tecnología

INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS

UNA VISIÓN GLOBAL DE BUSINESS INTELLIGENCE & ANALYTICS

Luis Joyanes Aguilar

Alfaomega

Marcombo

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

Marcombo

Inteligencia de negocios y analítica de datos
Luis Joyanes Aguilar

Primera edición: 2019

Primera edición: MARCOMBO, S.A. 2019

www.marcombo.com

ISBN: 978-84-267-2721-3
D.L.: B-5589-2019

Impreso en Servicepoint
Printed in Spain

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

Contenido disponible online

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 www.marcombo.info y complete el formulario con el código INT1.

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

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)

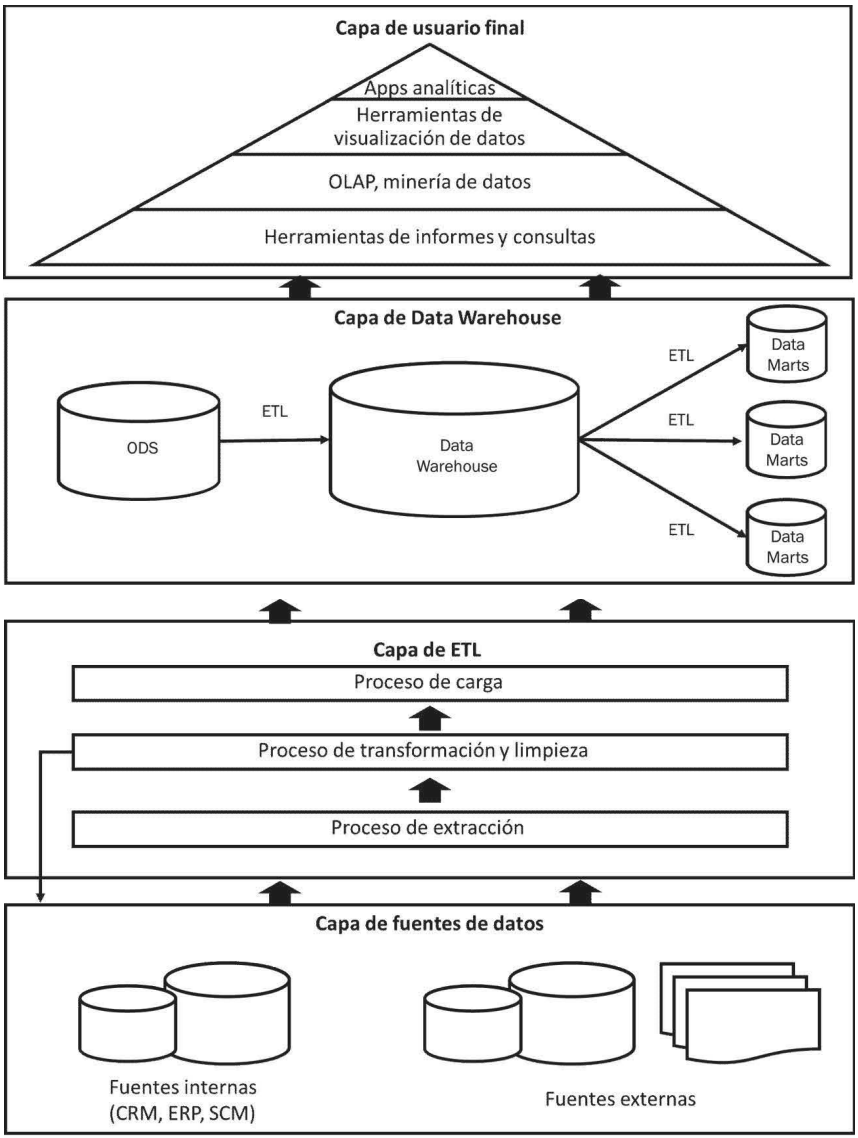

Figura 1.1. Arquitectura de inteligencia de negocios de cinco capas
Fuente: Lih Ong, Pei Hwa Siew y Siew Fan Wong (Ong *et al*, 2011)[9]

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).

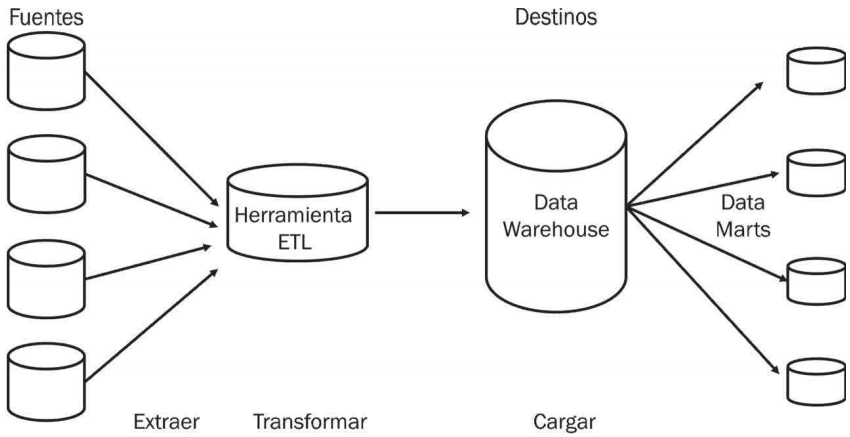

Figura 1.2. Flujo de datos en proceso ETL

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.

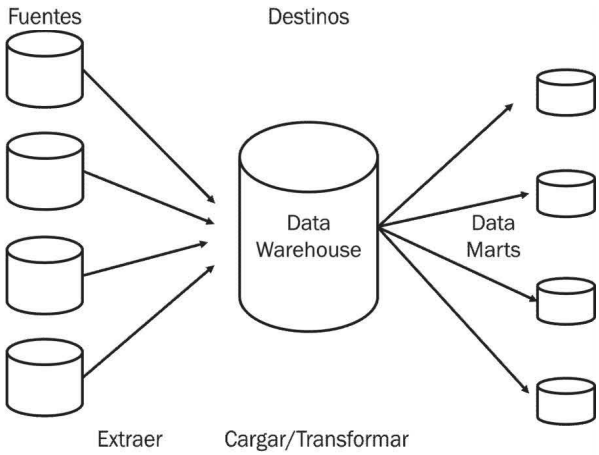

Figura 1.3. Proceso ELT (extraer, cargar, transformar)

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*).

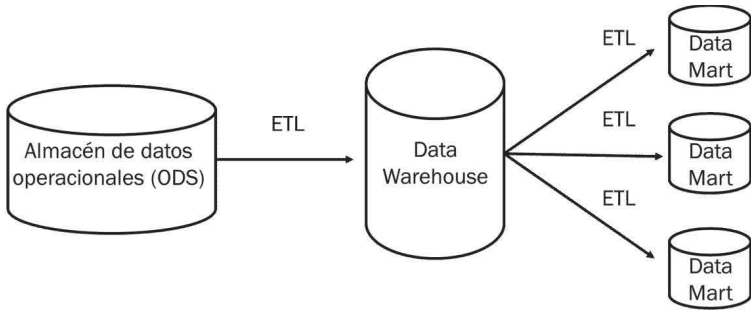

Figura 1.4. Capa de *Data Warehouse* (almacén de datos)

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)

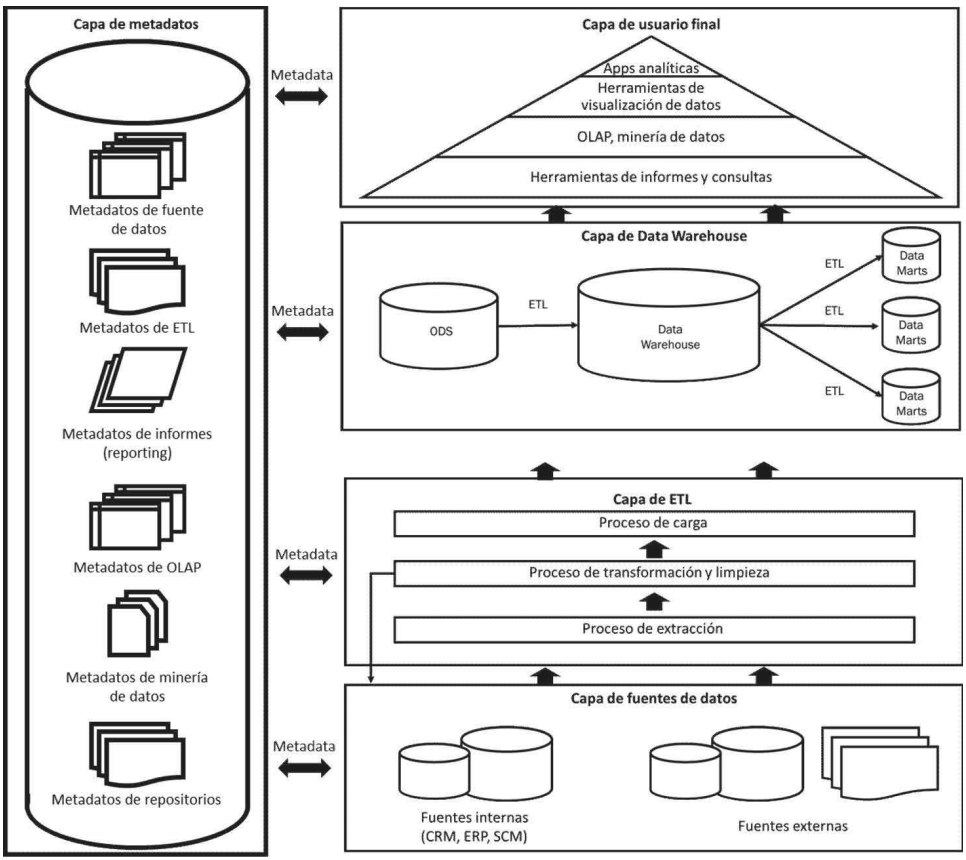

Figura 1.5. Capa de metadatos. Fuente: (Ong *et al*, 2011)[10]

- 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.

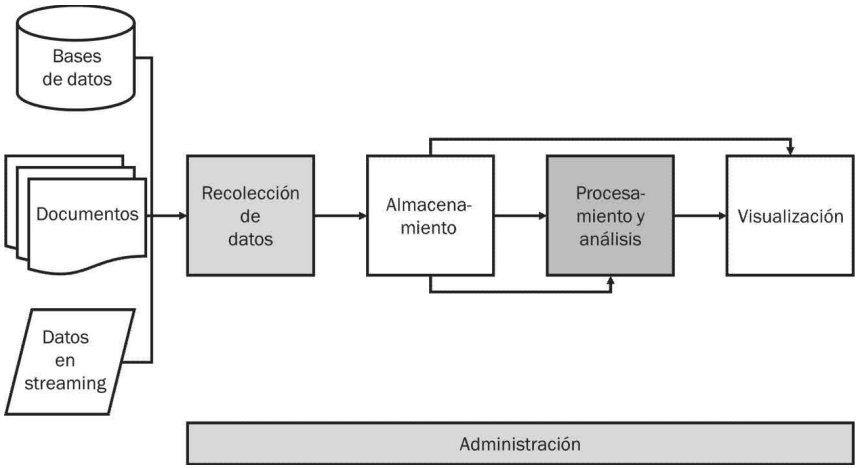

Figura 1.6. Arquitectura de Inteligencia de Negocios con integración de Big Data
Fuente Chunmei Duan, (2014)

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 *clústere*s 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—.