROCESO DE CIENCIA DE DATOS

Se compone de una serie de etapas que tienen cierta semejanza con procesos similares a Minería de Datos o procesos de Inteligencia de Negocios e, incluso, ingeniería de *software* avanzada.

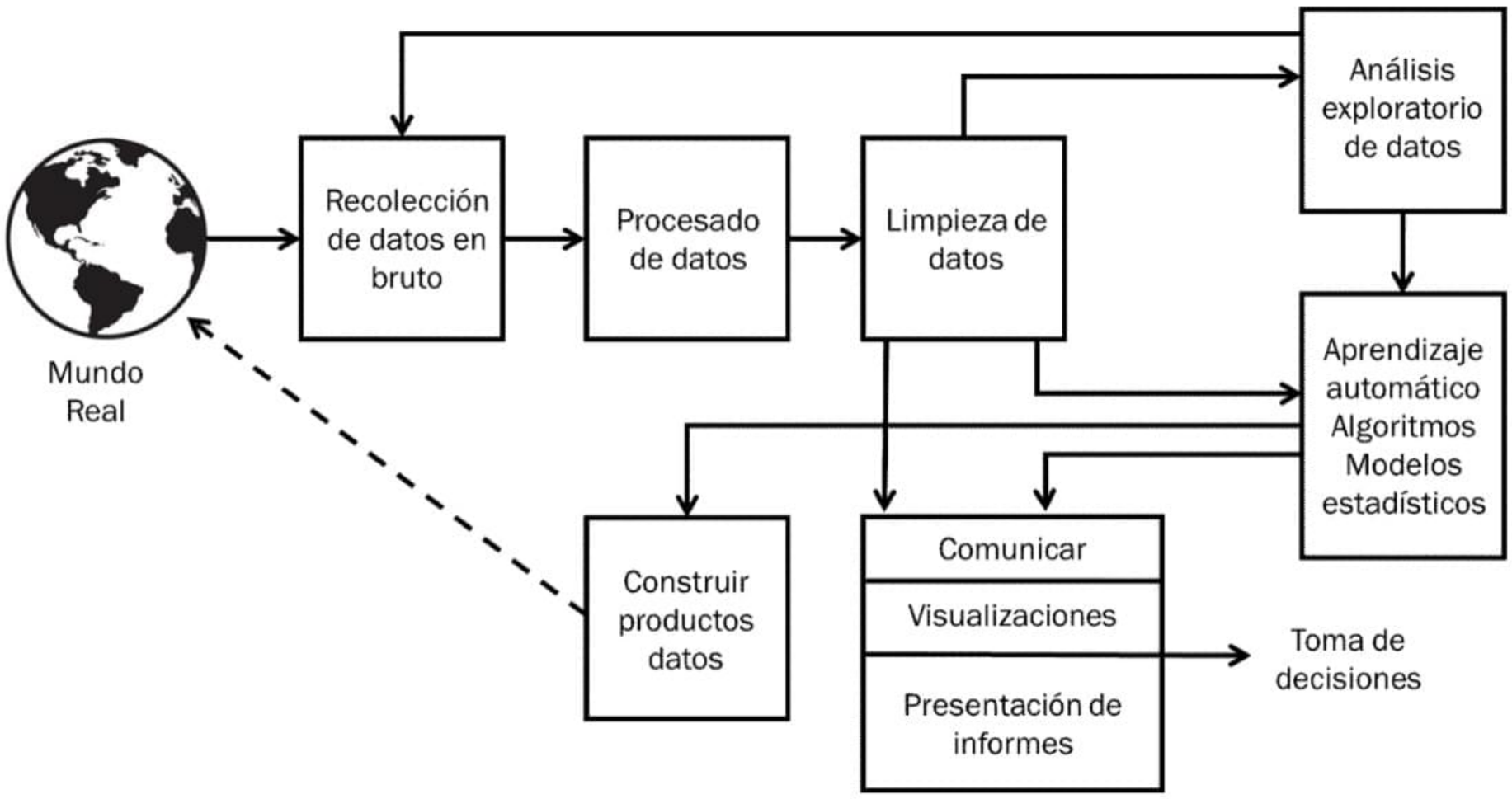


Figura 14.7. Proceso de Ciencia de Datos.
Fuente: Schutt y O´Neill (2014, pág. 41[10]).

Las etapas de un proceso de datos son:

a. Recolección de los datos en bruto, extraídos del mundo real.

b. Procesamiento de datos (herramientas como R, Python, SQL, se utilizan en esta etapa en sincronización con las etapas a y b).

c. Limpieza de datos.

d. Una vez realizada la limpieza de datos, se deberán hacer varias tareas de modo secuencial o en paralelo:

   1. Análisis exploratorio de datos.

   2. Realización de algoritmos de aprendizaje automático y creación de modelos estadísticos. Construcción de prototipos teniendo presente datos del mundo real con las realimentaciones necesarias.

   3. Comunicar, realizar visualizaciones y presentación de informes.

   4. Construir productos datos.

e. Toma de decisiones.

Los datos residen en el mundo real y, como primera etapa, se realiza la ingestión o recolección de los datos en bruto (del entorno). A continuación, se lleva a cabo el procesamiento de datos (normalmente con lenguajes R, Python o SQL) y se dejan limpios para su posterior análisis. Una vez que los datos han sido procesados y limpiados se puede hacer un análisis exploratorio de datos y un modelo estadístico, para lo cual se usan algoritmos de aprendizaje automático. Una vez realizadas estas etapas se puede interpretar, visualizar y comunicar los resultados mediante informes (reportes), presentaciones orales, publicación de

un artículo o noticias en medios de comunicaciones (newsletters) e, incluso, en conferencias o en charlas, y lógicamente se toman las decisiones que se consideren oportunas en los departamentos de la organización afectados.

En estadística, el Análisis Exploratorio de Datos (EDA, Exploratory Data Analysis) es un enfoque para el análisis de conjuntos de datos que resuman sus principales características utilizando, normalmente, métodos visuales. Se puede emplear o no el análisis exploratorio, pero principalmente la tarea más importante que realiza este tipo de análisis es ver todo aquello que los datos pueden decirnos, más allá del modelado formal o la tarea de la prueba de hipótesis. Un modelo estadístico es un tipo de modelo matemático que engloba un conjunto de suposiciones relativas a la generación de algún dato de muestra y datos similares de una población grande; a menudo representa, en un formato idealizado considerablemente, un proceso de generación de datos.

De modo alternativo, según las circunstancias, se puede construir o hacer un prototipo de un producto dato (data product), que no es más que un programa que realiza una tarea determinada, como un algoritmo de búsqueda para realizar rankings de productos, un sistema de recomendación o sugerencias de nuevas compras basadas en el historial de compras. Se pueden hacer nuevos productos datos mediante la realimentación de una nueva captura de datos del entorno del mundo real.

El científico de datos está involucrado dirigiendo cada una de las etapas del proceso de datos. Necesita formular cuestiones e hipótesis y hacer un plan para ver cómo se puede atacar el problema y llevarlo a cabo con su equipo humano, implicándose en el proceso de alto nivel.

14.5 EL CIENTÍFICO DE DATOS

El advenimiento creciente de datos ha conducido a nuevos perfiles profesionales. Aunque son muchas las nuevas profesiones que han ido emergiendo, sin duda el científico de datos es el profesional más reconocido y demandado por organizaciones y empresas que desean gestionar y explotar los datos existentes en las empresas y fuera de ellas.

Un científico de datos es un experto de Ciencia de Datos que resuelve problemas complejos de diferentes sectores (negocios, finanzas, marketing, ciencias de la vida, industria, logística, etcétera), haciendo uso de análisis de datos, y extrae conocimiento de valor de las compañías para una toma de decisiones acertada y eficiente.

El término fue acuñado por D.J. Patil[11], considerado como uno de los grandes científicos de datos actuales, y Jeff Hammerbacher[12] en 2008, que trabajaban en LinkedIn y Facebook, respectivamente.

Así, un científico de datos (data scientist) es un experto que ha de tener formación multidisciplinar para resolver problemas complejos a partir del análisis de datos, extrayendo conocimiento y conclusiones para la toma de decisiones. Cada día se le requiere también conocimiento de negocios, de la Web y, sobre todo, de sociología e incluso de filosofía, esencialmente de ética empresarial. Pero también una formación avanzada en matemáticas, estadística, programación y sus diferentes lenguajes, Analítica de Datos —básicamente, Analítica de Big Data—, aprendizaje automático y visualización de datos.

Científico de datos: la profesión más sexy del siglo XXI (HBR)

Un artículo muy influyente, que se ha convertido en referencia mundial, se publicó en el año 2012 en la prestigiosa revista *Harvard Busines Review*[13]: *Científico de datos. La profesión más sexy del siglo XXI*. Fue escrito por dos expertos mundiales en datos y conocimiento: Tom Davenport (experto mundial en Gestión del Conocimiento y Capital Intelectual) y D. J. Patil (reconocido como unos de los primeros científicos de datos del mundo; de hecho, en 2015, el presidente Obama lo nombró *Chief Data Science* de los Estados Unidos).

Davenport y Patil definen al científico de datos como un profesional que combina conocimientos de matemáticas, estadística y programación de computadoras, que se encarga de analizar los grandes volúmenes de datos (Big Data). Señalan que, a diferencia de la estadística tradicional que utiliza muestra de datos, el científico de datos aplica sus conocimientos estadísticos para resolver problemas de negocio, aplicando nuevas tecnologías que permiten realizar cálculos que hasta ahora no se podían realizar y que comprende todos los volúmenes de datos.

Comienzan su artículo analizando la situación de ese momento producida por el advenimiento de Big Data (el número de HBR estaba dedicado como tema central a Big Data) y la figura de Jonathan Goldman, doctor en ciencias físicas que llegó a LinkedIn en junio de 2006 aplicando métodos de científico de datos —de hecho, por esta circunstancia fue contratado— y convirtió a la red social LinkedIn14 en la referencia mundial en redes sociales y, en particular, de profesionales. A continuación, explican quiénes son los científicos de datos y cómo encontrar al que se necesita en la empresa, en las organizaciones o en las administraciones públicas. Asimismo, describen, a modo de decálogo, cómo encontrar al científico de datos que necesita una compañía y, en consecuencia, sus competencias y características. En este catálogo de recomendaciones figuran desde universidades idóneas para la formación en la disciplina en aquellos momentos —hoy ya no sólo en los Estados Unidos, sino en España y Latinoamérica, proliferan las universidades que imparten cursos y maestrías de Big Data y Ciencia de Datos— hasta características profesionales de los candidatos y, sobre todo, empresas que son casos de éxito, donde los científicos de datos las han convertido en empresas líderes y de referencia mundial. También recomiendan las visitas a sitios de referencia en Big Data y Data Science, como Kaggle o

Descargado en: eybooks.com

TopCoder, asociaciones profesionales especializadas en estas disciplinas. Terminan su artículo explicando la razón del título: el nuevo trabajo "caliente" de la década o "la profesión más sexy".

14.6 EL PERFIL DEL CIENTÍFICO DE DATOS

El científico de datos es uno de los perfiles profesionales más demandados en la actualidad y que más vacantes de empleo está creando en todos los países, dado que es imposible cubrir todos esos puestos, precisamente, por la carencia de profesionales con la formación multidisciplinar que hemos comentado.

Las expertas en Ciencia de Datos, Rachel Schutt y Cathy O´Neill, en su obra de referencia sobre *Data Science* (*Doing Data Science*, O´Reilly, 2014), consideran que el perfil del científico de datos y sus conocimientos y competencias principales necesitan estar implicados en todo el proceso de ciencias de datos (figura 14.7). La función del científico de datos, como muchas otras profesiones, tiene dos campos profesionales: el académico y de investigación y la industria o empresa. En cualquier caso, Schutt y O´Neill (2014) presentan un gráfico (figura 14.8) con los porcentajes de conocimiento que consideran debería tener el perfil del científico de datos en las disciplinas antes comentadas. Estiman que las mayores competencias deben tenerlas en estadística, seguida de matemáticas, *Computer Science* (informática o ingeniería de sistemas) y aprendizaje automático (Inteligencia Artificial).

Schutt y O´Neil (2014)[15] consideran que el perfil del científico de datos (figura 14.8) —posteriormente dedicaremos una sección especial a los roles profesionales relacionados con datos— debe tener niveles de destreza y dominio de las siguientes disciplinas:

- Estadística.
- Matemáticas.
- Informática (Ciencias de la Computación).
- Aprendizaje automático.
- Experiencia de dominio.
- Comunicación y presentaciones.
- Visualización de datos.

Estas autoras insisten en la necesidad del conocimiento de dos disciplinas innovadoras, que son de gran interés en la actual Ciencia de Datos: la comunicación y modos de presentación y la visualización (métodos y herramientas). Las autoras también reseñan en su libro que, dada la dificultad de encontrar un rol profesional que aglutine tantas destrezas de tantas disciplinas,

debería considerarse mejor un **equipo profesional de Ciencia de Datos**, puesto que será difícil encontrar el científico de datos perfecto. Más adelante comentaremos también las opiniones de estas autoras al tratar la función del científico de datos.

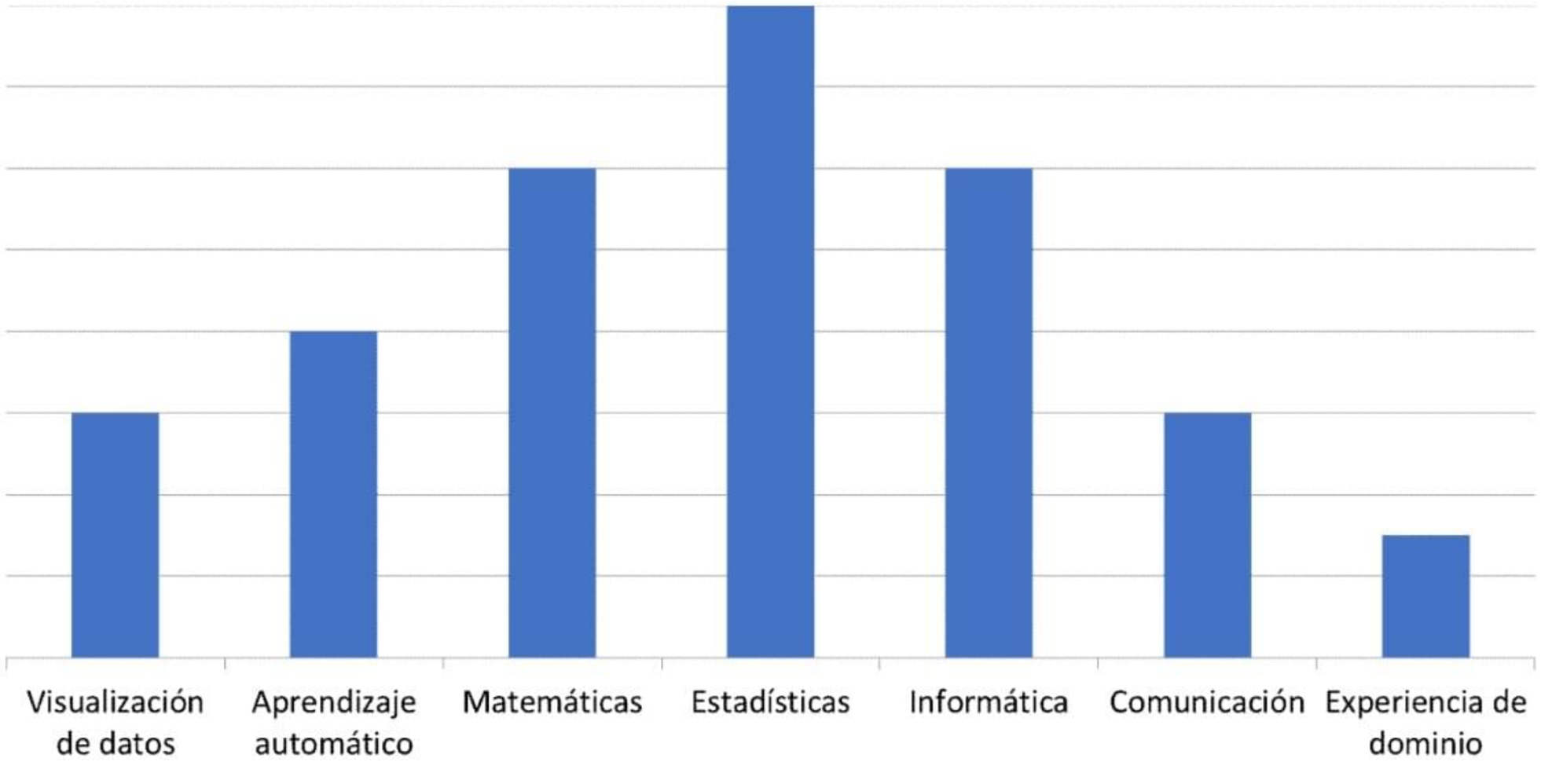


Figura 14.8. El perfil del científico de datos
Fuente: Schutt y O´Neill: 2014

El científico de datos en la academia

En el campo académico, las citadas autoras comentan la situación de la formación en la universidad, tanto en grado como en posgrado, y la imperiosa necesidad de ofertas académicas de las universidades, escuelas de negocios y empresas especializadas en *Big Data* y en *Data Science.* También dedican un tiempo a la investigación y a la propuesta de desarrollo de investigaciones y, en particular, a la elaboración de tesis doctorales (Ph. D.) sobre los temas de Ciencia de Datos y *Big Data*, así como tendencias tecnológicas asociadas.

Además, lógicamente, el científico de datos deberá tener conocimientos en función de su dedicación a todas las tareas específicas del científico de datos en la industria, las cuales comentamos a continuación.

El científico de datos en la industria y en la empresa

Según Schutt y O´Neill, el científico de datos en la empresa ha de establecer la estrategia de datos de la compañía en la que se implican una gran variedad de actividades, desde la ingeniería e infraestructura para recolección de datos hasta cómo decidir el uso de datos para una mejor toma de decisiones. Ha de establecer una comunicación fluida y transparente con los ingenieros, científicos

analistas, así como con los puestos directivos de la empresa, tales como el CEO y el CTO, y con los líderes de productos clave de la misma.

Asimismo, ha de ser una persona que conozca cómo extraer significado e interpretar los datos, que requerirán herramientas y métodos de estadística, aprendizaje automático, así como buenas relaciones personales y comportamiento humano a la altura de su rol profesional.

Técnicamente, requiere el conocimiento profundo de las técnicas de recolección de datos, limpieza y transformación de formatos de datos, ya que los datos en bruto no serán, normalmente, muy limpios y claros. También necesita conocimiento de técnicas y metodologías de ingeniería de *software*. Otras competencias muy importantes son conocer bien las **herramientas de visualización de datos,** para interpretar y producir buenos resultados e informes, y encontrar patrones, construcción de modelos y algoritmos, así como prototipos.

Por último, debe saber comunicarse con los restantes miembros del equipo, ingenieros, matemáticos, responsables de negocio, etcétera, con un lenguaje claro y comprensible, y utilizando eficientemente las herramientas de visualización antes comentadas.

14.7 HERRAMIENTAS DE PROGRAMACIÓN PARA CIENCIA DE DATOS

Las herramientas típicas, o caja de herramientas (*toolbox*), del científico de datos son muy variadas y responden a las necesidades de obtener la máxima eficacia en las múltiples disciplinas que componen la Ciencia de Datos y en los roles profesionales que ha de asumir su figura.

Dado que la programación de computadoras es un componente muy importante, los científicos de datos deben ser muy eficientes con lenguajes de programación tales como R, Python, SQL, Scala, Julia, Java, C/C++, etcétera.

Para las aplicaciones de estadística, matemáticas, algoritmos, modelado y visualización de datos es importante que, además de conocer herramientas de Tableau, Qlik, etcétera, también conozca y utilice bibliotecas y paquetes de *software* existentes —preferentemente de *software* abierto, pero también *software* propietario—, tales como D3, Scikit-learn, e1071, Pandas, Numpy, TensorFlow, Matplotlib, Shiny y ggplot2 (Castrounis 2017)[16].

Cada vez con mayor frecuencia, los científicos de datos deben ser capaces de utilizar herramientas y tecnologías asociadas con *Big Data*, tales como Hadoop, Spark, Pig, Mahout, Hive, etcétera.

Asimismo, y dado que los científicos de datos han de acceder y consultar bases de datos, deben conocer bases de datos relacionales RDBMS (SGBDR), bases de

datos NoSQL y "en memoria" y nuevas bases de datos híbridas (relacionales y NoSQL). Algunas de las más conocidas son PostgreSQL, MongoDB, Redis, HBase, VoltDB, MySQL Cluster, MemSQL, etcétera.

Caja de herramientas de KDnuggets

KDnuggets, el portal de referencia mundial —ya citado varias veces— sobre información profesional y actualizada de *Data Mining, Analytics, Big Data* y *Data Science*, publica cada año una estadística con las plataformas de Ciencia de Datos más utilizadas por las empresas y organizaciones, así como, en particular, los lenguajes de programación usados en desarrollos profesionales o dentro de las plataformas citadas.

El informe de 2018, *Software Poll KDnuggets Analytics/Data Science Tools* presenta las diez herramientas más populares (figura 14.9) en 2018 y realiza también un estudio comparativo estadístico con herramientas más populares de los años 2016 y 2017. Los resultados más destacados son: Python, es la herramienta (lenguaje de programación y biblioteca de funciones) más popular seguida de RapidMiner y R (otro lenguaje de programación); luego siguen herramientas tradicionales como SQL y Excel y destacan Tableau (herramienta, por excelencia, de visualización; ver capítulo 7) y Apache Spark, herramienta de la fundación Apache más eficiente que la popular Hadoop para aplicaciones en tiempo real.

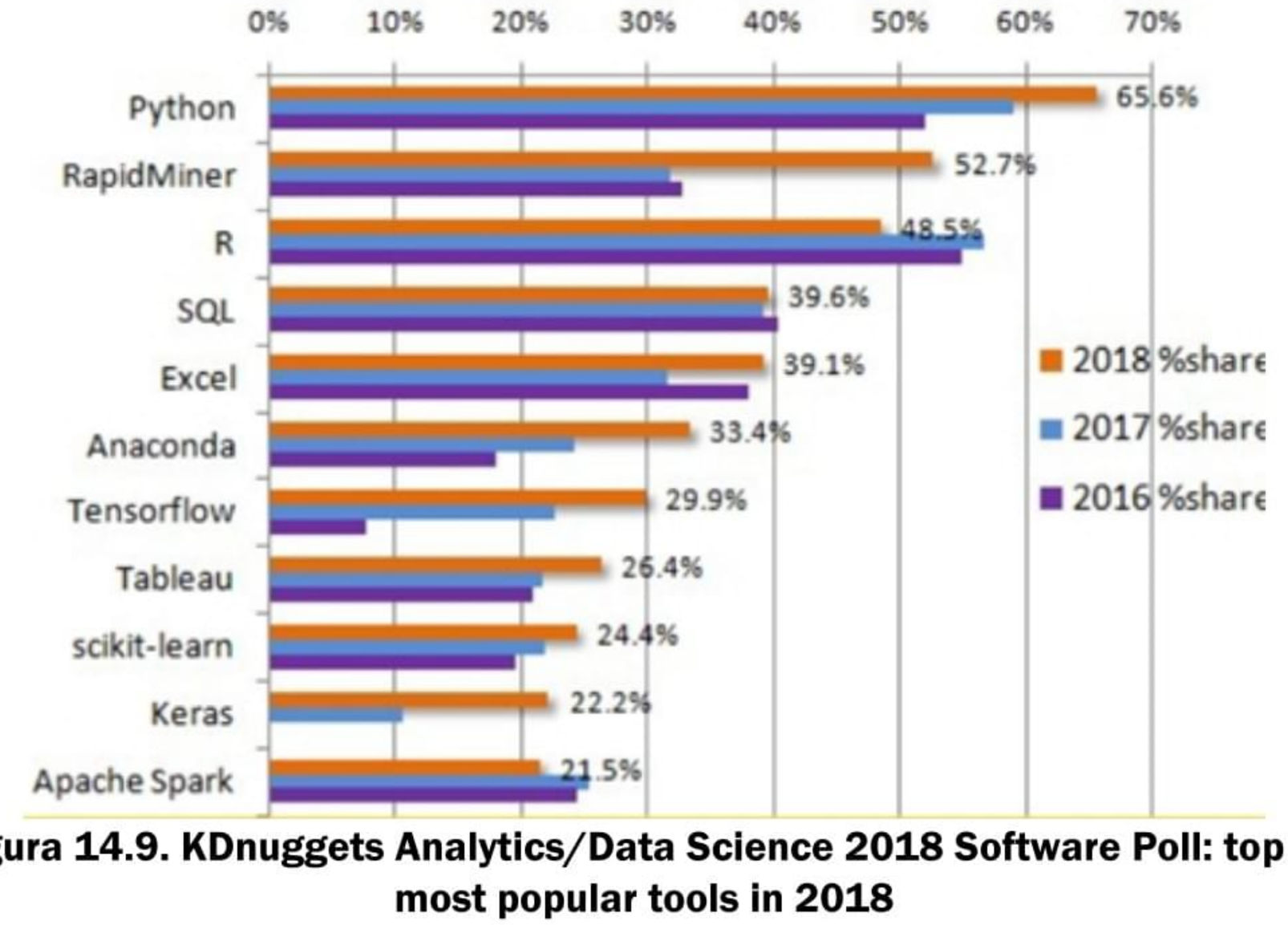


Figura 14.9. KDnuggets Analytics/Data Science 2018 Software Poll: top 10 most popular tools in 2018

Fuente: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

En cuanto a los lenguajes de programación específicos que se utilizan en desarrollos propios o en las plataformas, se destacan R, Python, Java, Scala, herramientas clásicas del sistema operativo Unix, Julia y el clásico C/C++. Tal vez por estas razones, los lenguajes de programación clásicos C/C++, Java y el propio del sistema operativo Unix, se siguen impartiendo en todos los niveles de programación en las carreras de ciencias e ingeniería.

UNA REVISIÓN DE PLATAFORMAS DE CIENCIAS DE DATOS

SQL

Es el lenguaje de programación estándar de las bases de datos relacionales y que debe conocerse profundamente para realizar la integración con los lenguajes especializados en estadística, aprendizaje automático, lenguaje de procesamiento natural y reconocimiento de voz, entre otros. Es un lenguaje normalizado de código abierto que se utiliza mucho en Ciencia de Datos, especialmente integrado con otras plataformas y lenguajes de programación. Además, han aparecido versiones de SQL que están comenzando a ser empleadas en el desarrollo de *Big Data*.

R

Es el lenguaje estadístico por excelencia y existen programas y paquetes de *software* para casi todas las materias utilizadas en organizaciones y empresas. Es una herramienta que se emplea mucho en aplicaciones de Ciencias de Datos y en gestión y desarrollo de *Big Data* y, cada vez más, en aplicaciones de aprendizajes automático y profundo.

PYTHON

Es la otra gran herramienta de Ciencia de Datos, junto con la plataforma R. Una de sus ventajas más importantes es la gran cantidad de bibliotecas y aplicaciones estadísticas. También es un lenguaje de programación fácil de aprendizaje que, unido a sus grandes bibliotecas, lo hace idóneo para el desarrollo de aplicaciones de Analítica de Datos.

KNIME

KNIME (*Konstanz Information Miner*: `www.knime.org`) es una plataforma de Minería de Datos que se usa como herramienta de Inteligencia de Negocios y para Ciencia de Datos. Se creó en la Universidad de Constanza (Alemania) y, posteriormente, se convirtió en una empresa con sede en Zurich (Suiza), donde ofrece servicios generales de desarrollo de la herramienta, formación y consultoría.

WEKA

Es un *software* de aprendizaje automático[17] desarrollado por la Universidad de Waikato, en Nueva Zelanda.

14.8 ROLES PROFESIONALES RELACIONADOS CON DATOS

Los términos relacionados con la Gestión y Análisis de Datos suelen ser muy variables y, en ocasiones, sinónimos, por lo que es difícil asociar las competencias a los muchos roles profesionales que surgen a menudo en el ámbito empresarial o de investigación. No obstante, es frecuente encontrar los siguientes roles: **analista de datos, ingeniero de datos, arquitecto de datos, científico datos, ingeniero de visualización**. Normalmente, el rol profesional va asociado a una, a todas o a algunas de las etapas de la arquitectura de *Big Data* o del proceso de Ciencia de Datos.

Analista de datos. Es el responsable de las etapas de procesamiento y análisis de datos. Su formación fundamental será matemáticas, estadística y economía y negocios (Inteligencia de Negocios y Analítica de Datos). Algunas de las tareas que el analista de datos ha de realizar son: acceso y consultas a diferentes fuentes de datos, proceso y limpieza de datos, resumen de datos, visualizaciones de datos e informes, etcétera. Las herramientas más utilizadas son de Inteligencia de Negocios y Analítica de Datos, tales como Microsoft Excel, Tableau, SAS, SAP, Qlik y MicroStrategy. Pueden realizar, sobre todo cuando alcanzan ya la especialización, herramientas de Minería de Datos, tales como IBM SPSS, Rapid Miner y KNIME.

Ingeniero de datos. Han adquirido gran importancia en la era de *Big Data* y, de hecho, ya tienen hoy en día competencias similares al ingeniero de *Big Data*. El ingeniero de datos no está tan concernido con la estadística, Analítica y modelado de datos como sus homólogos analistas de datos, y está más implicado en la ingeniería y arquitectura de datos, infraestructuras de computación, almacenamiento y flujo de datos, etcétera. Los ingenieros de datos son, por consiguiente, los responsables de la arquitectura de datos y de la instalación de la infraestructura necesaria. Deben tener conocimientos de computación avanzados y deben ser programadores expertos. Recientemente, se les está asociando con unos nuevos perfiles que están emergiendo, los expertos en **DevOps** (desarrollo y operaciones).

Ingeniero de *Big Data*. Son los desarrolladores de ***Big Data***. En general, son ingenieros de *software*, responsables de realizar los programas establecidos por los analistas, científicos de datos y arquitectos de *Big Data*. Serán los encargados de diseñar y construir los algoritmos, los sistemas de recolección y almacenamiento de datos, y realizar los programas de gestión de *software* de la empresa, donde se tengan en cuenta sus planes y líneas de negocio.

Arquitecto de *Big Data*. Es el responsable de toda la arquitectura y proceso de *Big Data*. Deberá tener una visión global del proyecto y el conocimiento de cada una de las áreas necesarias del proceso, desde la recolección de datos hasta la presentación de resultados a través de las herramientas de visualización. Su formación esencial será ingeniero informático o ingeniero de sistemas, normalmente especialista en ingeniería de *software*, pero también con conocimientos de otras ingenierías y de sistemas de información.

Ingeniero de visualización. Dada la importancia que han adquirido las técnicas y herramientas de visualización, se requiere una alta especialización en comunicación, presentaciones y visualizaciones de datos. Requiere una buena formación de ingeniería con una alta especialización en herramientas de visualización. También son demandados especialistas en diseño gráfico, así como en *marketing*, comunicaciones y medios sociales.

Científico de datos. Tiene una visión más horizontal de todo el proceso de Ciencia de Datos. Su tarea principal será la programación de algoritmos para el Análisis de Datos, pero debe conocer bien el negocio de la empresa (su plan de negocio, así como las líneas de negocio fundamentales). En definitiva, un científico de datos debe ser capaz de identificar aquellas variables relevantes para la empresa que ayuden a mejorar resultados, multiplicar el volumen de ventas, fidelizar a los clientes, ahorrar costes, etcétera. Los científicos de datos suelen mezclar, entre otros, conocimientos de matemáticas, estadística e informática, a los que es conveniente unir conocimientos de negocios, administración de empresas, ciencias de la salud y ciencias sociales. El científico de datos se ha consolidado ya como una profesión muy demandada en todo tipo de organizaciones y empresas (grandes y pequeñas). El científico de datos jefe, sobre todo en las grandes multinacionales, se está comenzando a reconvertir en el nuevo rol profesional de director de datos (**CDO,** *Chief Data Officer*), que es el responsable de toda la estrategia y política de datos de las organizaciones y con dependencia directa del presidente o director general.

Así, a los perfiles o roles profesionales ya implantados de Director de Tecnología (**CTO**, *Chief Technology Officer*), con unas competencias muy abiertas y transversales en toda la organización y empresa con independencia del clásico CIO, y también con un alto nivel directivo y estratégico, Director de Informática o de Sistemas de Información (**CIO**, *Chief Information Officer*) y Director de Seguridad de la Información (**CISO**, *Chief Information Security Officer*), hay que sumar dos perfiles profesionales que cada día serán muy demandados en la administración, organizaciones y en empresas de todo tipo:

- ***Chief Data Officer*** **(CDO) o Director de Datos.** Es un cargo dependiente del máximo ejecutivo de la empresa, a quien reporta directamente. Une las dos disciplinas hoy imperantes en la transformación digital: *Big Data y* ciberseguridad. Su misión es impulsar el crecimiento de la organización o empresa mediante la transformación digital de la misma. Ha de tener una amplia visión de la gestión de la empresa y del mundo digital, de Análisis de

Datos y de seguridad de la información, con el objeto de diseñar estrategias y políticas de ciberseguridad. Requiere una formación multidisciplinar de ingeniería, estadística, análisis digital, ciberseguridad, además de una amplia visión y conocimiento de la empresa.

- *Data Protection Officer* (DPO) Director o Delegado de Protección de Datos. Es un perfil jurídico, pero con formación informática y de seguridad de la información y ciberseguridad, aunque también podría ser un tecnólogo con formación jurídica en protección de datos y privacidad. Este perfil es exigible en la administración y en determinadas empresas de la Unión Europea desde el 25 de mayo de 2018, que ha entrado en vigor el Reglamento General de Protección de Datos (GDPR).

También comienza a tener bastante fuerza en grandes y medianas empresas el rol del **CTO** (*Chief Transformation Officer*), con las competencias de director de transformación digital de la corporación y con la responsabilidad de liderar dicha transformación digital y la digitalización necesaria.

14.8.1 CERTIFICACIONES PROFESIONALES EN *DATA SCIENCE*

En la actualidad, existe un gran número de universidades y escuelas de negocio que imparten cursos de maestría, especializaciones, diplomados y cursos especializados en las diferentes materias que componen la Ciencia de Datos.

Asimismo, y en beneficio de usuarios personales y empresas, existen muchos cursos gratuitos en plataformas **MOOC** (cursos masivos en línea y abiertos), donde es posible formarse y obtener diplomas y certificaciones que acrediten una formación avanzada. Las plataformas más populares y acreditadas son **Coursera**, **edX** y **MiriadaX**.

Las empresas distribuidoras de *Big Data* también han iniciado el lanzamiento de certificaciones profesionales que vienen acreditadas por el prestigio de la empresa correspondiente. Así, son cursos muy reconocidos los impartidos por empresas como SAP, IBM, Microsoft, HP, SAS, EMC y Coursera. Otras certificaciones profesionales ya acreditadas son:

- Certified Analytics Professional (**CAP**).
- Cloudera Certified Professional: Data Scientist (**CEP:DS**).
- EMC Data Science Associate (**EMCDSA**).

Además de las acreditaciones anteriores, las organizaciones y empresas valorarán, cada día con mayor intensidad, a los demandantes de empleo y a los propios empleados, las certificaciones internacionales en seguridad como **CISA, CISSP, CISM** o certificaciones propias de los grandes fabricantes y proveedores de soluciones de seguridad de la información, *hardware* y *software* propietario o de código abierto, con estándares como OpenStack y otros.

14.9 LA CIENCIA DE DATOS EN LA INTELIGENCIA DE NEGOCIOS

Las empresas y organizaciones se enfrentan a la implantación de técnicas de Inteligencia de Negocios, Analítica de Negocios (*Analytics*) y Ciencia de Datos, bien de modo independiente o de forma integrada. Las consultoras y empresas especializadas en TI y en negocios utilizan, para sus soluciones de *software* e infraestructuras de los sistemas, los tres términos de modo independiente y, en numerosas ocasiones, utilizan los términos como sinónimos. Nuestra opinión es que Inteligencia de Negocios integra a la Analítica de Negocios, aunque ambos sistemas operan de modo independiente pero integrado en los sistemas de información corporativos.

Inteligencia de Negocios. En la actualidad, podemos considerar dos tipos de BI: Inteligencia de Negocios tradicional o clásica e Inteligencia de Negocios moderna. La **Inteligencia de Negocios tradicional** se ha centrado siempre en el Análisis Descriptivo, Análisis de Datos históricos (transcurridos o pasados). Su fortaleza es la el estudio y el análisis de los datos históricos. En los resultados, se utilizan presentaciones de informes, cuadros de mando (*dashboards*) y visualizaciones interactivas, con una gran oferta de opciones de acceso a los datos y con una lectura y visión intuitivas. La Inteligencia de Negocios, con la generación de informes, la gestión del desempeño, la lectura de información y todos los mecanismos válidos para evaluar el desempeño empresarial, se ha convertido en la tendencia empresarial con mayor penetración de mercado.

Business Analytics* o *Analytics. Se centra principalmente en lo que sucederá en el futuro. Combina análisis estadístico avanzado y modelado predictivo para dar una idea de lo que se espera que suceda, de modo que se puedan anticipar los desarrollos o hacer cambios para mejorar resultados. La Analítica se basa en la capacidad de predecir. La predicción es el atributo principal de *Analytics*, se da en gran medida por la capacidad para realizar proyecciones, segmentaciones, modelamiento de datos, etcétera. La Analítica dispone de herramientas gratuitas, pero se recomienda que, dada la importancia de sus resultados, se opte por realizar estudios previos de soluciones sólidas y robustas, comparándolas con las gratuitas (*open source*), y estudiar su fiabilidad y eficiencia.

En la práctica, Analítica de Negocios es un proceso asistido por tecnologías, mediante el cual el *software* analiza los datos para predecir lo que sucederá (Analítica Predictiva) o lo que podría suceder, tomando un cierto enfoque (Analítica Prescriptiva), Por estas razones, algunas consultoras y proveedores de soluciones de Inteligencia de Negocios denominan también a esta analítica como Analítica Avanzada. El término Analytics es muy utilizado por consultoras y proveedores de soluciones de software.

Sus herramientas ordinarias son informes (reportes), consultas, cuadros de mando (*dashboards*) y alertas, soportadas en entregables fácilmente legibles y comprensibles, tales como gráficos, tablas, infografías, narraciones de datos (*data storytelling*), etcétera.

Ciencia de Datos. El advenimiento de *Big Data* ha implicado la necesidad de la gestión inteligente de datos (esencialmente no estructurados). Se apoya en un desarrollo avanzado de matemáticas, estadística, informática y otras disciplinas, pero sujetos, en la actualidad, a la innovación en medios sociales. Se diferencia de la Analítica en que se puede apoyar en la automatización del análisis de datos, mientras que la Analítica es un proceso basado, principalmente, en las experiencias personales. Una definición muy difundida de Ciencia de Datos se refiere a la disciplina que utiliza métodos cuantitativos de estadística y matemáticas junto con tecnologías (computadoras y *software*) para desarrollar algoritmos diseñados para describir patrones, producir resultados y encontrar soluciones óptimas a problemas complejos. Desde el año 2017 se está consolidando el término de Inteligencia de Negocios moderna (*Modern BI*)

Inteligencia de Negocios moderna *(Modern BI)*

A finales de 2016, la consultora Gartner y el proveedor de soluciones de Inteligencia de Negocios Tableau —uno de los más utilizados en el mundo— comienzan a utilizar en sus informes y estudios el término Inteligencia de Negocios moderna (véase apartado 15.3).

Las organizaciones comienzan a adoptar la Inteligencia de Negocios moderna para referirse a la adopción de plataformas de Analítica Avanzada. El poder del análisis pasa a ser un bien colectivo para muchas personas de la organización, al contrario de lo que sucede en la BI tradicional, donde sólo utilizan unas pocas y especializadas. Las plataformas modernas incorporadas a las soluciones de Inteligencia de Negocios son confiables y escalables, y permiten que las organizaciones capaciten, incluso, a personas que no son expertas en análisis para la explotación de datos administrados y la colaboración en el descubrimiento de resultados. Gartner ya predecía que a partir de 2017 la Inteligencia estaría en todas partes con el uso de herramientas de Inteligencia de Negocio más modernas, con soluciones de *software* innovadoras basadas en tecnologías de Analítica Avanzada.

Louis Columbus[18], un experto en Inteligencia de Negocios que escribe en la revista *Forbes*, explica las características de la Inteligencia de Negocios moderna o avanzada. Estas características han sido extraídas del estudio *Advanced and Predictive Analytics Market Study 2017*, realizado por la prestigiosa consultora Dresner Advisory Service —una comunidad de más de 3.000 organizaciones internacionales—; en este estudio se sostiene que las características y herramientas de las soluciones avanzadas de BI para el año 2017, y sucesivos, se sustentarán en las siguientes iniciativas y tendencias prioritarias:

- Reporting.

- Dashboards.
- Visualización avanzada.
- Autoservicio de usuario final (democratización de la Inteligencia de Negocios: cualquier usuario puede acceder a las herramientas de BI aunque no sea un especialista). En el apartado 15.12 se describirá el concepto de Inteligencia de Negocios de autoservicio como una tendencia de futuro.
- BI cognitivo.
- Analítica de Textos.
- Internet de las cosas.
- Analítica de Video.

Las tres analíticas que habrá que considerar son: Analítica "en memoria" (de bases de datos *in memory*), Analítica en Base de Datos (NoSQL) y Analítica en Hadoop (*Big Data*); las plataformas de BI y *Analytics* modernas deberán incluir estos tres tipos de analítica.

En resumen, la Inteligencia de Negocios moderna suele incluir las dos disciplinas, *Business Intelligence* y *Analytics*, en plataformas especializadas. Ésta es la razón por la que el clásico Cuadrante Mágico anual de Gartner de Inteligencia de Negocios pasó a denominarse, desde la edición de 2016, *Business Intelligence & Analytics Platforms* (Plataformas de *BI & Analytics*).

La Ciencia de Datos ha emergido en estos últimos años como un término multidisciplinar que engloba a numerosas disciplinas, como ya se ha comentado, centradas en el Análisis de Datos, pero unidas al conocimiento del negocio y a estrategias de gestión empresarial, que permitan convertir los datos en conocimiento para una eficiente toma de decisiones, conversión que, en numerosas ocasiones, estará automatizada mediante algoritmos de Inteligencia Artificial. La Ciencia de Datos se puede considerar como una evolución de la Inteligencia de Negocios (BI) y la Analítica de Negocios (BA) en la era de *Big Data*.

Sobre la base de estas consideraciones, Ciencia de Datos se ha convertido, según las grandes consultoras y los proveedores de soluciones de *software de negocios*, en un "término paraguas", que engloba principalmente las tecnologías de *Big Data* e Internet de las cosas, Minería de Datos, *machine learning* (aprendizaje automático), *deep learning* (aprendizaje profundo), Analítica de Datos y visualización de datos, además de los conocimientos ya mencionados de matemáticas, estadística, informática (ingeniería de sistemas) y el conocimiento del dominio o del negocio.

Profesiones relacionadas con los Análisis de Datos

En la actualidad, los roles profesionales más demandados por las empresas en el sector datos se suelen resumir en: analistas o consultores de Inteligencia de Negocios, analistas de datos, científicos de datos e ingenieros de datos.

14.10 RESUMEN

La Ciencia de Datos (*Data Science*) es una ciencia multidisciplinaria que requiere conocimientos de matemáticas y estadística, experiencia de dominio de datos y destreza de *hacking* (computación y desarrollos avanzados). A estas disciplinas se añaden conocimientos profundos de desarrollo de *software* tradicional y aprendizaje automático, además de conocimientos de Minería de Datos y visualización de datos.

Se hace una breve historia de la evolución de la definición de *Data Science*, basada en el diagrama de Venn original, referencia obligada en la disciplina, de Drew Conway —un científico de datos de gran prestigio—, publicado en 2010.

El científico de datos es un profesional especializado en Ciencia de Datos, que dirige todas las actividades de Ciencia de Datos de la compañía. Su perfil requiere una formación multidisciplinar: estadística, matemáticas, ciencias de computación (informática), comunicaciones, experiencia en dominio de datos, aprendizaje automático y profundo, reconocimiento de voz, etcétera.

Los lenguajes de programación y las plataformas más utilizadas en Ciencias de Datos son: SQL, R, Python, RapidMiner, Tableau, Qlik, KNIME y WEKA.

El proceso de Ciencia de Datos consta de las siguientes etapas: recolección de datos, procesamiento de datos, limpieza de datos, exploración de datos/modelos y algoritmos, comunicación, visualización de datos e informes (reportes), realización de productos de datos y toma de decisiones.

Los roles profesionales de Ciencia de Datos son muy variados, teniendo a la cabeza el científico de datos; también aparecen roles como analista de datos, ingeniero de datos, ingeniero de visualización, o los directamente relacionados con *Big Data*, ingeniero y arquitecto de *Big Data*.

Las soluciones de *software* de Inteligencia de Negocios (*Business Intelligence*), Analítica de Negocios (*Business Analytics* o *Analytics*) y Ciencia de Datos (*Data Science*) son tendencias que se han de estudiar para su implantación en organizaciones y empresas. La Inteligencia de Negocios moderna, según las consultoras y proveedoras de *software*, incluye la Analítica de Negocios y la Ciencia de Datos se conforma con una tendencia multidisciplinar, donde la figura del científico de datos sobresale como rol profesional muy importante a nivel corporativo.

GLOSARIO GARTNER

(gartner.com/it-glossary)

Con el objeto de comparar las tres disciplinas (*Business Intelligence, Analytics y Data Science*), ya estudiadas a lo largo del libro, veamos las definiciones que la consultora Gartner da en su *IT Glossary* —referencia obligada en temas de TI— para permitirnos comparaciones de un modo más formal.

Inteligencia de Negocios (*Business Intelligence*). *"Término paraguas que incluye las aplicaciones, infraestructura y herramientas, y mejores prácticas que facilitan el acceso al análisis de información para mejorar y optimizar decisiones y desempeño (rendimiento)".*

Analítica de Negocios (*Business Analytics).* "Se compone de soluciones utilizadas para construir modelos de análisis y simulaciones para crear escenarios, comprender realidades y predecir estados futuros. Analítica de Negocios incluye Minería de Datos, Analítica Predictiva, Analítica Aplicada y estadística y se entrega como una aplicación adecuada para un usuario de negocios. Estas soluciones estadísticas, con frecuencia, vienen con contenidos preconstruidos de la industria, que se centran en un proceso de negocios en la industria (por ejemplo, reclamos, aseguramiento o requerimientos regulatorios específicos)".

Analítica ***(Analytics).*** "Se ha convertido en un término general para una variedad de iniciativas relacionadas de aplicaciones y de Inteligencia de Negocios. Una primera opción, Analítica es el proceso de análisis de información desde un dominio específico, tales como la Analítica de un sitio web; una segunda opción se aplica a la amplitud de capacidades de Inteligencia de Negocios a un área de contenido específico (por ejemplo, ventas, servicios, cadena de suministros, etcétera). En particular, los vendedores de Inteligencia de Negocios para diferenciar sus productos de la competencia. De modo creciente, la Analítica se utiliza para describir el Análisis de Datos, matemáticas y estadística de grupos (*clusters*), segmentos, puntuaciones y predice escenarios que probablemente sucederán".

Ciencia de Datos. Este término no viene recogido en el *IT Glossary*, pero Gartner considera que el concepto es igual y está incluido en el término **Analítica Avanzada** (***Advanced Analytics***)**:** "Es el examen de datos autónomos o semi-autónomos o contenido, utilizando técnicas y herramientas sofisticadas, normalmente más allá de la Inteligencia de Negocios tradicional para descubrir conocimientos (*insights*) más profundos, hacer predicciones o generar recomendaciones. Las técnicas de Analítica Avanzada incluyen, entre otras, Minería de Datos y de Texto, aprendizaje automático, coincidencia de patrones, previsiones (*forecasting*), visualización, Análisis Semántico, Análisis de

Sentimientos, análisis de redes y clústeres, estadística multivariante, análisis de gráficos, simulación, procesamientos de eventos complejos, redes neuronales".

Plataformas de Inteligencia de Negocios (*BI platforms*). Facilitan a las empresas construir aplicaciones de Inteligencia de Negocios, proporcionando capacidades en tres categorías: análisis, tal como procesamiento analítico en línea (OLAP), gestión de metadatos y un entorno de desarrollo.

Servicios de Inteligencia de Negocios *(BI services).* Se ofrecen para diseñar, desarrollar y desplegar procesos empresariales e integrar, soportar y gestionar (administrar) las aplicaciones y plataformas tecnológicas relacionadas. Incluyen aplicaciones de negocios y de infraestructuras para plataformas de Inteligencia de Negocios, necesidades analíticas e infraestructuras de almacenamiento de datos (*Data Warehousing*). Las soluciones incluyen áreas tales como gestión del desempeño corporativo (CPM, *Corporate Performance Management*) y el análisis, además de la plataforma de BI tradicional y almacenes e infraestructuras de datos y áreas de calidad de datos.

NOTAS

[1] *Drew Conway. The Data Science Venn Diagram is Creative Commons licensed as Attribution-NonCommercial.* http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

[2] https://www.quora.com/What-is-data-science

[3] *IBM. What is data science?*

http://www.ibm.com/analytics/us/en/technology/data-science/

[4] *M. Tim Jones. IBM. Data science and open source. IBM. Learn about open source tools for converting data into useful information.* https://www.ibm.com/developerworks/opensource/library/os-datascience/

[5] *David Taylor. Battle of the Data Science Venn Diagrams. Octubre 2016.* http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

[6] *Brendan Tierney - Oralytics Blog.* http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html

[7] *Steven Geringer Raleigh.* http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

[8] *Mattew Mayo. The Data Science Puzzle, Explained. KDnuggetts.* http://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html.

[9] *Christi Eubanks. Gartner. Three Lessons CrossFit Taught Me About Data Science. 26 de mayo, 2016.* http://blogs.gartner.com/christi-eubanks/three-lessons-crossfit-taught-data-science/

[10] Rachel *SCHUTT y Cathy O´NEIL.Doing Data Science. Sebastopol (USA): O´Reilly, 2014. pp. 40-43.*

[11] *//oreily.ly/1aKXJwT. D. J. Patil y Hilary Mason, autores de Data Driven. Creating a Data Culture en la editorial O´Reilly.* http://www.oreilly.com/data/free/files/data-driven.pdf *(descarga gratuita)*

[12] //linkd.in/17sqKNZ

[13] *Thomas Davenport y D, J. Patil. Data Scientist: The sexiest Job of the 21st Century. Harvard Business Review. Octubre 2012.* https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[14] *LinkedIn ha sido comprada a finales de octubre por Microsoft, por la cantidad de 27.000 millones de dólares.*

[15] *Rachel Schutt y Cathy O´Neil Doing Data Science. Sebastopol: O´Reilly, 2014.*

[16] *Alex Castrounis. What is Data Science, and What Does a Data Scientist Do? (Marzo 2017)*

https://www.kdnuggets.com/2017/03/data-science-data-scientist-do.html

[17] *Weka.* www.cs.waikato.ac.nz/ml/weka

[18] https://www.forbes.com/sites/louiscolumbus/#18c5486639e0

CAPÍTULO 15

TENDENCIAS DE FUTURO EN INTELIGENCIA DE NEGOCIOS. PRIVACIDAD, PROTECCIÓN Y SEGURIDAD DE LOS DATOS

(PARTE 1)

CONTENIDO

15.1 Introducción
15.2 Inteligencia de Negocios en la nube: tendencias
15.3 Medidas de seguridad en el ciclo de vida de los datos
15.4 Los riesgos a la privacidad en la Inteligencia de Negocios
15.5 Ética y responsabilidad social de las empresas
15.6 El nuevo reglamento de protección de datos y de privacidad de la Unión Europea
15.7 Revisión general de tendencias futuras en Inteligencia de Negocios

OBJETIVOS
• Conocer las tendencias de la economía digital.
• Conocer las tendencias de Inteligencia de Negocios.
• Analizar el cambio profundo de la Cuarta Revolución Industrial en Inteligencia de Negocios.
• Analizar la necesidad de utilizar la ética en los sistemas de Inteligencia de Negocios.
• Conocer los conceptos fundamentales de la Ley de Privacidad y Protección de Datos de la Unión Europea que entró en vigor el 25 de mayo de 2018.
• Analizar las previsiones del futuro del empleo, las nuevas profesiones y las tendencias en el desarrollo e implantación de sistemas de Inteligencia de Negocios.
• Conocer las tendencias tecnológicas de impacto para 2018 y 2019.

15.1 INTRODUCCIÓN

La Inteligencia de Negocios actual se implanta en las empresas con dos modelos complementarios: **tradicional** o **clásica**, que se centra en los datos históricos y es el modelo más utilizado en la actualidad y desde hace varios años, y que analiza el estado actual de la empresa; y **moderna**, donde se integran las nuevas tendencias de Analítica como Inteligencia Artificial (con asistentes virtuales), Inteligencia de Negocio de autoservicio o democratización del BI, nuevas plataformas, etcétera. Las empresas han de integrar las soluciones de Inteligencia de Negocios tradicional existentes en la actualidad con las nuevas tendencias de la Inteligencia de Negocios moderna.

Las tendencias de Inteligencia de Negocios moderna para 2018, que predicen las consultoras como Gartner y Forrester o empresas proveedoras de soluciones como Tableau, exigen su conocimiento y las técnicas y herramientas necesarias para desplegar en organizaciones y empresas.

La privacidad de los usuarios y las empresas es una gran preocupación en los sistemas de Inteligencia de Negocios. Las empresas europeas deben cumplir el nuevo reglamento de la Unión Europea (GRDP), de cumplimiento obligatorio desde mayo de 2018. Este reglamento es también de obligado cumplimiento para empresas ajenas a la Unión Europea que realizan negocios con empresas europeas.

Las tendencias tecnológicas integradas en la Inteligencia de Negocios, Analítica de Negocios y Ciencia de Datos están aportando el aumento de puestos de trabajo y nuevos roles profesionales con nuevas competencias integradas con las competencias tradicionales. Así, se están consolidando profesiones como analista/consultor de Inteligencia de Negocios, científico de datos o ingeniero de

datos centrado en el desarrollo y despliegue de infraestructuras *hardware* y *software*.

La Inteligencia de Negocios moderna, con la imprescindible integración con la Analítica de Negocios o Analytics y la evolución hacia la Ciencia de Datos, configura el futuro de la empresa y la necesidad de su implantación y despliegue en las organizaciones, junto con las necesarias estrategias para su cumplimiento.

15.2 INTELIGENCIA DE NEGOCIOS EN LA NUBE: TENDENCIAS

La consultora Dresner Advisory Services, especializada en *Cloud Computing* y en Inteligencia de Negocios, viene realizando en los últimos años un informe profundo sobre las tendencias estratégicas de la Inteligencia de Negocios en la Nube (*Cloud* BI): *2017 Cloud Computing & Business Intelligence Market Study Report*[1] (abril 2017, 101 pp. en pdf). El informe requiere un registro previo de cliente, sin embargo, tenemos la suerte de que Louis Columbus, un prestigioso columnista de negocios de la revista *Forbes*, publicó un artículo (*2017 State Of Cloud Business Intelligence*[2]) en el que hacía un resumen excelente y suficiente para conocer las conclusiones más sobresalientes de las tendencias del modelo *Cloud* BI del citado informe, considerado el de mayor impacto en el año 2017, y con total seguridad lo será en los siguientes años.

Las tecnologías e iniciativas estratégicas de la Inteligencia de Negocios se agruparon en 33 categorías. Muchas de estas tendencias ya han sido analizadas con anterioridad en el otro estudio notable que hemos comentado, el de la consultora BARC. Las cinco tecnologías *top* de Dresner son:

- *Reporting* (creación de informes y reportes).
- *Dashboards* (cuadros de mando) personalizados.
- BI de autoservicio para usuario final (*End user self-service*).
- Visualización avanzada.
- *Data Warehousing*.

Las restantes tecnologías e iniciativas estratégicas más destacadas y clasificadas por orden de importancia en el estudio son:

- Minería de Datos, algoritmos avanzados, Analítica Predictiva.
- Descubrimiento de datos.
- Integración con procesos operacionales.
- Soporte de dispositivos móviles.
- *Data Storytelling* (Narración o Historia de Datos).

- BI embebido (incrustado en aplicaciones, portales, etcétera).
- *Big Data* (por ejemplo, Hadop y/o Spark).
- Soporte colaborativo.
- *Software* de código abierto.
- *Software* como servicio en la nube.
- *Social Media Analytics* (Analítica Social, BI Social).
- Internet de las cosas.
- Analítica de flujo continuo de datos (*Data Streaming*).
- Analítica e Inteligencia de posición (localización geográfica).
- Analítica de Texto.
- Consultas *ad hoc*.
- Integración de datos/herramientas de calidad de datos/herramientas ETL.
- Analítica de video.

15.3 MEDIDAS DE SEGURIDAD EN EL CICLO DE VIDA DE LOS DATOS

Los riesgos de seguridad que se pueden producir en un sistema de Inteligencia de Negocios afectan a todas las etapas de su ciclo de vida.

- **Extracción de datos.** Las numerosas fuentes de información de donde se obtienen los datos pueden ser: información sensible (redes sociales, foros, sensores, buscadores, entre otros), actividades diarias que generan datos (uso de las computadoras, teléfonos inteligentes y otros dispositivos), datos especialmente protegidos (datos de carácter personal), etcétera.
- **Almacenamiento de datos.** Al utilizar bases de datos distribuidas en *Big Data*, aparecen los problemas típicos de seguridad de las bases de datos: datos almacenados que no son correctos y producen resultados no deseados o información tratada que no sea comprometida para los usuarios.
- **Procesamiento de datos.** Se han de monitorizar y controlar las alertas que pueden producirse durante el procesamiento de datos.
- **Análisis de datos.** Se presentan los problemas comentados anteriormente sobre calidad, fiabilidad, y veracidad de los datos y los resultados del análisis para una correcta toma de decisiones.
- **Visualización de datos.** Los resultados presentados deberán estar correlacionados con las diferentes etapas del ciclo de vida de los datos.

En cuanto al acceso a los datos, debe asegurarse que ellos procedan de fuentes confiables y que la compartición de datos entre organizaciones o empresas cumplan, estrictamente, las normas de privacidad de las normativas legales vigentes. En caso de ceder datos a terceras partes, se debe controlar a qué información accede esa empresa y que sea, estrictamente, la autorizada.

Entre las medidas de seguridad que se deben tener presente en la administración de datos de los sistemas de Inteligencia de Negocios se han de considerar las siguientes:

- **Anonimato de datos**. Eliminar todo rastro de información que pueda identificar a una persona (por ejemplo, datos médicos o historias clínicas deben estar codificados para que no sean accesibles).
- **Cifrado de datos sensibles**. Sólo aquellas personas autorizadas tendrán acceso a los mismos.
- **Control de acceso y monitorización**. Establecer políticas de control de acceso. Velar porque se exija el cumplimiento de las políticas implementadas.
- **Aplicar soluciones de seguridad**, tales como análisis de riesgos y soluciones de seguridad, utilizando Inteligencia Artificial y análisis de *Big Data*, buscando anomalías o riesgos de seguridad y técnicas de cifrado de datos.

15.4 LOS RIESGOS A LA PRIVACIDAD EN LA INTELIGENCIA DE NEGOCIOS

Big Data e Internet de las cosas. como ya conoce el lector, trae grandes oportunidades a las empresas y a los usuarios en general. Estas grandes oportunidades incluyen el descubrimiento de patrones de comportamiento de los clientes de una organización o empresa, para ayudar a la creación de campañas de *marketing* más personalizadas, un servicio de atención al cliente personalizado, predecir tendencias económicas, obtener mejoras en los servicios prestados por la empresa, etcétera.

Los grandes datos originados por *Big Data* y las múltiples fuentes de datos de las que se alimentan —principalmente Internet de las Cosas y los datos almacenados en la nube—, entrañan un sinfín de oportunidades, pero también grandes riesgos en la privacidad y protección de datos. El análisis de los grandes volúmenes de datos se ha de realizar siguiendo estrictas normas legales y éticas, para salvaguardar la privacidad e intimidad de las personas.

Elena Gil (2016)[3], una reconocida experta en protección de datos y privacidad en el sector de *Big Data*, agrupa los riesgos más importantes de esta tendencia tecnológica (sin considerar los riesgos técnicos) en tres grandes grupos:

1. El riesgo de caer en conclusiones erróneas que nadie revisa: error por azar y error por confusión.
2. El riesgo que para las personas pueda tener la toma de decisiones automatizada sin un sesgo humano.
3. El riesgo para la privacidad de las personas

El primer riesgo se refiere al error por azar y por confusión que se puede producir en el análisis de datos, debido a la posible creación de modelos predictivos no correctos y a correlaciones de variables, y se pueden producir, por ejemplo, por malinterpretación de variables económicas que pueden afectar tomas de decisiones económicas de la empresa o de las administraciones públicas. Gil (2016) plantea que "después de analizar los datos es importante encontrar la verdadera relación entre las variables para poder crear un modelo predictivo. Es decir, es imprescindible poder diferenciar la *causalidad* de la *casualidad*".

El riesgo de la toma de decisiones automatizadas es, en la actualidad (y lo será cada día más a medida que se despliegue la Inteligencia Artificial en organizaciones y empresas). Uno de los más trascendentales para tener en cuenta, debido a su posible imprevisión es la toma de decisiones automatizadas a través de algoritmos sin intervención humana. Si bien en el desarrollo del algoritmo existe la intervención humana —los creadores de los algoritmos—, una vez que se ejecuta, en muchas ocasiones, puede no existir un control humano que compruebe las decisiones, lo que hace que esas decisiones se tomen de modo automático. Gil plantea que la confianza ciega en el algoritmo puede conducir a que las empresas tomen decisiones sobre las personas sin que se puedan conocer las razones en que se han apoyado. *Big Data* aumenta el riesgo relacionado con la toma de decisiones de forma automatizada.

El gran número de decisiones de la vida diaria quedan sujetos a algoritmos ejecutados de forma automatizada. Gil (2016: 53) señala que "el problema surge cuando los datos que son analizados por medio de los algoritmos no son precisos o veraces, pero los individuos no tienen incentivos para corregirlos porque no son conscientes de que están siendo utilizados para tomar decisiones que les afectan".

El gran pensador actual Yuval Noah Harari, en su citada obra *Homo Deus*, plantea los grandes retos que traerán a la humanidad los años futuros y cómo **el mundo va a cambiar radicalmente gracias a los algoritmos, el *Big Data* y la Inteligencia Artificial.** Según Harari, la ciencia converge en un dogma universal que afirma que los organismos, incluido el ser humano, no son más que algoritmos, y que la vida es procesamiento de datos. Pronto los algoritmos nos conocerán mejor que nosotros mismos. ¿Tiene sentido, entonces, que dejemos en sus manos nuestro futuro? ¿Deben tomar ellos nuestras decisiones? La respuesta de Harari es afirmativa.

Esta situación sobre los riesgos de *Big Data* en la privacidad requiere el cumplimiento de disposiciones legales, tal como parece ya contemplar el nuevo Reglamento General de Protección de Datos de la Unión Europea, de forma que no se autoricen decisiones trascendentales para una persona sobre la única base de un análisis automático de datos. Es decir, se han de evaluar la toma de decisiones automáticas que no entrañen supervisión humana.

15.5 ÉTICA Y RESPONSABILIDAD SOCIAL DE LAS EMPRESAS

La computación socialmente responsable es un asunto clave en los negocios actuales, ya que las acciones de los trabajadores de la información afectan a casi todos los departamentos de la empresa y a otros empleados.

La responsabilidad social corporativa o de la empresa se refiere al comportamiento ético y legal de la empresa como entidad. Las leyes definen una sociedad más justa, con comportamientos legales y adecuados, y enfatizan en las acciones de gobierno con respuestas a comportamientos adecuados. La ética se refiere, por el contrario, a principios o estándares morales que ayudan en la guía del comportamiento, acciones y decisiones.

Los dilemas éticos son elecciones difíciles que implican objetivos en conflictos, responsabilidades y lealtades, que pueden ser cubiertos por las leyes o no. Existen numerosas situaciones en que los trabajadores de la información se enfrentan a dilemas éticos; enumeramos a continuación algunas de las más usuales, reconocidas por empresas y trabajadores:

- Visualizar el correo electrónico de los miembros de un equipo o de subordinados.
- Recomendar el envío de listados de correos (*mailing*) de clientes a otros negocios.
- Utilizar un navegador durante el horario laboral para realizar otras tareas comerciales ajenas a la empresa.
- Implementar un sistema que gestione los expedientes de crisis de las empresas (en España son conocidos como ERE).
- Consultas de redes sociales, *blogs*, etcétera, personales durante la realización de la jornada laboral.

15.5.1 EL DILEMA ÉTICO

Schwab (2016: 128-129) al analizar el impacto de la cuarta revolución industrial y tratar el apartado del individuo, dedica una sección especial[4] al planteamiento

de preguntas éticas en la nueva sociedad, al considerar que los avances tecnológicos nos están empujando a nuevas fronteras de la ética. Así, Schwab se hace las preguntas siguientes en torno a la Inteligencia Artificial y la Robótica:

- ¿Qué hacemos? ¿Confiar en el asesoramiento prestado por un algoritmo o en el que ofrecen familiares, amigos o colegas?
- ¿Consultamos a un médico robot, controlado por Inteligencia Artificial con una tasa de éxito en el diagnóstico casi perfecto, o nos quedamos con nuestro médico humano de toda la vida y sus consejos de confianza?
- ¿El desarrollo futuro de la sociedad nos puede conducir a una situación en la cual los mismos seres humanos comiencen a actuar como robots?

Evidentemente, confiamos en el poder predictivo de la Inteligencia Artificial y, en especial, en los algoritmos y técnicas de aprendizaje automático y aprendizaje profundo. Sin duda, la confianza ciega en los algoritmos es uno de los grandes riesgos éticos, tanto a nivel personal como en la toma de decisiones que toman las empresas las empresas, sin que los empleados y clientes puedan saber cómo se han tomado.

Afortunadamente, según Elena Gil (2016), el nuevo Reglamento de Protección de Datos de la Unión Europea ya ha introducido una disposición por la que se prohíbe tomar decisiones trascendentales para una persona sobre la única base de un análisis automático de datos.

15.5.2 LA ÉTICA EN LOS ALGORITMOS

El desarrollo de algoritmos y la puesta a disposición de clientes particulares o empresas entraña también problemas, dado que, en muchas ocasiones, se trabaja sin ningún código ético y se pueden proyectar malos hábitos o conductas, tales como racismo, homofobia, discriminaciones, etcétera.

Es necesario que, al igual que sucede con los desarrollos de software, se requiera el cumplimiento de las normas éticas ordinarias y las conductas reguladas por las asociaciones internacionales de desarrollo de software, como ACM, IEEE, ISACA, etcétera, y que deben ser de obligado cumplimiento.

Evidentemente, el desarrollo de algoritmos dentro de la ya comentada economía de algoritmos planea muchos retos y oportunidades, así como también grandes riesgos éticos. Por ejemplo, en el caso de vehículos autónomos, cuyo funcionamiento principal se basa en algoritmos, ¿cómo se ha de contestar a preguntas acerca del uso del vehículo autónomo, como puede ser el caso de un accidente?: ¿Quién es el responsable? ¿El conductor autónomo, el automóvil, el programador o el algoritmo? Cabe comentar el caso de los Estados Unidos, donde existe una norma reguladora, relativa al caso de los vehículos autónomos, que ha sido promovida por la NTTSA, organismo oficial que controla la seguridad de las

carreteras del país, y que reconoce al sistema automático de conducción (de Google, en este caso concreto) como el conductor.

15.5.3 EL DEBATE ÉTICO DE LA INTELIGENCIA ARTIFICIAL EN LA UNIÓN EUROPEA Y EN ESTADOS UNIDOS

En la Unión Europea y en Estados Unidos, fundamentalmente, se está planteando desde hace varios años y a todos los niveles (político, académico, profesional...) un debate ético sobre el uso de la Inteligencia Artificial

Los gobiernos de Estados Unidos y el Parlamento Europeo han debatido sobre los usos de la Inteligencia Artificial para establecer los límites éticos y legales de esta tecnología, que cada vez afecta más a ámbitos de nuestra vida. Actualmente, una de las esferas que más preocupa es el gran avance de los automóviles autónomos y sin conductor, que toman decisiones por sí mismos y pueden tener consecuencias no deseadas: ¿Hace falta poner límites a los sistemas autónomos de Inteligencia Artificial? ¿Quién es el responsable si algo sale mal? Los eurodiputados están abogando por seguros obligatorios de este tipo de vehículos para indemnizar a las víctimas de los accidentes que provoquen.

En marzo de 2017 un grupo de científicos internacionales expertos en Inteligencia Artificial redactaron y publicaron “La declaración de Barcelona donde se aprobaron las primeras recomendaciones sobre “un uso adecuado y ético” de la inteligencia artificial y proponía medidas para evitar los posibles usos maliciosos de los sistemas basados en inteligencia artificial, en cuya aplicación señalaba “Europa va rezagada” (La Vanguardia, 11 de marzo, 2017).

La Inteligencia Artificial basada en datos requiere de un uso eficiente de las técnicas y analíticas de *Big Data*, junto con una gran potencia informática para alcanzar niveles de rendimiento adecuados; pero estas circunstancias influyen considerablemente en los comportamientos de los algoritmos, de los robots y de las técnicas de aprendizaje automático, y ello requiere de una normalización de las leyes y directivas europeas y de los países restantes, así como planteamientos éticos que se deben exigir a diseñadores, desarrolladores, empresarios, distribuidores y usuarios en general.

15.6 EL NUEVO REGLAMENTO DE PROTECCIÓN DE DATOS Y DE PRIVACIDAD DE LA UNIÓN EUROPEA

El 27 de abril de 2016, el Parlamento y el Consejo Europeo aprobaron en Bruselas el Reglamento General de Protección de Datos (UE 2016/679) RGPD (GDPR en inglés), el cual se centra en el tratamiento de datos personales, y su

libre circulación, y deroga la Directiva 95/46/EC que ejercía como anterior reglamento.

Ha entrado en vigor en todos los países miembros de la Unión Europea el 25 de mayo de 2018 el GDPR. Este nuevo reglamento protege más la privacidad y se da más control a los internautas sobre su información privada tanto en redes sociales como en teléfonos inteligentes, banca en línea, etcétera, de manera que puedan decidir qué información desean compartir.

15.6.1 NOVEDADES DEL NUEVO REGLAMENTO

Estrella Barrionuevo, abogada de la Asesoría Jurídica de Telefónica España, se adelantó a la entrada en vigor del reglamento y publicó un artículo en el *blog aunclicdelastic* (del *blog* de Telefónica, `blogthinkbig.com`) con las novedades más sobresalientes, en su opinión, del nuevo Reglamento de Protección de datos de la Unión Europea. Un resumen de estas novedades es el siguiente[5]:

- Introducción del concepto **pseudonimización** como categoría intermedia entre los datos personales y los datos anónimos.
- En cuanto al consentimiento, se concreta en forma específica que será necesario que el usuario realice una acción afirmativa para consentir, de manera que las casillas **premarcadas** o la inactividad del usuario no constituirá un consentimiento válido, lo que elimina la posibilidad del conocimiento tácito.
- Se modifica la edad por defecto para que los menores puedan consentir por sí mismos y no a través de quien ostente su patria potestad: pasa de los 14 años, que establece actualmente la LOPD, a 16, aunque este límite podrá reducirse por parte de los estados miembros, sin bajar de los 13 años.
- Nuevos derechos para el usuario:
    - Derecho al olvido.
    - Derecho a la portabilidad de los datos personales.
    - Se crea el nuevo rol profesional de **DPO** (*Data Protection Officer*), Delegado o Director de Protección de Datos para administraciones públicas y entidades que traten datos personales a gran escala.
    - Se establece la obligación de realizar análisis de riesgos y evaluaciones de impacto para determinar el cumplimiento normativo.
    - Se amplía la obligación de comunicar las brechas o incidentes de seguridad, tanto a los afectados como a la AEPD, a todos los operadores del mercado que traten datos de carácter personal, en un plazo de 72 horas.

15.6.2 RECOMENDACIONES DE LA AEPD SOBRE EL NUEVO REGLAMENTO

La AEPD (Agencia Española de Protección de Datos) ha elaborado y publicado un documento simplificado, cuyo objetivo es intentar resolver las posibles dudas de los ciudadanos, organizaciones y empresas, relativas a la puesta en marcha del mencionado Reglamento de la UE. El modo elegido para redactar el documento ha sido el de pregunta-respuesta y, para ello, ha seleccionado un conjunto de doce preguntas clave con sus respectivas respuestas, y cuya lectura y consulta recomendamos, dada la relevancia de la fuente: AEPD[6]. Estas preguntas son:

1. La entrada en vigor del Reglamento, ¿supone que ya no se aplica la Ley Orgánica de Protección de Datos Española?
2. ¿Cuál es, entonces, el significado de que el Reglamento haya entrado en vigor?
3. ¿A qué empresas u organizaciones se aplica?
4. ¿Qué implica para los ciudadanos que el Reglamento amplíe el ámbito de aplicación territorial?
5. ¿Qué nuevas herramientas de control de sus datos poseen los ciudadanos?
6. ¿A qué edad pueden los menores prestar su consentimiento para el tratamiento de sus datos personales?
7. ¿Qué implica la responsabilidad activa recogida en el Reglamento?
8. Entonces, ¿supone una mayor carga de obligaciones para las empresas?
9. ¿Cambia la forma en la que hay que obtener el consentimiento?
10. ¿Deben las empresas revisar sus avisos de privacidad?
11. ¿En qué consiste el sistema de “ventanilla única”?
12. ¿Tienen las empresas que empezar a aplicar ya las medidas contempladas en el Reglamento?

15.6.3 EL DELEGADO DE PROTECCIÓN DE DATOS (DPO)

La principal responsabilidad del DPO[7] consiste en garantizar el cumplimiento de la normativa de privacidad y protección de datos de su organización, institución, empresa o corporación. Sus funciones serán independientes del Director de Seguridad y no subordinadas al mismo; su diferencia principal será que el DPO se dedicará en exclusiva a sus funciones y no como hasta ahora, que, normalmente, las tareas de protección de datos recaían en el Director de Seguridad.

El DPO deberá ser designado atendiendo a sus cualidades profesionales y, en particular, a sus conocimientos especializados de la legislación y las prácticas en materia de protección de datos, y a su capacidad para ejecutar los cometidos contemplados en el Reglamento. Así, existe la obligación de contratar un

Delegado de Protección de Datos (**DPO**) en organizaciones e instituciones públicas y en entidades con más de 250 trabajadores. En el caso de entidades con menos de 250 empleados, será obligatorio el DPO cuando necesiten un seguimiento sistemático y periódico de los datos personales tratados para la monitorización o investigación de mercados, análisis de riesgos o datos crediticios o de solvencia patrimonial, así como cuando traten los citados datos como catalogados de especialmente protegidos.

Competencias profesionales

Las competencias de los DPO son:

1. Tendrán que ser profesionales que puedan acreditar formación y conocimientos especializados en materia de protección de datos.
2. Sus funciones, básicamente, serán asegurar el cumplimiento normativo de la protección de datos, haciendo compatible el funcionamiento de la organización, la consecución de los objetivos lícitos y legítimos de su actividad y la garantía del derecho a la protección de datos y la seguridad de la información.
3. Serán los interlocutores necesarios con la Autoridad de Control de la Protección de Datos.
4. Puede establecerse a través de contratación externa o mediante designación dentro de la plantilla de la organización.

Evaluación de impacto en la privacidad

El Reglamento General contempla otro instrumento para garantizar el cumplimiento; se trata de la **evaluación de impacto en la privacidad** (*Privacy Impact Assessments*, PIA), aplicable en el caso de que sea probable que un tratamiento suponga un riesgo elevado para los derechos y las libertades de personas físicas.

El auge de nuevos modelos de negocio, comunicaciones y medios tecnológicos —tales como las tecnologías *wearables*, la expansión del *Internet of things* (IOT), la progresiva implantación de soluciones de cruzamiento masivo de datos o *Big Data*, el procesamiento de datos sensibles de carácter religioso o ideológico, el tratamiento de datos biométricos, la geolocalización, las nuevas fronteras en el ámbito de la ciberseguridad, el *fingerprinting* o la tecnología de reconocimiento facial en redes sociales— dan lugar a nuevos riesgos que pueden tener consecuencias con carácter simultáneo en distintas localizaciones, lo que da valor no solamente al desarrollo de este marco unificado a nivel europeo, sino también a la necesaria existencia de los *Data Protection Officer* en el seno de las organizaciones.

Así, el Reglamento General exige, en su artículo 33, la evaluación de impacto relativa a la protección de datos cuando sea probable que un tipo de tratamiento (en particular, si utiliza nuevas tecnologías, por su naturaleza, alcance, contexto o fines) suponga un alto riesgo para los derechos y libertades de las personas.

Privacidad desde el diseño

Piñar (2016)[8] destaca el principio de privacidad desde el diseño contemplado en el nuevo reglamento de la UE: "La nueva regulación garantizará que la salvaguarda de la protección de datos se incorpore a los productos y servicios desde sus primeros estadios de desarrollo *(Data protection by design).* Se fomentarán las técnicas '*Privacy-friendly*', como la seudoanonimización, para salvaguardar los beneficios de la innovación en *Big Data*, a la vez que se protege la privacidad."

Este principio de privacidad desde el diseño (art. 25.1) significa, según Mañas, que en el diseño de aplicaciones que traten datos personales se tiene que garantizar la privacidad de los mismos desde el principio. Esto implica, por ejemplo, que, en materia de redes sociales, los perfiles de privacidad de los usuarios estarán por defecto cerrados a otros usuarios, debiendo ser el usuario quien los abra a otros.

Mañas destaca también la importancia del consentimiento en las operaciones de tratamiento de datos, que ha de ser "libre, específico, informado e inequívoco y el responsable del tratamiento de los datos deberá poder probar que el titular 'consintió el tratamiento de sus datos'".

15.6.4 GUÍA DE PRIVACIDAD Y SEGURIDAD EN INTERNET (AEPD/INCIBE)

El 7 de octubre de 2016, la Agencia Española de Protección de Datos (AEPD) y el Instituto Nacional de Ciberseguridad (INCIBE) presentaron la *Guía de Privacidad y Seguridad en Internet*[9] con consejos y recomendaciones prácticas para el ciudadano, y cuyo objetivo era ofrecer información práctica sobre cómo reducir esos riesgos a los que pueden exponerse cuando utilizan determinados servicios.

En la presentación oficial de la guía, se destaca que "buena parte de los servicios más populares de la Red se prestan utilizando gran cantidad de información y datos personales que aportan los propios usuarios. La Agencia e INCIBE consideran que la privacidad y la seguridad son aspectos fundamentales que deben tenerse en cuenta para minimizar los riesgos que pueden llegar a producirse en un mundo hiperconectado".

La guía tiene un total de 18 fichas prácticas en las que se abordan temas tales como la importancia de proteger los dispositivos portátiles, generar y gestionar contraseñas, en qué consiste la verificación en dos pasos, cómo realizar copias de seguridad o proteger el correo electrónico, configurar el navegador para que no almacene todos los pasos que se dan mientras se navega por Internet, gestionar la información que se almacena en la nube, configurar el perfil en las redes sociales o dónde recurrir para educar a los menores en un uso seguro de Internet, sin olvidar temas como el *phishing* o la protección de redes WiFi.

Los temas tratados en las fichas se señalan en la tabla 15.1.

Ficha 1 Tus dispositivos almacenan mucha información privada ¿Te habías detenido a pensarlo?
Ficha 2 ¿Por qué son tan importantes las contraseñas?
Ficha 3 ¿Son suficientes las contraseñas?
Ficha 4 No esperes a tener un problema para realizar copias de seguridad.
Ficha 5 ¿Será fiable esta página?
Ficha 6 ¿Tengo obligación de dar mis datos cuándo me los piden?
Ficha 7 ¿Cómo puedo eliminar datos personales que aparecen en los resultados de un buscador?
Ficha 8 ¿Cómo puedo usar el navegador para que no almacene todos los pasos que doy por Internet?
Ficha 9 ¿Quién puede ver lo que publico en una red social?
Ficha 10 Identificando timos y otros riesgos en servicios de mensajería instantánea.
Ficha 11 Toda la información que se publica en Internet ¿es cierta?
Ficha 12 ***Phishing***: el fraude que intenta robar nuestros datos personales y bancarios.
Ficha 13 ¡¿Qué le pasa a mi conexión de Internet?!
Ficha 14 Quiero proteger mi correo electrónico.
Ficha 15 ¿Qué tengo que tener en cuenta si guardo mi información personal en la nube?
Ficha 16 ¿Puedo compartir ficheros por Internet de forma segura?
Ficha 17 No tengo claro para qué está utilizando mi hijo Internet, ¿qué puedo hacer?
Ficha 18 ¿Las pulseras y relojes que miden la actividad física son seguros?

Tabla 15.1. Títulos de las fichas de la Guía de Privacidad y Seguridad en Internet

Fuente: Agencia Española de Protección de Datos/Instituto Nacional de Ciberseguridad

https://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/2016/Privacidad_y_Seguridad_en_Internet.pdf

Cada ficha desarrolla el tema correspondiente con la información esencial e incluye enlaces y contenidos que se encuentran desarrollados en las páginas de

la AEPD y de INCIBE (en la página de la OSI, Oficina de Seguridad del Internauta). En todas las fichas se incluye, de modo muy práctico y de fácil lectura, un catálogo de "Consejos y Recomendaciones".

15.7 REVISIÓN GENERAL DE TENDENCIAS FUTURAS EN INTELIGENCIA DE NEGOCIOS

A continuación, se describe una síntesis de las tendencias futuras y disruptivas de mayor impacto en la inteligencia de negocios y analítica de datos, extraídas de los informes ya analizados y de otros informes publicados por consultoras de prestigio mundial como Gartner, Forrester, IDC, Accenture, etc. y organizaciones internacionales como el WEF (Foro Económico Mundial).

Descubrimiento/Visualización de datos

La visualización y el análisis exploratorio de datos (descubrimiento de datos) para los usuarios de negocios ha evolucionado considerablemente y es una de las tendencias más importantes en BI (IN). El aumento de los grandes volúmenes de datos, su variedad y su disponibilidad en tiempo real requiere de técnicas más eficientes que las tradicionales de tablas y gráficos estáticos, y más orientadas al usuario con formatos de presentación interactivos, ágiles y rápidos de generación. Los *scatterplots*, mapas de árboles, diagramas en red, gráficos en 3D, gráficos interactivos, infografías, animaciones, los modernos *Data Storytelling*, etcétera, son más adecuados para visualizar grandes volúmenes de datos y presentar resultados.

Una nueva tendencia que está surgiendo es la interacción visual y la navegación de datos. La visualización y el análisis visual de datos, unidos a las técnicas de descubrimiento de datos con interfaces de fácil uso, ayudan a los usuarios de los negocios a profundizar en los datos.

El descubrimiento de datos, en particular, se centra en proporcionar soluciones intuitivas a los usuarios de negocios, de modo que puedan crear sus propios modelos de datos y, en su caso, utilizar métodos predictivos o de Minería de Datos para analizarlos. El uso y análisis de datos interactivos y exploratorios es una característica clave del descubrimiento eficiente de datos.

La visualización de la información de la empresa es cada día más variada y se dirige a una mejor comprensión de los datos más complejos.

Inteligencia de Negocios de autoservicio

Es una de las tendencias prácticas de mayor impacto en las competencias de los empleados y en el uso eficiente de las técnicas y herramientas de la Inteligencia de Negocios moderna. La Inteligencia de Negocios de autoservicio (*self service* BI)

es una característica ya muy arraigada en las empresas digitales y está trayendo grandes beneficios a su personal corporativo. Es una forma de Inteligencia de Negocios en la que el usuario tiene la capacidad de cumplir de forma independiente sus propias necesidades de información. Con la Inteligencia de Negocios de autoservicio, los profesionales no técnicos pueden generar sus propios informes, realizar consultas y dirigir sus propios análisis sin la asistencia del personal especializado de TI.

BI de autoservicio se puede definir como el conjunto de herramientas que permiten a los usuarios de negocios tener independencia del departamento de TI en la obtención de la información, pudiendo no sólo extraerla sino analizarla con gran eficacia. En esencia, los usuarios de negocios pueden acceder a la información que necesitan para el desempeño de sus tareas en cualquier lugar donde se encuentren, en el momento que sea necesario y con herramientas fáciles de utilizar. Se trata de asegurar que la Inteligencia de Negocios sea accesible y esté disponible en cualquier momento y lugar. La interactividad, la visibilidad y la compatibilidad entre sistemas, aplicaciones y dispositivos se deben tener en cuenta para que el uso de los datos pueda tener lugar en condiciones óptimas. Así, se ha de garantizar la entrega de los datos de la Inteligencia de Negocios a dispositivos móviles y las tendencias BYOD, pero siendo consciente el empleado de los riesgos que entraña su uso y que deberán ajustarse a las normativas de seguridad dictadas por los departamentos de TI.

Los modelos centralizados y verticales tienden a desaparecer dando paso a infraestructuras más flexibles, adaptables y horizontales que fomentan la colaboración y el intercambio, promoviendo la creatividad en el marco de una cultura basada en la transparencia y objetividad de los datos. En la práctica, este nuevo modelo de BI debe permitir la independencia del usuario.

Existe una tendencia creciente en las empresas para facilitar a los usuarios de negocios —con independencia de su rol corporativo— la construcción o el diseño de sus propias consultas, informes (reportes), interfaces e, incluso, modelos de datos. También se requiere que los usuarios puedan (si lo necesitan) preparar sus propios cuadros de mando (*dashboards*), utilizando componentes de diferentes fuentes, y adaptarlos, ajustarlos y combinarlos para sus necesidades personales. Los departamentos de TI, responsables del funcionamiento de los sistemas de BI, deben diseñar sistemas de modo que los usuarios no deban estar dependiendo del control de dichos departamentos, sino que dispongan de la libertad suficiente para realizar sus competencias profesionales con independencia de dichos departamentos, aunque cumpliendo las normas técnicas y de seguridad que ellos obligatoriamente han de dictar.

Sistemas de Inteligencia de Negocios de autoservicio significa la democratización de los datos utilizados por los usuarios corporativos en función de su rol profesional.

La democratización del *software* de BI de autoservicio significa que, cada vez con mayor frecuencia, los empleados acceden a los informes de BI independientemente del departamento al que pertenecen y sin necesidad de ser directivos o empresarios.

Gestión de Datos Maestros y de calidad de datos

La importancia de estas actividades debe facilitar que los empleados sólo puedan tomar sus decisiones correctas basadas en datos precisos, exactos y de calidad. El objetivo de la Gestión de Datos Maestros (MDM) es conducir e intercambiar los datos maestros, tales como los de clientes o proveedores de datos maestros de productos, en múltiples sistemas de información. Las empresas trabajan con sistemas de información tales como CRM, ERP, SCM, GIS, utilizan servicios web o necesitan mezclar los sistemas para los datos corporativos o cooperar eficazmente con sus *partners*.

En la era de *Big Data* —en la que los datos se incrementan de modo exponencial, son fáciles de generar y se utilizan con gran profusión— se ha de tener la seguridad de utilizar o producir datos de calidad en los nuevos servicios y productos. Las soluciones de *software* deben estar bien definidas, de modo que la calidad de los datos y la Gestión de los Datos Maestros han de ser factores críticos de éxito.

Data Storytelling (Narración de Datos)

Este nuevo concepto, reservado para el periodismo, la publicidad o la política, ha comenzado a ser una herramienta y una técnica muy eficiente para su utilización en Inteligencia de Negocios. La Narración de historias de Datos (Data Storytelling) es una técnica que combina la visualización de datos (Data Visualization o Dataviz) y la narración de historias (Storytelling).

La técnica del *Data Storytelling* está al alcance de todos: personal de *marketing*, financieros, comerciales, directivos y todos aquellos que tienen la necesidad de hacer buenas presentaciones basadas en datos, donde se desee impactar al auditorio (prospectos, clientes, socios, colaboradores), llevándolos al interior de la historia que se presenta: exposición de los resultados financieros de la organización, explicación del lanzamiento de un nuevo producto, presentación de la apertura de una nueva sucursal, por mencionar algunos de los casos donde se puede aplicar esta nueva tendencia.

El Data Storytelling es una nueva técnica de representación gráfica de datos y abarca la combinación de varios elementos clave, tales como datos, imágenes y narrativa. El desafío está en combinar las imágenes correctas y la narrativa con los datos correctos para construir historias animadas parecidas a una película, con datos en movimiento para disfrutar y entender. El profesor Hans Rosling está considerado como el maestro del Data Storytelling y creó una herramienta llamada Gapminder, que permite crear presentaciones como las que él realiza.

Hoy en día existe un gran número de herramientas en el mercado que proponen funcionalidades de Data Storytelling.

Las organizaciones, como ya hemos visto en el proceso de Inteligencia de Negocios, recolectan, procesan, analizan, extraen valor y visualizan los resultados. Ahora se trata de añadir la técnica clásica de comunicación mediante una narración de la historia de los datos. Es una nueva etapa para considerar en la extracción de valor de Big Data.

La Narración de Datos ha surgido como un método sofisticado de explicar el significado y los conocimientos (insights) obtenidos en su análisis, para promover la acción por realizar en la dirección correcta. Las herramientas de Inteligencia de Negocios y de Analítica de Datos son la fuente dominante de las corporaciones, tal como llevamos comentando a lo largo del libro. Se trata ahora de preparar y presentar historias de datos como una herramienta integrada que facilita su interacción. La narración de historias analíticas interactivas mejora la credibilidad de las historias y presenta a los directivos y ejecutivos el cómo ganar en conocimiento para su toma de decisiones eficiente, de un modo más intuitivo que con el análisis estático y visual. En resumen, la Historia o Narración de los Datos pretende dotar a los mismos de una voz propia, que resalte los aspectos más relevantes del valor de los datos.

La herramienta Data Storytelling es una herramienta de visualización muy potente que utiliza narrativas de datos.

Inteligencia espacial y de geolocalización

La Inteligencia Espacial es la capacidad del ser humano de procesar la información en tres dimensiones, de pensar en imágenes. En el mundo empresarial, se aplica este concepto para definir a la Inteligencia de Negocios que incluye datos de geolocalización —es decir, con mapas, ubicación y coordenadas espaciales— y que, por consiguiente, nos muestra los resultados a golpe de vista. Y todo ello en tiempo real. Añadiéndole la geolocalización, existe la posibilidad de verlo todo, por ejemplo en un mapa, a golpe de vista, por lo que es mucho más rápido y sencillo. Esto es lo que llamamos Inteligencia Espacial.

La Inteligencia Espacial, de posición o ubicación, ha sido en los últimos años un mercado de nicho liderado por empresas de servicios de información geográfica (GIS) y de datos espaciales, entre otras, que utilizaban herramientas muy especializadas y que tenían poca o ninguna integración con soluciones de Análisis de Datos o con las herramientas de informes/reportes o cuadros de mando, dado que la mayoría de los datos y que gran parte de los conjuntos de datos incluyen —en la actualidad— datos geolocalizados de posición y algún tipo de información geográfica (por ejemplo, ciudad, código postal, zip, longitud, latitud, etcétera).

Las tendencias actuales de visualización y descubrimiento de datos han generado una demanda para la geovisualización y análisis, de forma que las

soluciones que representan datos de un modo visual son capaces de dibujar datos, mapas u otros objetos para proporcionar conocimientos (*insights*) adicionales.

Los proveedores de *software* ofrecen soluciones con enfoques diferentes para proporcionar inteligencia de posicionamiento en mapas y almacenes de datos o bases de datos NoSQL y "en memoria". La mayoría de los proveedores están ofertando soluciones para soportar diferentes capas de mapas, con información fiable y segura en esos mapas.

La implementación de la Inteligencia de Localización se consigue mediante el análisis de *Big Data*, utilizando un sistema GIS, y se está convirtiendo en una estrategia de Inteligencia de Negocios apoyada en la geolocalización. La geolocalización es la capacidad para obtener la ubicación geográfica real de un objeto, como un teléfono inteligente, un radar o una computadora, conectado a Internet. En esencia, la geolocalización es la consulta real de la ubicación del objeto. La Inteligencia de Localización permite ubicar a los clientes, visualizar su localización en mapas, entender y analizar su comportamiento y las relaciones existentes con ventas, mercadeo y lugares de influencia. Los consumidores y profesionales incorporan tecnologías GIS en la vida cotidiana, de forma que las soluciones de BI deben adoptar esta tendencia.

Las ventajas de la incorporación de la Inteligencia de Localización a una empresa son numerosas:

1. La ubicación está estrechamente relacionada con los temas sociales.
2. Los datos y los mapas pueden ayudar a planificar el futuro del negocio.
3. La ubicación puede potenciar en gran medida la Inteligencia de Negocios y ha evolucionado mucho, ofreciendo nuevas plataformas con tableros y análisis detallados, que permiten a las empresas visualizar y extraer datos específicos para contar con la ventaja competitiva de disponer de una visibilidad precisa y muy fiable.
4. La combinación de la Inteligencia de Negocios con la localización inteligente es fundamental para optimizar la experiencia de cliente e integrarla en las técnicas y herramientas de visualización y Análisis de Datos.
5. La localización inteligente permite segmentar a los clientes reales o potenciales, por código postal, país, etcétera, y, en consecuencia, revisar las tendencias pasadas, presentes y futuras de los clientes en distintas áreas.
6. La aplicación de la ubicación inteligente a la Inteligencia de Negocios permite a las empresas visualizar con exactitud aquellas zonas geográficas donde son más sólidas las ventas, qué región puede ser la más receptiva a ofertas específicas de *marketing* y dónde es necesario mejorar el servicio del cliente.

La Inteligencia de Negocios, unida a la publicidad móvil y el *marketing* en redes sociales y móviles, se beneficia de las herramientas de localización inteligente y, gracias a la ***consumerización*** de GIS, cualquier empresa se puede promocionar a

nivel local. La localización inteligente en los negocios proporciona a las empresas la capacidad de detectar y mostrar la localización asociada a ciertos sectores de información, para tomar decisiones estratégicas asociadas al negocio de una forma más consistente, e incrementa las posibilidades de obtener resultados más positivos.

Aplicaciones de la Inteligencia de Localización para la empresa

- **Tiempo real**. Muchas herramientas permiten la integración de datos de diferentes fuentes que se actualizan en tiempo real.
- **Visualización de alto impacto**.
- **Análisis de relaciones espaciales**. Responder con fiabilidad a cuestiones tales como: ¿Qué calle es más transitada a una hora determinada? ¿Qué lugar del barrio está más equidistante de los proveedores que necesita mi negocio? ¿Dónde ubicar mi negocio? ¿Dónde se da el clima más propicio para nuestro proyecto?
- **Asociaciones espaciales**. Algunas herramientas de geolocalización de datos permiten asociar, mediante líneas y figuras geométricas, distintas capas de información que muestran cuál es la más relevante para nuestro negocio y contextualizar el entorno.
- **Integración con otras tecnologías**. Por ejemplo, aplicaciones de realidad aumentada, cuya integración con plataformas de Análisis de Datos geolocalizados pueden permitir mostrar en tiempo real objetos asociados a información relevante.

Caso de estudio Google Maps

Google Maps facilita la incorporación de la información geográfica a los sistemas de Inteligencia de Negocios. Desde mediados de marzo de 2017, Google Maps permite compartir los datos de ubicación del usuario en tiempo real en dispositivos IOS y Android.

Inteligencia de Negocios Móvil (*Mobile* BI)

El aumento de dispositivos móviles (3.400 millones de usuarios de teléfonos inteligentes en el año 2017) exige que las soluciones de Inteligencia de Negocios estén preparadas para los dispositivos móviles y, sobre todo, con el uso frecuente de la tendencia BYOD (“trae tu propio dispositivo a la empresa o la organización”).

La movilidad es una de las tendencias más destacadas de la transformación digital de las empresas y en consecuencia una rama muy notable de la inteligencia de negocios móvil de soporte para dispositivos celulares.

Inteligencia de Negocios en la nube

Las soluciones de Inteligencia de Negocios en la nube (*Cloud* BI) aumentan considerablemente gracias a sus numerosas ventajas y facilidad de instalación.

Para hacer uso de servicios en la nube, sólo se requiere la suscripción mediante el correspondiente registro y pago de una cuota mensual (u otro periodo), que dependerá del número de usuarios y el volumen de información que requiera el cliente. Con las plataformas de BI en la nube, sólo se paga aquello que se necesita y se consume, por lo que puede personalizarse la aplicación y utilización comenzando con versiones básicas y aumentando a medida que se requiera.

Las ventajas de las soluciones de Inteligencia de Negocios en la nube son las asociadas a la nube (*cloud*), integradas con las aplicaciones de Inteligencia de Negocios del proveedor de las soluciones de *software*:

- Aplicaciones existentes para todo tipo de usuarios y empresas de cualquier tamaño y sector.
- Fácil de utilizar y ágil, ya que no se requieren conocimientos previos, y la formación suele ser rápida y sencilla.
- Rápida implementación. Según la aplicación, la duración media puede ser de días o de semanas.
- Pago por uso. Precio accesible y adaptado al cliente y a todo tipo de negocios. El modelo de *software* como servicio (SaaS) normalmente requiere una cuota mensual que, como se comentó anteriormente, dependerá del uso previsto.
- Reducción de costes en tecnologías de la información en la empresa.
- Más escalabilidad (posibilidad de ampliar o reducir las prestaciones contratadas).
- Toma de decisiones basadas en la realidad.
- Conocimiento de la empresa y gestión efectiva en tiempo real.
- Mayor capacidad para obtener datos en tiempo real, lo que facilita la colaboración en cualquier ubicación.
- Identificar tendencias, comparando con periodos anteriores o previsiones futuras.
- La computación en la nube es mucho más segura y los datos y la información almacenados en ella son accesibles desde cualquier lugar, en cualquier momento y con cualquier dispositivo.

La mayoría de las empresas proveedoras de soluciones de *software* de Inteligencia de Negocios, tanto de herramientas particulares como de soluciones integradas, ofrecen soluciones en la nube, tanto para grandes como para pequeñas y medianas empresas.

Además de las ventajas anteriormente citadas, y desde un punto de vista estricto de Inteligencia de Negocios, los indicadores positivos de BI en la nube son:

- Posibilidad de almacenar grandes volúmenes de datos.
- Análisis predictivo de alta precisión y eficacia.

- Velocidad de realización de consultas e informes.

Aunque la mayoría de las ventajas compensan claramente la aceptación de las soluciones de la Inteligencia de Negocios en la nube, también existen riesgos por motivos de seguridad, por lo que la empresa deberá realizar un estudio previo y fiable antes de tomar la decisión de migrar a la nube.

Contenido disponible online

Encontrará:

- Parte 2 del capítulo

- Bibliografía y Recursos de la obra.

NOTAS

[1] http://dresneradvisory.com/products/2016-cloud-computing-business-intelligence-market-study-report. *(abril 2017, 101 pp. en pdf).*

[2] *Louis Columbus, 9 abril 2017. 2017 State Of Cloud Business Intelligence, Forbes.*

https://www.forbes.com/sites/louiscolumbus/2017/04/09/2017-state-of-cloud-business-intelligence/#1f66901b45c2

[3] *GIL, Elena. Big data, privacidad y protección de datos. Madrid: AEPD/Agencia Estatal BOE, 2016. El excelente libro de Elena Gil, consiguió el Accésit 2015 de la XIX edición del Premio Protección de Datos personales de investigación de la Agencia Española de Protección de Datos (AEPD).*

[4] *Klaus Schwab. Op. cit., pp. 128-129. Cuadro H. En el límite ético.*

[5] aunclicdelastic.blogthinkbig.com/nuevo-reglamento-europeo-de-proteccion-de-datos- *[consultado el 27 de julio, 2016]*

[6] www.agpd.es

[7] *Efren Díaz Díaz.*
http://tecnologia.elderecho.com/tecnologia/privacidad/Protection-Officer-DPO-Reglamento-Proteccion-Datos-UE_11_945055002.html

[8] *José Luis Piñar Mañas, catedrático de Derecho Administrativo y ex director de la AEPD, en un artículo publicado en:*
http://noticias.juridicas.com/actualidad/noticias/11050-contenido-y-novedades-del-reglamento-general-de-proteccion-de-datos-de-la-ue-reglamento-ue-2016-679-de-27-de-abril-de-2016/

[9]https://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/2016/Privacidad_y_Seguridad_en_Internet.pdf

Descargado en: eybooks.com