PRESENTACIÓN

En el momento en que tuve la oportunidad de escribir este libro, lo primero que me vino a la cabeza fue que podía aportar yo a la ya extensa y variada literatura existente sobre el tema que nos ocupa. Las estanterías de las librerías temáticas, las plataformas de aprendizaje en línea y los blogs especializados rebosan contenidos alrededor del mundo del *Big Data*. Si a esto sumamos el ingente número de materiales y recursos elaborados por las compañías que se dedican al tema, ya sea en el desarrollo de *software* o en la prestación de servicios, nos encontramos con un área de conocimiento y una práctica empresarial, a priori, sobradamente documentada.

Sin embargo, es en esa abundancia donde para muchos está el problema; y es precisamente en ella donde yo encontré el primero de los argumentos que necesitaba para empezar a escribir. El grado de especialización de los textos sobre tecnologías de la información es parejo al de sistemas, arquitecturas, metodologías o marcos de desarrollo que la componen. En cierta manera, eso es lo esperable. Ahora bien, en estos contenidos tan específicos, al lector que se adentra por primera vez en la materia le cuesta enormemente posicionar y entender los distintos elementos que, como en el caso de *Big Data*, componen un ecosistema de tecnologías, servicios y soluciones de por sí complejo. Es verdad que existen libros (y muy buenos) con una intención más generalista, pero su tendencia es a girar alrededor de la infraestructura para el almacenamiento y el procesado de los datos, dejando las distintas formas de explotación y análisis para la estantería sobre inteligencia y analítica de negocio. Por lo tanto, la conveniencia de aportar una visión mucho más panorámica, conceptual y completa sobre todo el ciclo de vida del dato actuó como una motivación para que me lanzara a la escritura.

El segundo argumento tiene que ver con devolver lo aprendido. En estos ya 30 años alrededor del mundo del dato, primero desde la investigación, después desde el sector y la empresa, y siempre con incursiones en la docencia, uno no solo crece profesionalmente a base de experiencias, sino que desarrolla también una forma de entender y explicar las cosas. Y es esto lo que precisamente uno puede aportar a los que se inician en esta compleja y apasionante materia: su punto de vista, los elementos que considera más

relevantes, lo que le costó entender en su momento, aquello que más le llamó la atención y le llegó a entusiasmar.

Existió desde el principio una tercera motivación, en ningún caso menor: contribuir a hacer más amplia la literatura en castellano sobre tecnologías de la información. Tengo que admitir que la tarea no ha sido fácil. En un campo en el que el lenguaje oral está dominado por los anglicismos, trasladar estos conceptos al papel intentando poner un mínimo de rigor, pero sin caer en traducciones sin sentido que no aportan nada, es más complicado de lo que parece. En este sentido, me gustaría agradecer y reconocer el trabajo que instituciones como la Fundéu-RAE vienen haciendo de cara a promover y facilitar el buen uso del lenguaje en los medios e internet.

Respecto a la organización del libro, conceptualmente está divido en tres partes. La primera, compuesta por cinco capítulos, comienza con una visión general sobre las necesidades alrededor del tratamiento del dato, presentando *Big Data* como una disciplina que permite a las organizaciones explotar grandes volúmenes de datos heterogéneos para soportar la toma de decisiones. Continúa con un tema que yo considero central, ya que actúa como guía y referencia de todo lo que vendrá después: las arquitecturas y patrones para la organización de los datos. Los sistemas de almacenamiento y persistencia vienen a continuación, dando entrada a las distintas formas de procesamiento del dato, por lotes y en tiempo real.

La segunda parte gira entorno a la explotación analítica de la información a partir del dato una vez transformado y consolidado, con un capítulo por cada forma de análisis: prescriptivo, predictivo, prescriptivo y cognitivo. Si bien la primera parte se centra más en temas de infraestructura e ingeniería de datos, esta segunda se abre a aplicaciones y usuarios de negocio. Finalmente, en una corta tercera parte, formada por un único capítulo, planteo un tema transversal a todo el libro: la gestión y el gobierno del dato. Es quizá poco espacio para un aspecto tan importante que todas las organizaciones deben abordar. Sin embargo, he preferido incluirlo, aunque sea de forma breve, antes que dejarlo fuera y dar más extensión a otros de los temas más troncales, pero ya tratados.

Por último, no quisiera terminar esta presentación sin agradecer a mi buen amigo y colega Jaime Requejo el haber compartido conmigo su punto de vista sobre el planteamiento y la exposición del análisis descriptivo, tema del que es un consagrado y reconocido especialista.

Vila de Gràcia, Barcelona, a 15 de mayo de 2023.

ACERCA DEL AUTOR

Víctor López Fandiño es doctor en ingeniería industrial por la Universitat Ramon Llull, Barcelona, con una especialización en quimiometría sobre la aplicación de las redes neuronales artificiales al análisis estadístico multivariante. Con más de 30 años de experiencia en el sector de las tecnologías de la información, ha desarrollado la mayor parte de su carrera profesional en IBM, pasando por las divisiones de consultoría, *software* y, más recientemente, encargándose de la habilitación técnica de los socios tecnológicos de la compañía en España. Paralelamente, ha colaborado con distintas escuelas de negocio y universidades en la impartición de seminarios, cursos de especialización y asignaturas sobre explotación y análisis de los datos.

Siempre ha trabajado en áreas relacionadas con la gestión de la información, especialmente en temas de minería de datos, *data warehousing* y analítica de negocio, disciplinas por las que tiene una certificación como *Distinguished Technical Specialist*, otorgada por The Open Group.

1

BIG DATA: DEL DATO A LA INFORMACIÓN

Hablar hoy en día de *Big Data* en un contexto empresarial es hablar simplemente de datos: el uso del calificativo es, en una gran mayoría de casos, innecesario. En relativamente poco tiempo, empresas de todos los tamaños y niveles de facturación han tomado conciencia del volumen real de datos que les rodea. Estos datos no solo surgen de la propia actividad del negocio, sino que provienen también de fuentes externas que proporcionan un contexto y un sentido a esa actividad. Aunque probablemente sí en cuanto a cantidad, tampoco son datos necesariamente nuevos. A pesar de que muchos de ellos son intrínsecos al propio negocio, su puesta en valor, explotación y rentabilización no se ha producido hasta hace no muchos años.

En este primer capítulo vamos a sentar las bases de lo que entendemos por conceptos tan habituales, pero al mismo tiempo tan complejos, como dato, información y conocimiento. Estudiaremos el ciclo de vida de los datos desde su generación hasta su explotación, centrándonos en esta última desde un punto de vista analítico. Plantearemos también las distintas necesidades y los retos que se derivan de la gestión del dato.

1.1 DATOS, INFORMACIÓN Y CONOCIMIENTO

Muchas veces los conceptos de **dato** e **información** se utilizan de forma equivalente. Hablamos de gestión de los datos y de gestión de la información de forma intercambiable y recursiva, quizá dándole un matiz y un contexto más operacional a la primera y más analítico a la segunda. Al mismo tiempo, la mayoría coincidiremos en que el **conocimiento** es un concepto que requiere una mayor elaboración, estando dotado de un mayor nivel de abstracción. Si además incluimos la **sabiduría** dentro del conjunto, entonces la complejidad conceptual aumenta todavía más.

La relación entre estos cuatro conceptos es algo profundamente estudiado, tanto desde el punto de vista de las ciencias de la información como de la epistemología, si bien no existe un consenso claro entre las distintas escuelas. Una forma habitual de representar esta relación es a través de la llamada **pirámide DIKW** (*Data*, *Information*, *Knowledge*, *Wisdom*).

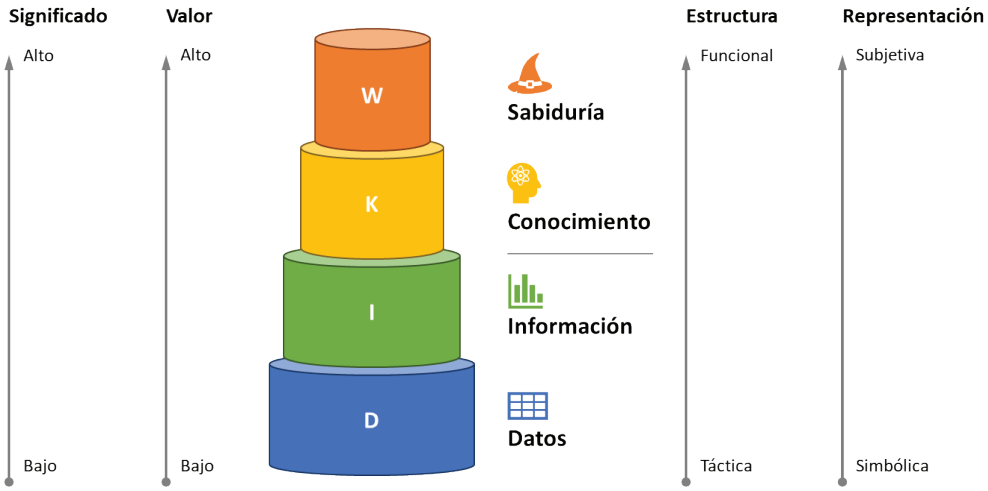

Figura 1-1. Pirámide DIKW.

Además de establecer una jerarquía, el modelo que hay detrás de esta pirámide (Figura 1-1) proporciona una definición más o menos consensuada de cada uno de estos cuatro conceptos, de forma que cada uno se apoya en el del peldaño anterior. La Tabla 1-1 contiene estas definiciones.

Desde el punto de vista de las tecnologías de la información, a medida que ascendemos por la pirámide nos enfrentamos con conceptos menos programables y susceptibles de ser manipulados mediante algoritmos, aunque la **inteligencia artificial** (**AI**, *Artificial Intelligence*) se empeñe día a día en contradecir esto. También en este ascenso vamos incorporando roles a esta cadena de valor. Desde una perspectiva empresarial, los **usuarios de negocio**, aquellos más cercanos a la toma de decisiones, son los encargados de generar conocimiento, entrando en este nivel de la pirámide y liderando el resto de la subida. Hasta ese punto, son los **ingenieros** los responsables de la captación de los datos y su elaboración de cara a facilitar la generación del conocimiento. Es evidente la importancia de cada uno de estos dos roles en esa cadena de valor.

Concepto	Definición	Características
Datos	Colección de hechos elementales codificados mediante símbolos y registrados sin una organización concreta	• Propios de las cosas • Discretos y objetivos • Operacionales
Información	Colección de datos procesados y organizados con el propósito de ser útiles a su destinatario, aportándole significado	• Entendible • Categorizada • Comunicable
Conocimiento	Información puesta en contexto, que combinada con experiencia, valores y reglas permite la toma de decisiones	• Propio de las personas • Subjetivo • Analítico
Sabiduría	Conocimiento acumulado que permite a las personas actuar de forma crítica y práctica ante una situación determinada	• Depende de un sistema de valores • Comporta un juicio ético

Tabla 1-1. Conceptos en la pirámide DIKV.

Al margen de lo que significa la pirámide del DIKV en la teoría general de la información y el conocimiento, a nosotros nos resulta útil para acotar el dominio en el que nos vamos a mover a la hora de plantear los sistemas de *Big Data*.

Podemos definir ***Big Data***[1] como el conjunto de operaciones, técnicas y tecnologías orientadas al procesamiento de grandes y variados volúmenes de datos, con el fin de generar información válida sobre la que desarrollar conocimiento y soportar las decisiones de negocio. Es decir, nos centramos en la base de la pirámide con el objetivo de habilitar y facilitar el tercer peldaño[2]. Por consiguiente, los sistemas de *Big Data* son aquellos componentes de *hardware* y *software* encargados de soportar esas operaciones.

1 La Real Academia Española admite y recomienda el uso del término **macrodatos**, en lugar de *Big Data*. Sin embargo, he optado por mantener el anglicismo en la medida en que nos referimos a una disciplina y su uso está ampliamente extendido.

2 La sabiduría es un concepto tremendamente abstracto que la pirámide DIKV intenta con mayor o menor éxito conceptualizar (de hecho, no son pocos los autores que la dejan fuera) y que nosotros no tocaremos. Tampoco entraremos en la gestión del conocimiento como rama de las tecnologías de la información.

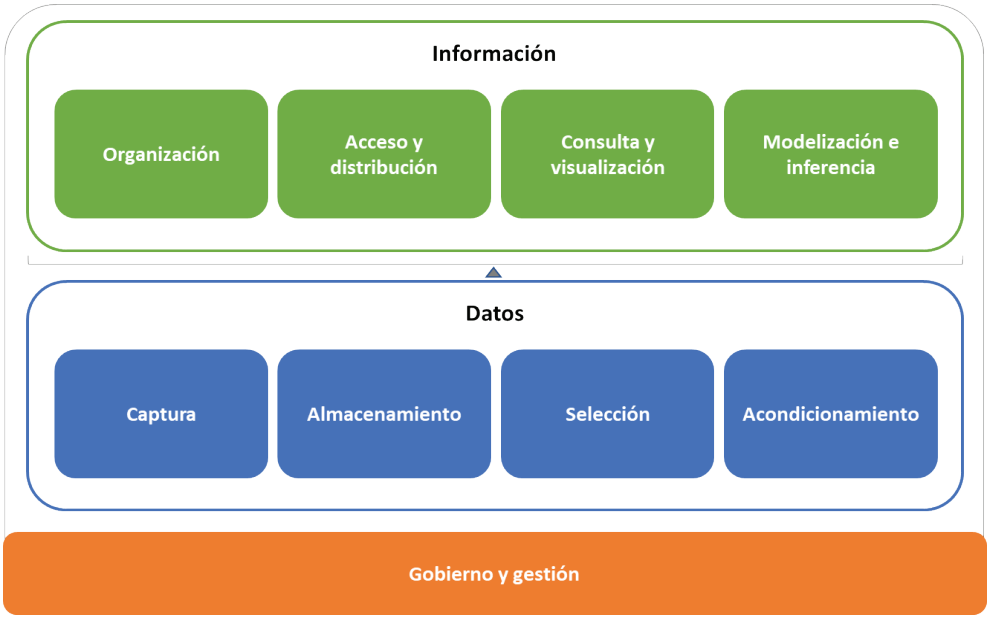

Figura 1-2. Operaciones de Big Data sobre datos e información.

La Figura 1-2 recoge esta idea. Sobre una capa de gobierno, encargada de proporcionar una serie de servicios comunes y unificados que van desde la trazabilidad hasta el control de acceso, se construye una infraestructura para la gestión del dato que permite su elaboración y transformación en información.

Si bien es posible caracterizar las distintas operaciones que componen esta cadena de valor, no siempre es fácil establecer cuando el dato deja de serlo y pasa a constituirse en información. Hay zonas difusas, especialmente en lo referente a la organización y distribución, donde las diferencias no están tan claras. Por ello, es habitual asociar la información con los sistemas encargados del acceso o consumo por parte de los usuarios finales, mientras que los datos quedarían confinados a aquellos que son internos, responsabilidad de los ingenieros y el departamento de tecnología.

Adicionalmente a este planteamiento, que es necesario para identificar y posicionar los elementos con los que vamos a tratar, no hay que perder de vista que los sistemas solo saben operan con datos, y además de forma agnóstica y descontextualizada[3]. En este sentido, entendemos los datos como las unidades básicas de almacenamiento sobre las que operan los ordenadores a través de procesos y aplicaciones. Esta visión operativa es compatible, y reconciliable, con nuestro modelo conceptual, especialmente desde el momento en que entendemos la información como un conjunto de datos procesados y organizados. Por lo tanto, cuando hablemos de datos en general estaremos englobando también la información como concepto, añadiendo siempre los matices pertinentes.

3 Nuevamente, la inteligencia artificial está marcando un antes y un después al respecto.

1.2 CARACTERIZACIÓN DEL DATO

Sobre este punto de partida, el dato puede ser caracterizado de muchas maneras, tomando tanto ejes técnicos como de negocio.

1.2.1 Datos en cuanto al tipo

Vamos a comenzar por una clasificación muy técnica y granular, pero que subyace en toda narrativa alrededor del procesamiento de los datos. Desde el punto de vista del tipo de operaciones que pueden hacer los ordenadores sobre los datos, podemos hablar de dos grandes clases:

- **Tipos simples**. También denominados tipos primitivos, representan un único valor. Cada tipo simple establece que valores puede tomar el dato, y dentro de que rango, así como las operaciones que se pueden realizar. Los tipos simples se dividen en lógicos, caracteres y numéricos, cada uno de ellos representado por un número determinado de bits. Mediante un tipo simple se puede codificar el salario de un empleado, el estado civil de un ciudadano o el indicador de que un cliente no desea recibir publicidad, por ejemplo.
- **Tipos compuestos**. Como resultado de la combinación de los tipos simples aparecen tipos compuestos, que representan un conjunto de valores a modo de estructura. Dentro de estos tipos nos podemos encontrar vectores, matrices, listas, conjuntos, registros, etc. Con tipos compuestos podemos representar el nombre de un producto, la imagen de la matrícula de un coche, en forma de matriz de bits, el audio de la transcripción de una conversación, o entes más complejos, como un coche o una persona.

1.2.2 Datos en cuanto al formato

Siguiendo en el ámbito técnico, pero ya dando forma a la idea de colección, podemos hablar de datos en cuanto a la forma de organizarlos (Figura 1-3). Como veremos, esta caracterización tiene mucha importancia a la hora de hablar del formato de los datos y la manera de almacenarlos.

- **Datos estructurados**. Una colección de datos está estructurada cuando presenta un modelo o esquema organizativo. Es decir, todos los elementos de la colección responden a una misma organización, tanto en cuanto a tipos como a significado. La primera idea que se nos viene a la cabeza cuando hablamos de datos estructurados es la de una **base de datos SQL**, donde las colecciones se materializan en forma de tablas y sus relaciones como referencias. Cada tabla responde a un esquema de tipos prefijado, de forma que todos los registros de la tabla tienen la misma estructura. Por ejemplo, un cliente puede estar almacenado como un tipo compuesto en una tabla, constituyendo un registro. Este tipo estará formado por un conjunto de tipos simples, representando cada uno un atributo

sociodemográfico o conductual. Todos los registros (clientes) de la tabla responden a los mismos atributos[4]. En cualquier caso, la base de datos relacional no es el único medio de persistencia de las colecciones de datos estructurados, ya que estas pueden almacenarse en ficheros planos con separadores, hojas de cálculo u otros formatos propietarios.

- **Datos semiestructurados**. Los datos semiestructurados se definen por diferencia, es decir, son aquellos que no son estructurados, pero que presentan cierta organización. Su primer rasgo identificativo es que no responden a una estructura tabular en forma de una colección de registros compuestos por atributos, como veíamos anteriormente. El formato de estas colecciones se basa en una organización jerárquica que agrupa los datos de forma semántica, incluyendo una serie de etiquetas que delimitan los valores y sirven como descripción de la estructura. Esto no implica necesariamente una falta de rigor en la definición, ya que estas colecciones pueden implementar un esquema susceptible de ser validado. Por el contrario, ofrecen más flexibilidad a la hora de definir la organización, permitiendo al mismo tiempo el análisis (*parsing*) de los datos. El correo electrónico (MIME), y los formatos **XML** y **JSON** son ejemplos de cómo organizar datos en colecciones semiestructuradas.

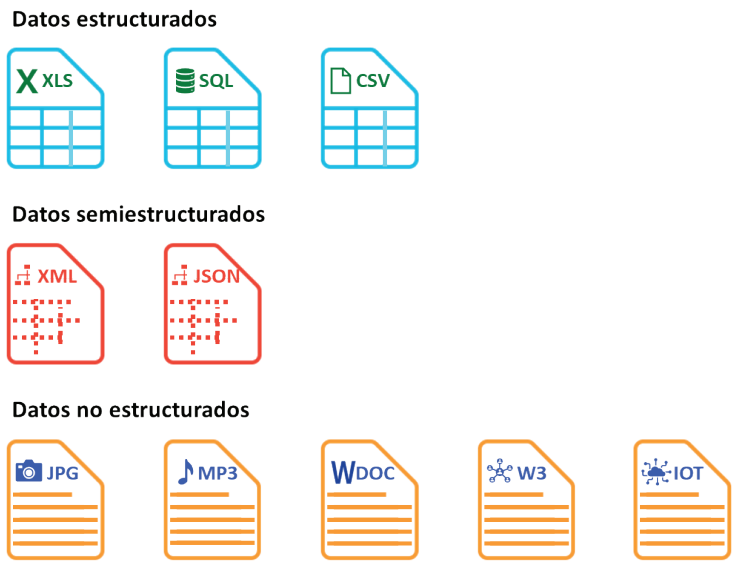

Figura 1-3. Datos en cuanto a formato

4 Esto no implica que todos estos atributos deban estar informados para todos los clientes. El esquema de la tabla, que responderá a un modelo de entidad, marcará, entre otras cosas, que datos admiten la ausencia de un valor.

- **Datos no estructurados**. En el otro extremo nos encontramos con colecciones de datos carentes de estructura. Aquí situamos **datos textuales**, como documentos, mensajes o registros de aplicación (*logs)*, y **datos no textuales**, incluyendo audio, vídeo e imágenes. En cualquier caso, estos formatos sí tienen una organización interna en forma de tipos compuestos, conformada además a un estándar (JPG, MP3, AVI, etc.). El calificativo de no estructurado aparece debido a la carencia ya de un esquema que facilite el acceso y la consulta. Estos tipos de datos, al igual que los semiestructurados, se acostumbran a persistir en sistemas de almacenamiento de objetos y **bases de datos NoSQL** especializadas[5].

1.2.3 Datos en cuanto al generador

Otro eje para considerar es el que tiene en cuenta quien es el creador de los datos. Tenemos dos posibilidades

- **Datos generados por personas**. Estas son unas de las colecciones que más rápidamente está creciendo, no tanto a nivel corporativo, sino por el gran volumen de interacciones en las redes sociales y el comercio electrónico. Aquí incluimos operaciones de compra, correos electrónicos, documentos de texto, hojas de cálculo, mensajes, video, imágenes, audio, etc. Si bien aquí hay una gran variedad de datos no estructurados, el volumen asociado a las transacciones comerciales directas entre empresas y consumidores (B2C, *Business-to-Consumer*) se apoya mayoritariamente en datos estructurados.
- **Datos generados por máquinas**. Son aquellos producidos por dispositivos digitales o aparatos mecánicos, sin que medie la intervención humana, y normalmente asociados a procesos industriales o científicos. Aquí podemos incluir imágenes generadas por sistemas de vigilancia o satélites, datos de sensores en entornos y aplicaciones de **IoT** (*Internet Of Things*), transacciones automáticas entre empresas (B2B, *Business-to-Business*) o registros de aplicaciones y sistemas, estos últimos suponiendo un volumen muy grande que no para de crecer[6]. El número de datos no estructurados en esta categoría tiene cada vez más peso.

---

5 En múltiples fuentes se indica que los datos no estructurados, y por extensión los semiestructurados, son aquellos que no pueden residir en una base de datos relacional. Esto no es correcto, ya que cualquier motor de SQL moderno puede almacenar imágenes, audio o documentos XML dentro de un campo de una tabla, permitiendo distintos niveles de interrogación y relación.

6 De hecho, el procesamiento de registros (*logs)* era una de las aplicaciones de referencia cuando se empezó a hablar de *Big Data*.

1.2.4 Datos en cuanto al tamaño

Si medimos los datos en términos de tamaño, no nos queda más remedio que relativizar la nomenclatura; lo que son volúmenes pequeños para una empresa pueden suponer un desafío para otra. La Figura 1-4 muestra algunos ejemplos de volúmenes de datos para hacernos una idea relativa de los tamaños.

Con el riesgo que conlleva delimitar unos rangos, podemos establecer las siguientes categorías, órdenes de magnitud y ejemplos:

Figura 1-4. Escala de almacenamiento de datos con algunos ejemplos.

- **Datos pequeños** (*gigabytes*). Un ejemplo podría ser una base de datos de proveedores, conteniendo información de contacto con varios miles de registros. Estos datos se pueden procesar con un *software* ofimático en ordenadores personales[7].
- **Datos medianos** (*terabytes*). Una base de datos conteniendo transacciones comerciales, con detalle de pedidos, facturas y devoluciones. Aquí estaríamos hablando de varios millones de registros, procesados con tecnologías convencionales.

7 Aunque aquí hablamos de datos pequeños en función de su tamaño, el término ***small data*** se usa comúnmente como continuación, o incluso alternativa, al *Big Data*, remarcando que la correcta visualización, comprensión y comunicación de los datos solo puede tener lugar sobre conjuntos pequeños.

- **Datos grandes** (*petabytes*). Los datos derivados de una aplicación de comercio electrónico, que incluyen rutas de navegación del usuario, tiempo de sesión, búsquedas, incidencias, etc. Las volumetrías estarían aquí sobre los billones de registros, requiriendo ya sistemas distribuidos y entornos de computación escalables.
- **Datos muy grandes** (*exabytes*). Aquí incluimos el procesamiento de datos de satélites, genómica, imágenes médicas, entornos de correlación de eventos de seguridad, inteligencia artificial, etc., requiriendo sistemas específicos y dedicados de computación.

Lógicamente, estas consideraciones tienen un impacto en los mecanismos de almacenamiento de los datos, pero no solo por el volumen en sí, sino también por su temperatura, medida esta como la frecuencia de acceso. Hablamos de **datos calientes** (*hot storage*) cuando estos requieren un acceso frecuente e instantáneo, siendo cruciales para el negocio. Por el contrario, los **datos fríos** (*cold storage*) son aquellos inactivos la mayor parte del tiempo, no requiriendo un acceso inmediato y permaneciendo archivados. Esta diferenciación implica métodos de almacenamiento separados, optimizados para cada caso, y que tienen una importante repercusión en el coste de la infraestructura.

1.2.5 Datos en cuanto a su rol

Las cuatro primeras clasificaciones que hemos visto son de carácter básicamente técnico. Sin embargo, para comprender bien el papel y el valor que los datos aportan al negocio es necesario ponerlos en un contexto más funcional.

Los datos que manejan las empresas tienen un trasfondo corporativo. Esto quiere decir que son compartidos de forma controlada por empleados, socios y proveedores a lo largo de diferentes organizaciones y departamentos, en diferentes geografías. Algunos de ellos son accesibles también por los clientes, como parte de las transacciones comerciales, y otros deben estar disponibles de cara a cumplir con marcos regulatorios, o incluso requerimientos judiciales. Desde este punto de vista corporativo, podemos clasificar los datos en cuatro categorías:

- **Datos maestros**. Son aquellos que detallan las entidades principales del negocio, y que son compartidos y utilizados por distintas aplicaciones. Ejemplos de datos maestros son clientes, empleados, productos u oficinas. Los datos maestros deben tener una concepción transversal del negocio, ya que implican a todos los departamentos, y son críticos para su funcionamiento. Por este motivo, su gestión debería centralizarse.
- **Datos operacionales**. Son los derivados del propio funcionamiento del negocio, consecuencia de las transacciones comerciales con clientes y proveedores. Los sistemas que producen estos datos son críticos, ya que su caída implicaría el paro de las actividades. Los datos operacionales necesitan datos maestros para tener sentido. Por ejemplo, una compra en un supermercado genera datos operacionales

dentro de un escenario formado por un cliente, una tienda, una serie de productos, un vendedor, etc. El dato operacional en si hace referencia al detalle de la facturación de la compra y las unidades vendidas, los puntos generados bajo el programa de fidelización, o el tiempo invertido por el vendedor en escanear los artículos. El escenario, por su parte, está formado por datos maestros.

- **Datos externos**. Los datos externos son aquellos no generados por el negocio, pero que tienen una relación con él, siendo susceptibles de influir y aportar valor. Aquí podemos incluir datos meteorológicos, que nos informan de la previsión de lluvias y su impacto a la hora de establecer los periodos de siembra, datos de redes sociales, diciéndonos el sentimiento que generan nuestros productos y servicios, o datos encargados a proveedores o agencias externas, que nos detallan la propensión de voto por código censal en las próximas elecciones municipales. Los datos externos pueden actuar como fuente para los datos maestros, aportando un perfil sociodemográfico de nuestros clientes en función de su lugar de residencia, por ejemplo.
- **Datos analíticos**. Aquí podríamos hablar ya de información en lugar de datos, en el sentido que hemos mencionado en el apartado anterior. Si el dato operacional tiene sentido dentro de un escenario, también debe analizarse en este para su comprensión. El dato analítico se genera a partir de los datos operacionales, denominados ahora **hechos**, dentro del contexto de los datos maestros, denominados aquí **dimensiones**, y relatado a lo largo de una **perspectiva temporal**. Es decir, el dato analítico siempre es dimensional, siendo el tiempo una de las dimensiones más importantes y ubicua. El dato analítico puede recircular, enriqueciendo los datos maestros; sería el caso, como ya veremos, de nuevos atributos de los clientes generados a través de un modelo de segmentación.

Ligar el dato analítico al tiempo como dimensión no implica que los datos operacionales no puedan monitorizarse y estudiarse en tiempo real[8]. Sin embargo, cuando nuestro análisis está enfocado a soportar la toma de decisiones desde un punto de vista estratégico y táctico, es imprescindible tomar una cierta profundidad histórica. Por el contrario, la monitorización del dato operacional estaría más enfocada a la toma de decisiones operativas.

8 Aún en este caso, el tiempo sigue siendo un eje imprescindible.

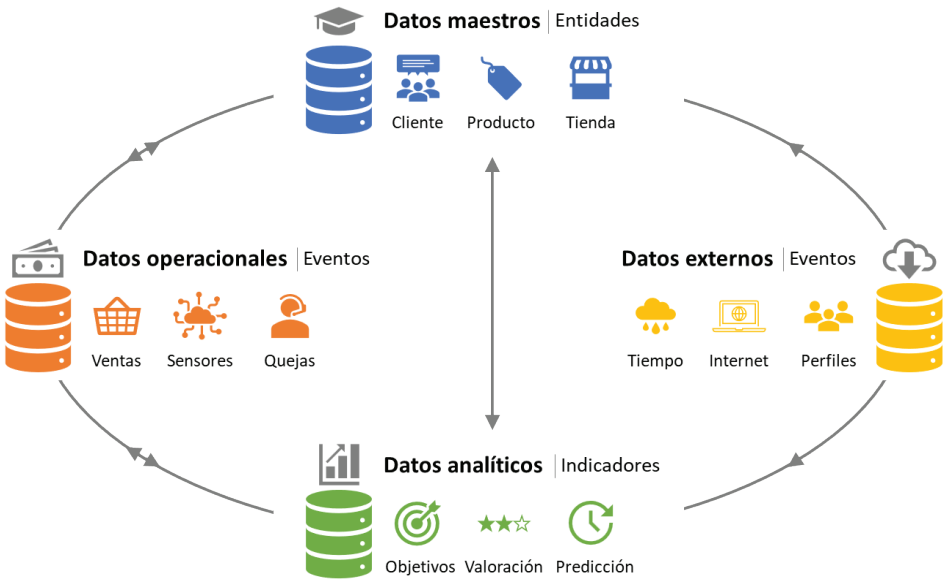

Figura 1-5. Datos corporativos

Cuando hablamos de **gestión de datos corporativos** (EDM, *Enterprise Data Management*) nos estamos refiriendo a los procesos involucrados en el manejo de estos cuatro tipos de datos. La Figura 1-5 muestra las relaciones entre ellos; la existencia de conexiones y dependencias bidireccionales añade complejidad en la gestión.

1.2.6 Datos en cuanto a su latencia

En términos de gestión de eventos, la **latencia** se define como el tiempo total transcurrido entre que un dato es generado y es puesto a disposición de las aplicaciones y los usuarios para ser consumido. Desde un punto de vista técnico, la latencia puede descomponerse en latencia de red, almacenamiento, procesado, etc. A estos tiempos podemos añadir, ahora funcionalmente, una latencia de análisis, decisión e implementación. Existe el convencimiento de que a mayor latencia menor es el valor que aporta el dato al negocio, pero esto debería matizarse.

Podemos hablar de dos tipos de datos en términos de latencia:

- **Datos en tiempo real**. Son aquellos captados en el momento en que son obtenidos. Esto implica dos cosas: por un lado que la captación es próxima a la inmediatez, produciéndose de forma constante; por el otro, que la integración del dato tiene lugar en la propia captación. El procesamiento en tiempo real es una característica inherente a los datos operacionales. Un sistema de reserva de entradas, un monitor de transacciones fraudulentas o un sistema de detección y apagado de incendios, son ejemplos de aplicaciones que tienen que trabajar en

tiempo real. El procesamiento de datos en tiempo real suele medirse en segundos, milisegundos o incluso menos.

- **Datos en lotes**. En este caso, los datos son almacenados en un lote (*batch*) cuando son recibidos, permaneciendo así durante un cierto periodo de tiempo o hasta que alcanzan un volumen determinado. Después son procesados de forma planificada como un conjunto y entregados en destino. Este periodo de tiempo puede ir desde una hora hasta varios meses. Aquí el tiempo de entrega no solo no es crítico, sino que además es, en muchos casos, necesario. Por ejemplo, un sistema de facturación de agua o electricidad tiene que operar necesariamente por lotes, acumulando las lecturas del contador durante un mes para poder emitir la factura. Aunque el dato analítico se ha asociado a este tipo de procesamiento, imprescindible para dar esa profundidad histórica de la que hablábamos antes, la necesidad de acortar los procesos de toma de decisiones, y la disponibilidad de datos operacionales para hacerlo, ha hecho que el procesado en tiempo real tenga tanta relevancia o más que el procesado por lotes.

Aunque la industria demanda cada vez más el aprovisionamiento y el procesado de datos en tiempo real, el tratamiento por lotes tiene una serie de ventajas desde el punto de vista de la integración de los datos.

Una de ellas es la eficiencia que se gana al unificar todo el tratamiento en un único proceso, no teniendo que gestionar cada dato individual cada vez que este se adquiere o genera. Además, permite un uso eficiente de los sistemas al poder planificar los lotes en periodos en que la carga de trabajo es baja.

Riesgo bajo	Riesgo medio	Riesgo alto
• Datos y contenidos autorizados, disponibles a través de páginas web públicas de la empresa • Ofertas de empleo • Notas de prensa • Material de *marketing* aprobado para uso público • Datos de contacto no designados como privados por la empresa o por el empleado	• Identificadores internos de empleados • Datos de desarrollo, investigación y patentes no publicados • Contenidos con propiedad intelectual licenciada de un tercero o restringida por contrato • Datos de recursos humanos relacionados con los empleados • Contratos no públicos • Datos financieros • Correo electrónico, informes, presupuestos y planes internos	• Identificadores personales (números de la Seguridad Social, DNI, carné de conducir, pasaporte, etc.) • Contraseñas y claves de acceso a sistemas, aplicaciones, etc. • Información identificable sobre salud o pólizas • Números de tarjetas de crédito o débito • Números de cuentas bancarias • Exportaciones controladas • Donaciones y regalos

Tabla 1-2. Ejemplos de tipos de datos en función de su sensibilidad.

En cualquier caso, será el proceso de negocio y el usuario final el que dicte la latencia más apropiada para cada caso, siendo habitual tratar un mismo dato de ambas formas con el fin de satisfacer necesidades y tiempos diferentes.

1.2.7 Datos en cuanto a su sensibilidad

Por último, los datos pueden ser también clasificados según su connotación en términos de privacidad e intimidad. Esta asignación marcará quien puede acceder a los mismos y durante cuánto tiempo, estableciendo también las condiciones en que deben ser almacenados.

Las empresas suelen emplear distintas categorías para gestionar sus datos en cuanto a su sensibilidad y acceso. Estas categorías tienen en cuenta tanto datos recogidos de las interacciones con clientes y proveedores, como aquellos generados por el propio negocio. Varían de un país a otro en función de las regulaciones que se aplican (GDPR, SOC2, HIPAA, PCI SSC, etc.)[9], y cada empresa introduce además sus propias matizaciones. De forma general, podemos hablar de tres clases de datos en términos del riesgo que supone su manipulación indebida:

- **Riesgo alto**. Son aquellos datos cuyo acceso, revelación o manipulación no autorizada puede suponer acciones legales y cargos criminales contra la empresa, poniéndola en riesgo. Estos datos están tipificados y gobernados por regulaciones nacionales e internacionales, y es obligatorio informar de cualquier incidencia o incumplimiento en cuanto a su gestión.
- **Riesgo medio**. Son datos igualmente regulados cuyo uso o revelación está sujeto a una serie de restricciones contractuales. Su uso no autorizado puede tener un efecto adverso tanto en la empresa y sus empleados, como en clientes, proveedores y socios comerciales.
- **Riesgo bajo**. Aquí incluimos datos disponibles al público en general. Su acceso, uso o alteración no tienen un impacto negativo ni en la empresa ni en su ecosistema.

La Tabla 1-2 da algunos ejemplos de datos pertenecientes a estas tres categorías.

---

9 **GDPR** (*General Data Protection Regulation*): ley de la Unión Europea para la protección global de la privacidad y los datos; **SOC2** (*Service Organization Control 2*): marco internacional para la gestión de datos de clientes; **HIPAA** (*Health Insurance Portability and Accountability Act*): ley estadounidense para la protección y seguridad de los datos médicos confidenciales de los pacientes; **PCI SSC** (*Payment Card Industry Security Standards Council*): directiva para la seguridad en el uso de tarjetas de pago establecida por los principales proveedores, como American Express, MasterCard y Visa Inc.

1.3 *BIG DATA* EN CONTEXTO

Una vez que hemos caracterizado el dato a lo largo de siete ejes (habría más), estamos ya en disposición de plantear lo que significa *Big Data* como concepto y modelo de procesamiento de datos para el negocio.

Hablar de *Big Data* es hacerlo de un antes y un después. Que este paso haya supuesto una revolución, en el sentido de un cambio de paradigma tecnológico, como lo será la **computación cuántica**, es algo ya más discutible. En cualquier caso, es habitual que cuando hablamos de *Big Data* lo veamos como una respuesta a una situación en la que las corporaciones se vieron desbordadas por la cantidad de datos que les rodeaban, planteándose como aprovecharlos. Si bien no hay nada de incorrecto en esta formulación, sí que adolece de un matiz importante que la haría menos reactiva: el análisis de grandes colecciones de datos permite adquirir un conocimiento que no es posible abordando solo conjuntos pequeños.

Figura 1-6. Las cinco uves del Big Data.

Las empresas se empezaron a dar cuenta de esto hace ya más de 25 años, cuando la **minería de datos** (*data mining*) se incorporó como disciplina dentro de las prácticas de la **inteligencia de negocio** (**BI**, *Business Intelligence*). Sin embargo, en aquella época fallaban dos elementos importantes: el primero de ellos era la cultura empresarial, conservadora y reacia a incorporar a la toma de decisiones elementos externos que no acababa de entender. El segundo era el estado de la tecnología: la minería de datos se hacía sobre muestras debido a que las aplicaciones no escalaban; no había volúmenes significativos de datos no estructurados, ni tampoco oferta comercial madura para tratarlos; los especialistas en modelización escaseaban y, además, todavía no existía una oferta de infraestructura analítica que fuera asequible y económica.

Quizás lo más propio sería hablar de *Big Data* como un fenómeno que se da alrededor de 2010 cuando convergen, en un orden por determinar, cuatro factores importantes:

- La aparición y consolidación de nuevos modelos de negocio que se basan en la disponibilidad y variedad de datos que hay en internet, creando a su vez más datos sobre los que se realimentan. El auge del **internet de las cosas**, con la explosión de dispositivos y sensores conectados en red, la generalización del acceso y participación en las redes sociales o la preponderancia del comercio electrónico son algunos ejemplos.
- Un cambio en la cultura empresarial, conducido por una nueva generación de profesionales de formación interdisciplinar y habituada a la tecnología.
- Un avance tecnológico que, girando entorno a la **computación en la nube** (*cloud computing*), proporciona nuevas funcionalidades y capacidades con unos modelos de consumo muy flexibles y con costes asequibles.
- Una oportunidad de negocio en sí mismo, explotada por los departamentos de *marketing* de las grandes corporaciones tecnológicas.

1.3.1 El modelo de las cinco uves

Sea como fuere, hay un cierto consenso a la hora de caracterizar el *Big Data* como la confluencia de cinco propiedades que presentan los datos que manejan las organizaciones. Es el modelo de las **5 uves** (Figura 1-6):

- **Volumen**. La búsqueda de nuevas oportunidades de negocio y la necesidad de diferenciarse de la competencia, hace que la cantidad de datos que hay que analizar para obtener resultados sea cada vez más grande. Como ya hemos comentado, unidades como el *petabyte* y el *exabyte* son ya habituales para referirnos a los volúmenes generados por el comercio electrónico, las redes sociales y dispositivos portátiles (Figura 1-7).
- **Velocidad**. En términos de *Big Data*, la velocidad presenta dos aspectos. El primero hace referencia a la celeridad con que se producen los datos. Todos somos conscientes de que la generación de contenido es un continuo que no se detiene nunca. El segundo aspecto pone el foco en la prontitud con la que los datos deben ser procesados y analizados para sacarles provecho. La combinación de ambos hace, en definitiva, que los flujos de datos que atraviesan las empresas no solo sean voluminosos, sino constantes. Esto supone un reto en términos de latencias de acceso y velocidades de respuesta, pero también en lo referente a la seguridad. Un aspecto igual de importante que la velocidad es la dirección. Tradicionalmente los datos fluían desde los orígenes a los destinos, donde eran almacenados y consumidos por parte de los usuarios finales. Esto ha dejado de ser así, ya que nos podemos encontrar recirculaciones que van desde los entornos analíticos hacia los operacionales, involucrando tanto datos como modelos predictivos.

Figura 1-7. Orígenes de datos para Big Data.

- **Variedad**. Parece que existe un consenso en que entorno al 80-90% de los datos que tienen las organizaciones son no estructurados, y estos porcentajes se llevan escuchando ya varios años: correos electrónicos y mensajes, manuales, transcripciones de audio, video e imágenes, etc. En determinados sectores estos tipos de contenido son todavía más relevantes. El ámbito legal puede ser uno de los extremos, con lo que representan las leyes y la jurisprudencia; el sector de la salud no se queda lejos, con toda la literatura médica, la investigación y los historiales clínicos. Son precisamente estos dos campos donde más están contribuyendo las tecnologías de *Big Data*, y donde se espera un mayor crecimiento.
- **Veracidad**. Aquí nos estamos refiriendo a poder asegurar la certeza de los datos, y este es un aspecto crucial. Si no se puede confiar en los datos, la toma de decisiones está comprometida. El problema de la calidad de los datos no es algo ni mucho menos nuevo, pero parece que a las empresas les cuesta todavía tomar conciencia de su importancia. El volumen y la velocidad vienen a dificultar el control de la veracidad, pero es probablemente la variedad de orígenes el factor más complicado, ya que obliga a un mayor foco en las tareas de reconciliación, limpieza y consolidación del dato. En este sentido, el aseguramiento de la calidad se basa en la trazabilidad del dato y el análisis de su impacto, contabilizando y documentando su origen, las transformaciones por las que pasa y su destino.
- **Valor**. Aunque la presentamos en último lugar, esta es la uve que lo justifica todo. Como decíamos al principio del capítulo, el objetivo final del tratamiento de los datos es convertirlos en información que a su vez nos permita desarrollar un conocimiento para soportar la toma de decisiones. Esto implica poder conocer y comprender a los clientes, diseñando nuevos productos y servicios para satisfacer sus necesidades, retener y fidelizar a los más rentables y optimizar los procesos de negocio para disminuir costes y aumentar el margen de beneficios.

La puesta en valor de lo que los datos representan para una empresa ha provocado una reorientación de todo el negocio a su alrededor. Es lo que se ha venido a denominar **empresas orientadas por los datos** (*data driven companys*).

1.3.2 Empresas orientadas por los datos

Esta idea va mucho más allá de una reingeniería de los procesos, suponiendo un cambio cultural y organizativo. Se trata de orientar el funcionamiento del negocio alrededor de la capacidad que tienen los datos para describir lo que ha pasado, estimar lo qué podrá suceder y determinar cómo anticiparse a ello. Debajo de este cambio de orientación está el mantra de que no se puede conocer lo que no se puede medir, y mucho menos mejorarlo[10].

Para negocios nacidos en la era digital este sería, en principio, su estado natural. Sin embargo, no lo es así para empresas más longevas, donde la resistencia al cambio es mayor (Figura 1-8). De forma general, podemos singularizar a las compañías que se encuentran en este estadio de orientación por los datos por las siguientes características:

- Concienciación de que el dato, después de los empleados y los clientes, es el principal activo de la empresa, y su gestión es una tarea colectiva.
- Los procesos, internos y externos, son necesariamente medibles, y la evaluación de la marcha del negocio y de sus empleados está basada en objetivos cuantificables.
- Los usuarios de negocio tienen una formación interdisciplinar, con un alto grado de autonomía para el acceso y la elaboración del dato. El responsable de los sistemas de información forma parte del consejo directivo, participando en la toma de decisiones estratégicas de la empresa.
- Adopción de modelos híbridos de computación en la nube con diferentes proveedores, consumiendo infraestructura, plataforma y software como servicio con pago por uso.
- Fuerte inversión en tecnología, especialmente en todo lo referente a procesamiento y análisis de datos.

Un aspecto igualmente diferenciador es la involucración y el conocimiento que los usuarios del departamento de tecnologías de la información tienen del negocio, lo que contribuye decisivamente a una innovación constante.

---

10 Esta frase proviene de una cita de Lord Kelvin (1824–1907): «*I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind*».

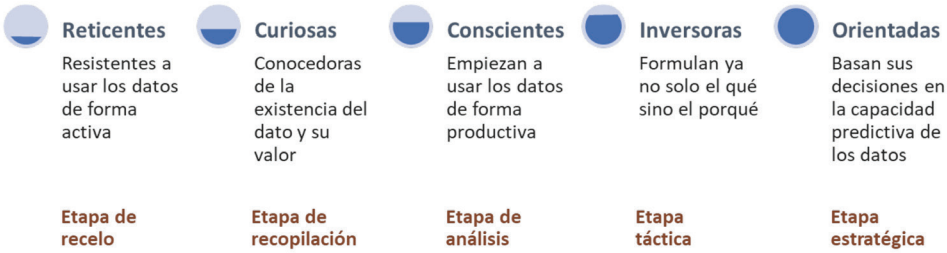

Figura 1-8. Evolución de las empresas con relación a los datos.

Nota. Adaptado de *The Evolution of the Data-Driven Company* [Figura], por Christopher S Penn, 2019, www.christopherspenn.com (*https://www.christopherspenn.com/2019/08/the-evolution-of-the-data-driven-company/*).

Propiedad	Contexto	Retos
Volumen	Escalabilidad	• Almacenamiento en paralelo, sin elementos compartidos • Elevado coste de la infraestructura necesaria
	Comunicaciones	• Capacidad y rendimiento de la infraestructura de red • Segregación de redes y localidad en el procesamiento del dato[11]
	Computación híbrida	• Movimiento de grandes volúmenes de datos entre distintos proveedores y el centro de datos local
Velocidad	Latencia de acceso	• Elevado consumo de ancho de banda de forma constante
	Agilidad de uso	• Necesidad de infraestructura escalable, distribuida y homogénea
	Tiempo de respuesta	• Mayor acceso al dato en memoria • Elevada concurrencia
	Seguridad	• Impacto de los mecanismos de seguridad en la latencia de acceso
Variedad	Tipología del dato	• Introducción de nuevas plataformas y tecnologías especializadas
	Integración	• Reconciliación y acceso a los datos en repositorios dispares • Consistencia del dato
Veracidad	Calidad	Complejidad en las estrategias de control de la calidad del dato
	Conformidad	Mayor esfuerzo en las políticas de gobierno del dato

Tabla 1-3. Retos del Big Data en cuanto a gestión y procesamiento.

11 La localidad del dato (*data locality*) hace referencia al hecho de procesar el dato donde reside, en lugar de moverlo y hacerlo en una ubicación central, minimizando así el tráfico de red y mejorando el rendimiento.

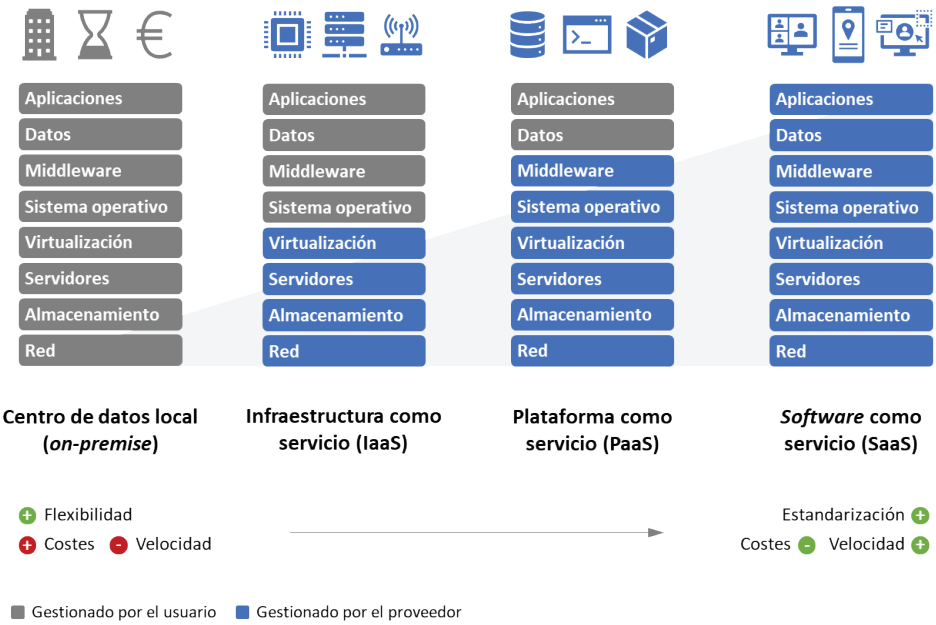

Figura 1-9. Modelos de servicio en la nube.

1.3.3 Computación en la nube

Podemos ver el *Big Data* como una respuesta a una serie de retos tecnológicos que plantearon en su momento las características de los datos que hemos estado viendo. La Tabla 1-3 los expone y pone en contexto. Muchos de ellos los iremos viendo a lo largo de los diferentes capítulos.

A la hora de gestionar grandes y variados volúmenes de datos, que se generan y mueven a gran velocidad, la consolidación del modelo de **computación en la nube** (*cloud computing*) fue decisiva. Lo fue en el ámbito tecnológico, sin duda, pero también muy especialmente en el económico.

La computación en la nube permite abstraer los distintos niveles que componen una aplicación, de forma que son gestionados por el proveedor de forma transparente y facturados por su uso[12]. Esto ofrece mucha flexibilidad a la hora de mover a la nube cargas de trabajo existentes, agilizando al mismo tiempo la creación de aplicaciones nativas. Básicamente nos podemos encontrar 3 modelos de consumo que se pueden combinar entre sí (Figura 1-9), siendo habitual que un mismo proveedor los ofrezca todos:

12 Es lo que se denomina una **facturación por suministro** (*utility services*), similar a la que realizan las compañías de gas, agua o electricidad

- Un modelo de **infraestructura como servicio** (**IaaS**, *Infrastructure as a Service*), que permite el aprovisionamiento, configuración y consumo de servidores físicos y virtuales, almacenamiento y elementos de red. Proporciona también funcionalidades de balanceo de carga, autoescalado alta disponibilidad o seguridad perimetral. Es el modelo que ofrece más flexibilidad, pero también un mayor esfuerzo de administración. Permite desplegar rápidamente aplicaciones existentes sin prácticamente necesidad de modificarlas.
- Un nivel por encima se sitúan los **servicios de plataforma** (**PaaS**, *Platform as a Service*). Aquí la abstracción cubre todos los elementos, exceptuando el propio código y los datos de las aplicaciones. Este nivel es muy amplio, ya que podemos encontrar desde componentes todavía muy próximos a la infraestructura, como plataformas de contenedores, hasta **entornos «sin servidor»**[13] (*serverless*) orientados a la gestión de eventos. También encontramos aquí servicios de base de datos, integración de aplicaciones, aprendizaje automático, inteligencia artificial, componentes analíticos o herramientas de desarrollo. Es decir, en este nivel dispondremos de piezas que podremos ensamblar mediante código para conformar nuestra aplicación
- Finalmente encontraríamos el modelo de ***software* como servicio** (**SaaS**, *Software as a Service*), donde consumimos, ya como usuarios finales, una aplicación (un cliente de correo electrónico, un procesador de textos o un gestor de campañas de *marketing*, por ejemplo). El proveedor se encarga de toda la gestión del ciclo de vida, incluyendo la liberación de nuevas versiones y la gestión de incidencias. La principal cualidad de este nivel sería la instantaneidad de la contratación y el acceso. Por el contrario, es el modelo menos flexible de todos, ya que las posibilidades de adaptación y programación suelen ser limitadas por motivos de estandarización.

El modelo en la nube permite separar el almacenamiento de la capacidad de proceso. En un centro de datos local es necesario dimensionar la potencia de cómputo de acuerdo con los picos de trabajo, aunque la mayoría del tiempo los procesadores estén ociosos. La flexibilidad que ofrece la nube permite desplegar un servidor físico, una máquina virtual o un contenedor en cuestión de segundos o minutos, acomodando rápidamente nuevas cargas de trabajo, planificadas o esporádicas. Esto es especialmente relevante en los sistemas de *Big Data*, donde los recursos consumidos son elevados, aunque de forma irregular en el tiempo.

Otra característica importante que tiene que ver con la anterior, es la capacidad de la nube para separar y gestionar de forma individual los distintos elementos que integran una solución. Esto permite descomponer aplicaciones monolíticas en distintos **microservicios**, que se desarrollan, actualizan y escalan de forma independiente,

13 El proveedor se encarga de dimensionar dinámicamente los recursos necesarios para ejecutar el código, facturándose por petición y/o memoria consumida.

proporcionando una gran flexibilidad y favoreciendo su reutilización. Esto conduce también al modelo de **nube híbrida** (*hybrid cloud*), donde distintos componentes residen y se ejecutan en nubes de distintos proveedores o en el centro de datos local, comunicándose entre sí de forma segura.

Adicionalmente, la existencia de múltiples **interfaces de programación** (**API**, *Application Programming Interface*) a nivel de servicio permite automatizar la mayoría de las tareas de gestión, tanto de forma planificada como en respuesta a eventos. El aprovisionamiento y la cancelación de servidores, la gestión de certificados y autorizaciones, el ensamblado en el despliegue de aplicaciones o la inferencia de modelos analíticos son algunos ejemplos de procesos que se pueden automatizar y orquestar.

Como veremos a lo largo de los diferentes capítulos, los sistemas de *Big Data* se pueden implementar empleando elementos de los tres niveles que acabamos de plantear. De esta manera se benefician de la facilidad de escalado y la flexibilidad en el despliegue, así como de la forma de facturación del modelo, adoptando los mejores componentes de cada proveedor, todo ello gestionado de forma controlada y segura.

1.3.4 Gestión y gobierno del dato

El beneficio de implantar en las empresas una disciplina alrededor del **gobierno del dato** (*data governance*) era algo que ya se había hecho patente en las últimas décadas. La consolidación del *Big Data*, con los retos que conlleva en cuanto a gestión de los datos, no hizo más que convertir ese beneficio en necesidad.

Cuando hablamos de gobierno del dato nos estamos refiriendo al conjunto de procesos, roles, políticas, estándares y mediciones que permiten el uso eficiente de los datos dentro de una organización. Va mucho más allá de un conjunto de buenas prácticas, convirtiéndose en un marco que orquesta y regula una empresa orientada por los datos, tal y como veíamos anteriormente. El gobierno del dato cubre un campo tremendamente amplio, y se apoya en un conjunto de tecnologías y soluciones que necesariamente deben cubrir todo el ciclo de vida del dato: captación, almacenamiento, transformación, consumo y eliminación, tal y como veíamos en la Figura 1-2.

Un aspecto transversal en el gobierno del dato es la **gestión de metadatos**. Los metadatos son propiedades de los datos, adicionales a su contenido, que nos sirven para documentarlos, tanto desde el punto de vista técnico como de negocio. Los datos parten con una serie de metadatos descriptivos que nos informan, entre otros, sobre su origen, creador o formato, y a lo largo de su ciclo de vida van incorporando otros metadatos acerca de las transformaciones que han sufrido, sus relaciones con otros datos, su calidad, significado de negocio, restricciones de uso, informes donde son consumidos, etc. Los metadatos son la base de cualquier actividad relacionada con el gobierno del dato, desde la catalogación hasta la publicación, pasando por el control del linaje y el análisis de impacto.

Componente	Función
Gestión de metadatos	Generación y mantenimiento de metadatos técnicos y de negocio a lo largo de todo el ciclo de vida
Catalogación y clasificación	Gestión de glosarios, diccionarios y categorías, incluyendo el etiquetado y la catalogación de activos
Definición de estándares	Unificación de los elementos de gobierno, desde la arquitectura de datos hasta la nomenclatura
Integración	Captura, transformación, almacenamiento y entrega del dato para su consumo
Gestión de la calidad	Cuantificación, monitorización e implantación de medidas para corregir y asegurar la calidad del dato
Linaje de activos	Control y análisis de impacto transversal sobre datos, procesos, modelos e informes
Búsqueda y autoservicio	Mecanismos de solicitud y acceso autónomo a los datos por parte de los usuarios finales
Control de acceso y privacidad	Gestión de aprobaciones en el tratamiento del dato, control de usuarios y políticas de enmascaramiento
Control regulatorio	Automatización de los procesos de auditoría, tanto interna como externa
Monitorización y medición	Seguimiento, documentación y medición de los diferentes procesos involucrados en la gestión

Tabla 1-4. Componentes de un marco para la gestión y el gobierno del dato.

La Tabla 1-4 resume los principales componentes que deberían formar parte de una correcta estrategia unificada de gobierno. Iremos viendo muchos de ellos a lo largo del libro, especialmente en el Capítulo 10.

1.4 ETAPAS DE ANÁLISIS EN LA EXPLOTACIÓN DE LA INFORMACIÓN

El aspecto fundamental y fundacional del *Big Data* es el análisis de grandes volúmenes de datos. Es decir, poner la información (llegados a este punto podemos referirnos al dato ya de esta manera) a disposición de los usuarios y analistas para tomar decisiones documentadas sobre la operativa, la táctica y la estrategia del negocio. La **analítica de negocio** (BA, *Business Analytics*) es el conjunto de aproximaciones, métodos y tecnologías encaminadas a la explotación de la información con ese fin[14].

14 Al hablar del contexto del *Big Data* hicimos referencia a la inteligencia de negocio (BI). Existe una importante confusión en el mercado y la literatura respecto a las diferencias y similitudes entre esta y la analítica de negocio. Muchos consideran a la segunda como un subconjunto de la primera, centrada en la modelización predictiva, el pronóstico (*forecasting*) y la visualización de datos. Para otros es exactamente al revés, y hay quien las ubica una al lado de la otra, encargándose la inteligencia de la parte descriptiva y la analítica de la predictiva. Con una perspectiva histórica, mi opinión es que son básicamente la misma cosa. Sin embargo, creo que hablar de analítica de negocio es más explícito o, en todo caso, menos pretencioso.

Podemos hablar de 4 categorías dentro de la analítica de negocio (Figura 1-10). Lo habitual es que las empresas las vayan abordando de forma progresiva, construyendo una sobre la otra. Por este motivo, se suelen hacer corresponder con grados de madurez empresarial en cuanto a la orientación por los datos[15]. Aunque dedicaremos un capítulo a cada una de estas categorías, es importante plantearlas brevemente en conjunto.

Figura 1-10. Estadios de madurez analítica dentro del negocio.

1.4.1 Analítica descriptiva

La analítica descriptiva proporciona información sobre el rendimiento pasado del negocio y su contexto, respondiendo a preguntas como:

- Cuál fue el número de piezas defectuosas en cada una de las fábricas durante el último trimestre.
- Cómo ha variado la rentabilidad promedio por metro cuadrado de las tiendas respecto al último año.
- Qué relación existe entre los días de lluvia y el incremento en la venta de paraguas.

15 Se pueda abordar directamente un análisis predictivo, de propensión al abandono de clientes, por ejemplo. Sin embargo, lo habitual es tener antes resuelta la analítica descriptiva.

Aunque es un análisis reactivo, esto no quiere decir que no pueda ser inmediato, ya que puede involucrar datos en tiempo real que se acaban de producir. Este tipo de análisis se vale de informes, tanto planificados como a medida[16], y de cuadros de mando interactivos (*dashboards*) que permiten al usuario consultar y navegar de forma fácil por la información en modo autoservicio (Figura 1-11).

Figura 1-11. Ejemplo de cuadro de mando interactivo desarrollado con IBM Cognos Analytics.

1.4.2 Analítica prescriptiva

Aquí nos adentramos ya en el terreno de las recomendaciones, tanto en forma de planificación, presupuestación u optimización en general. Este tipo de análisis tiene un enfoque más operativo y de proceso, ya que busca detallar la mejor solución para una situación determinada. Por ejemplo:

16 Los informes planificados son aquellos que su ejecución y distribución está programada, realizándose de forma periódica. Los informes a medida, o informes *ad hoc*, son aquellos generados de forma individual por un usuario para satisfacer una necesidad concreta.

- Organizar los turnos y las rotaciones de las tripulaciones en una compañía aérea, teniendo en cuenta restricciones operativas y laborales que condicionan la planificación.
- Establecer las ubicaciones más adecuada para situar una serie de centros logísticos con el fin de abastecer los puntos de venta lo más rápido posible, incurriendo en los mínimos costes.
- Determinar y planificar qué tipo de planta generadora debe ponerse en funcionamiento en cada franja y en que secuencia para abastecer la demanda de energía, teniendo en cuenta el tiempo de arranque y los periodos de mantenimiento obligatorios.
- Definir la estrategia de precios más adecuada para el petróleo, considerando los niveles de producción en cada momento, la demanda y la situación geopolítica.

La tecnología para este tipo de análisis se basa en la combinación de reglas de negocio con modelos matemáticos de optimización y simulación (*what-if analysis*), incluyendo técnicas de programación lineal y no lineal, o algoritmos genéticos y cadenas de Markov.

1.4.3 Analítica predictiva

La analítica predictiva se basa en el descubrimiento de patrones, tendencias y relaciones que permiten explicar un comportamiento a partir de datos históricos con el fin de anticiparse a él en el futuro.

Como disciplina, la **minería de datos** proporciona técnicas basadas en el análisis estadístico multivariante y el **aprendizaje automático** (*machine learning*) que van desde la regresión, la clasificación o la segmentación (*clustering*) hasta el análisis de series temporales o la detección de asociaciones y patrones secuenciales.

Por ejemplo, la Figura 1-12 muestra un modelo de árbol de decisión. Su objetivo es clasificar clientes en bandas de ingresos anuales, teniendo en cuenta una serie de atributos como el país de nacimiento, la educación, el estado civil, la edad, el sexo, la plusvalía obtenida mediante la venta de activos, etc. Para ello evalúa estas características generando reglas que permiten dicha asignación de forma automática. La idea es que el modelo generalice ese patrón, de forma que pueda ser aplicado a nuevos clientes cuyos ingresos son desconocidos, soportando determinadas acciones comerciales.

El verdadero valor de los modelos predictivos está en su puesta en producción. Esto quiere decir aplicarlos en entornos operacionales, integrándolos con otros sistemas y aplicaciones del negocio, con el fin guiar acciones directas como la detección de fraude en transacciones, la concesión de cupones de descuento en supermercados o la selección de público objetivo para campañas, por ejemplo.

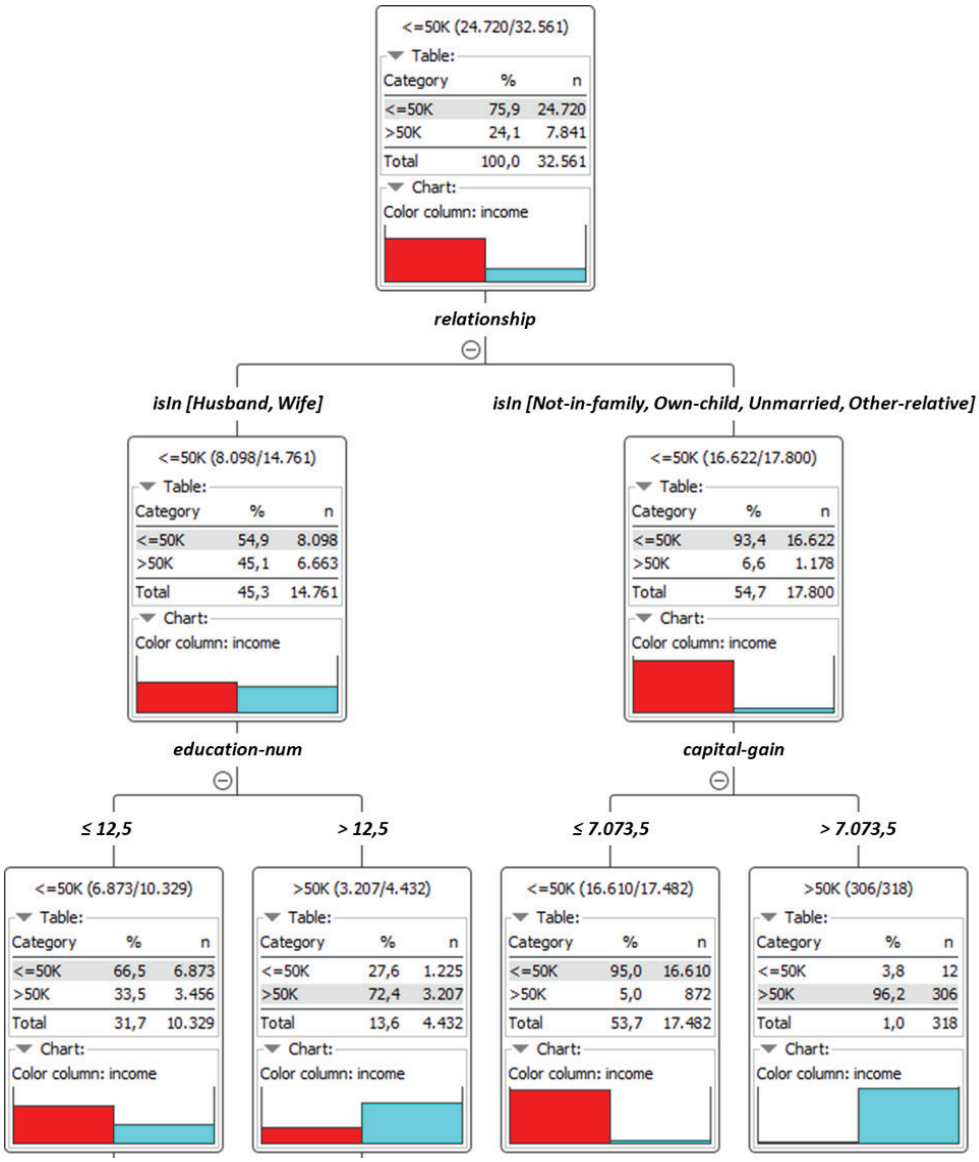

Figura 1-12. Árbol de clasificación binario desarrollado con KNIME Analytics Platform.

1.4.4 Analítica cognitiva

La **computación cognitiva** refleja el estado del arte de la tecnología en cuanto al procesamiento y el análisis de la información. La **inteligencia artificial**, a través de los **sistemas simbólicos** y del **aprendizaje profundo** (*deep learning*), está avanzando rápidamente en tareas como el **procesamiento del lenguaje natural** (**NLP**, *Natural Language Processing*), el **reconocimiento del habla** o la **clasificación de imágenes**.

El objetivo es, ni más ni menos, que desarrollar sistemas con capacidad para entender, razonar e interactuar emulando a los seres humanos. Además de su aplicación directa en campos como la robótica, la visión artificial o los asistentes virtuales, estas tecnologías marcan la diferencia a la hora de analizar resultados. Por ejemplo:

- Facilitando la preparación e integración de la información, recomendando como transformar y enlazar datos de fuentes dispersas mediante la comprensión semántica de su contenido.
- Revelando patrones y relaciones entre los datos que son difíciles de detectar, sugiriendo nuevos cruces de información.
- Recomendando las mejores formas de representar y visualizar los datos, interpretando su significado y contexto.
- Respondiendo a preguntas en lenguaje natural, de forma clara y concisa, acelerando la navegación sobre los datos

Veremos como el aprendizaje profundo permite desarrollar e implementar este tipo de sistemas. También en qué consisten los retos asociados con el procesamiento de imágenes y del lenguaje natural, especialmente en lo referente a sesgo y equidad.

1.5 ESCENARIOS DE APLICACIÓN DEL *BIG DATA*

Vamos a concluir el capítulo con una relación de aplicaciones de *Big Data* en distintos sectores y abordando distintos problemas de negocio. El objetivo no es otro que ilustrar los conceptos que hemos presentado, facilitando su comprensión y valorando su potencial. La Tabla 1-5 las contiene.

Área	Objetivo	Descripción
Medios de pago	**Detección de fraude en compras**	Predecir la probabilidad de que una transacción con tarjeta de pago sea fraudulenta. Mediante el análisis en tiempo real del perfil de esta, es posible generar recomendaciones que pueden llegar incluso a denegarla de forma automática
Finanzas	**Monitorización del mercado bursátil**	Seguimiento en tiempo real de las cotizaciones en bolsa, con el fin de anticipar movimientos en los precios de las acciones y guiar las estrategias de compraventa
Finanzas	**Gestión del riesgo en las inversiones**	Mejorar la estimación del resultado y la rentabilidad de las inversiones mediante modelos predictivos, mitigando al mismo tiempo el riesgo asociado
Seguros	**Personalización y tarificación de pólizas de salud**	Elaborar un perfil del asegurado en base a su historial médico, hábitos y actividad registrada por los dispositivos móviles. De esta manera es posible adaptar las coberturas de su póliza y establecer la tarifa más adecuada para él y la compañía
Seguros	**Evaluación del riesgo en seguros del automóvil**	Mejorar la segmentación y asignación de los clientes a distintas bandas de riesgo de acuerdo con la probabilidad de tener un accidente
Seguros	**Detección de fraude en visitas médicas**	Detectar asociaciones y patrones repetitivos que indiquen secuencias de consulta involucrando distintos especialistas sin que medie una causa
Seguros	**Inspección de partes de daños en seguros del hogar**	Implementación de un sistema de identificación y validación de daños mediante reconocimiento de imágenes, que permita agilizar su cuantificación y acelerar el pago de las indemnizaciones
Servicio al cliente	**Monitorización de llamadas**	Análisis de las transcripciones de las llamadas para identificar el grado de satisfacción de los clientes al principio y a la conclusión, analizando también el tono conversacional de los agentes
Servicio al cliente	**Clasificación de encuestas de satisfacción**	Enrutamiento automático de encuestas de satisfacción al departamento más apropiado según el contenido de campos de texto libre
Distribución	**Ubicación de los productos en los lineales o en la tienda en línea**	Identificar asociaciones de productos que son comprados de forma conjunta con el fin de ubicarlos juntos en la tienda física o en la disposición de la página web, facilitando también el diseño de promociones a medida
Distribución	**Recomendación de productos**	Construir un motor de emparejamiento que asocie cada cliente con un conjunto de productos según su perfil e historial de compra

Tabla 1-5. Relación de aplicaciones del Big Data por área.

Área	Objetivo	Descripción
Distribución	**Generación de cupones promocionales**	Análisis en tiempo real del contenido de la cesta de la compra al pasar por el terminal de punto de venta, y generación de un cupón de descuento en base a su contenido y la probabilidad de redención de este
Marketing	**Análisis de sentimiento ante nuevos lanzamientos**	Analizar el sentimiento de los mensajes de Twitter y otras redes sociales para determinar lo que los usuarios están opinando sobre la compañía y sobre los nuevos productos que acaban de ser lanzados
Logística	**Diseño de almacenes y optimización de rutas**	Diseño de la ubicación de los productos en cada almacén teniendo en cuenta la composición de los pedidos, minimizando el número de ellos necesarios para satisfacer el envío
Telefonía	**Prevención del abandono y fidelización de clientes**	Utilizando datos históricos, identificar perfiles de clientes que concentran una alta probabilidad de abandono, determinando sendas de desvinculación a medio plazo y prediciendo su salida a corto, con el fin de retener a los más rentables y fidelizarlos
Telefonía	**Identificación de fraude en llamadas telefónicas**	Con el uso de modelos de detección de anomalías es posible analizar los registros de llamadas para identificar usuarios compinchados en la realización de llamadas a servicios de tarificación especial
Energía	**Predicción y adaptación de la demanda**	Mediante el empleo de sensores en las centrales de producción y de contadores inteligentes en los puntos de consumo, analizar en tiempo real la demanda para adaptar y predecir la generación
Salud	**Identificación de anomalías en imágenes médicas**	Implantación de sistemas de clasificación automática de imágenes que son capaces de identificar tumores y otras complicaciones después de un proceso de entrenamiento con datos históricos
Salud	**Extracción de términos en historiales médicos**	Procesado de historiales médicos y clínicos textuales con el fin de identificar y sistematizar síntomas, dolencias y tratamientos, detectando relaciones causa-efecto de interés
Seguridad	**Análisis de archivos de registro de sistemas**	Recopilación, monitorización, análisis, correlación, búsqueda y almacenamiento de archivos de registro (*logs*) provenientes de múltiples y variados sistemas para detectar posibles amenazas de seguridad y responder de forma anticipada y automática
Medios	**Gestión de estadísticas en eventos deportivos**	Seguimiento y documentación de la actividad de los jugadores durante un partido, combinando estadísticas al segundo y datos históricos que son presentados en tiempo real

Tabla 1-5 (continuación). Relación de aplicaciones del Big Data por área.

1.6 RESUMEN DEL CAPÍTULO

En este capítulo introductorio hemos planteado en que consiste el *Big Data* como concepto, las necesidades que aborda, los retos a los que se enfrenta en términos de la gestión del dato, y los beneficios que aporta soportando distintas categorías de análisis y aplicaciones.

- ***Big Data*** es el conjunto de operaciones, técnicas y tecnologías orientadas al procesamiento de grandes y variados volúmenes de datos, con el fin de generar información válida sobre la que desarrollar conocimiento y soportar las decisiones de negocio.
- Los retos a los que se enfrenta el *Big Data* en la gestión y procesamiento de los datos tienen que ver con su **volumen**, su **velocidad** de generación y consumo, la **variedad** en cuanto al formato y el aseguramiento de su calidad y **veracidad**.
- El modelo de **computación en la nube** permite un despliegue de las soluciones de *Big Data* de una forma flexible y económica, dando acceso a capacidad de cómputo y almacenamiento bajo demanda, así como a las últimas tecnologías a nivel de infraestructura, componentes y aplicaciones.
- El **gobierno del dato** es un aspecto fundamental a la hora de establecer y asegurar una cultura empresarial orientada por los datos. Se compone de un conjunto de procesos, roles, políticas, estándares y mediciones que permiten el uso eficiente de los datos dentro de una organización.
- Podemos dividir la explotación de la información en cuatro categorías de análisis: **descriptivo**, **predictivo**, **prescriptivo** y **cognitivo**. Cada uno de ellos supone un estadio de madurez de las organizaciones en cuanto a la toma de decisiones basada en la evidencia.

En el siguiente capítulo abordaremos los distintos patrones arquitecturales de datos que nos podemos encontrar a la hora de desplegar una infraestructura y plataforma de *Big Data*. Como veremos, la caracterización de los datos con lo que trabajemos tendrá mucho que ver.

2

ARQUITECTURAS Y PATRONES PARA *BIG DATA*

Las **arquitecturas de datos** vienen a describir la forma en que estos son organizados y gestionados dentro de una entidad. Cubren todo su ciclo de vida, desde la ingestión inicial hasta la explotación final, pasando por el procesamiento y el almacenamiento, todo ello de forma gobernada.

En este capítulo vamos a plantear las dos principales formas de organizar los datos que podemos encontrar en las empresas: la **arquitectura centralizada**, monolítica y agnóstica de las distintas áreas que hay en la organización, y la **arquitectura orientada por dominios**, descentralizada en cuanto a la propiedad y la elaboración del dato, pero unificada en lo referente al gobierno. Esta última supone una visión muy actual (por lo tanto, con poca implantación todavía) y rompedora de cómo organizar los datos, con grandes implicaciones a nivel de cultura empresarial. Veremos la forma en que ha ido evolucionando la arquitectura centralizada, creando y adoptando los distintos patrones de los que se compone, y llegaremos a ver como estos mismos se acomodan y adaptan en la reciente concepción orientada por dominios. El objetivo no es otro que introducir conceptos que iremos desarrollando en los sucesivos capítulos.

2.1 PATRONES ARQUITECTURALES

Una arquitectura de datos debe dar respuesta a requerimientos de negocio. Como tal, se centra en plantear una visión abstracta de alto nivel sobre cómo afrontar estos, admitiendo posteriormente distintos diseños por parte de arquitectos e ingenieros de cara a su implementación. Es decir, la arquitectura no dice nada de los productos o servicios concretos sobre los que se acabará materializando, pero sí sobre qué componentes hay que tener en cuenta, qué funciones deben asumir y cómo deben estar relacionados, y lo hace mediante una combinación de patrones arquitecturales.

Un **patrón arquitectural** viene a ser una buena práctica; un modelo que, debido a su uso reiterado, ha demostrado ser satisfactorio para resolver una problemática en un contexto concreto. En el nuestro, disponemos de patrones para resolver las necesidades que acabamos de enumerar en lo referente al ciclo de vida del dato. Es importante resaltar que, mientras una arquitectura concreta utiliza una serie de patrones de una determinada forma, otra podrá orquestar esos mismos patrones de una manera diferente[17].

2.1.1 Tipologías de patrones

Antes de nada, vamos a plantear las tipologías básicas a la hora de clasificar los distintos patrones necesarios para componer una arquitectura de datos.

Tipología	Propósito	Ejemplos
Integración	Reconciliar los datos provenientes de distintos orígenes con el fin de garantizar un acceso y entrega consistente, todo dentro de un marco gobernado	• ETL & ELT • Lambda • Kappa • Federación
Modelado	Organizar y representar los datos de la forma más eficiente para su consumo y almacenamiento	• Entidad-Relación • Multidimensional • *Schemaless*
Almacenamiento	Persistir los datos de cara a su posterior acceso de acuerdo con su formato y la forma en que serán consumidos	• OLTP • *Data warehouse* • ODS • *Data lake*
Análisis	Extraer conocimiento de la información para soportar la toma de decisiones dentro del negocio	• Análisis descriptivo • Análisis prescriptivo • Análisis predictivo • Análisis cognitivo

Tabla 2-1. Tipos de patrones dentro de una arquitectura de datos.

La Tabla 2-1 los condensa, pero básicamente se corresponden con las distintas etapas del ciclo de vida del dato. Algunos de estos patrones ya los comentamos en el capítulo anterior, especialmente los relativos al análisis.

Por ejemplo, el **análisis predictivo** los podemos ver como un patrón arquitectural que proporciona una solución reutilizable, en forma de metodología, componentes y técnicas, dirigida a dar respuesta a la necesidad del negocio en cuanto a la detección de patrones y tendencias que subyacen en sus grandes volúmenes de datos. Como veremos en su correspondiente capítulo, podemos a su vez dividir el análisis predictivo en subpatrones que nos indican como abordar problemáticas concretas, como la detección de patrones secuenciales, la identificación de anomalías y desviaciones o la clasificación de datos.

17 Estamos adoptando aquí una noción muy sintetizada de arquitecturas y patrones, siendo este un campo muy amplio y matizado dentro de la ingeniería de sistemas.