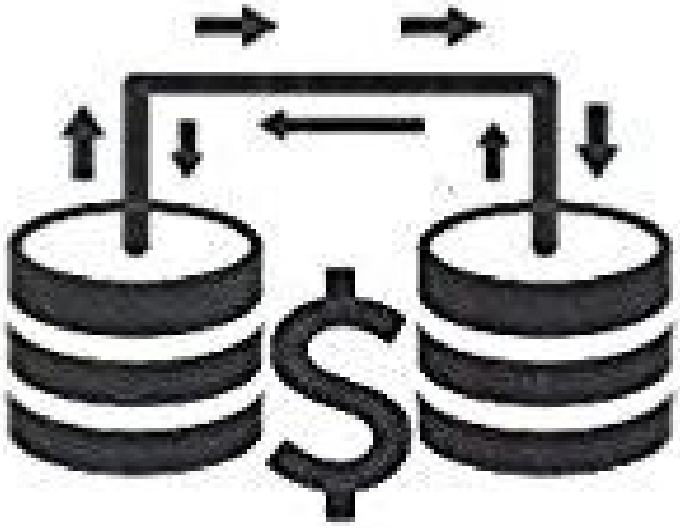
El profesional de la información



BIG DATA

GESTIÓN Y EXPLOTACIÓN DE GRANDES VOLÚMENES DE DATOS

MONTSERRAT GARCÍA-ALSINA

EDITORIAL UOC

Big data

Gestión y explotación de grandes volúmenes de datos

Montserrat García-Alsina

Director de la colección: Javier Guallar

Diseño de la colección: Editorial UOC
Diseño de la cubierta: Natàlia Serrano

Primera edición en lengua castellana: marzo 2017
Primera edición en formato digital (epub): noviembre 2017

Rambla del Poblenou, 156
08018 Barcelona
http://www.editorialuoc.com

Realización editorial: Oberta UOC Publishing, SL
ISBN: 978-84-9116-712-9

Montserrat García-Alsina

Doctora en Sociedad de la Información y el Conocimiento, profesora de la Universitat Oberta de Catalunya e investigadora del grupo de investigación KIMO (Knowledge and Information Management in Organizations) de la misma universidad.

A QUIÉN VA DIRIGIDO ESTE LIBRO

Este libro te interesa si quieres saber...

- Qué es *big data*
- Porqué ya no hablamos de información sino de datos
- Cómo se gestionan los datos y quién lo puede hacer
- Con qué instrumentos y tecnología se cuenta
- Para qué se gestionan grandes volúmenes de datos
- Por qué ya no nos sirve lo que teníamos hasta hace poco
- Qué relación existe entre gestión de macrodatos, gestión de información, gestión de documentos, inteligencia competitiva, gestión del conocimiento y gobernanza de la información

Índice

PRESENTACIÓN

A lo largo de este libro se expone en qué consiste lo que conocemos como *big data* y se presentan diversos aspectos relacionados con la gestión de grandes volúmenes de datos, así como la vinculación de estos con la gestión de la información y el conocimiento.

En las últimas décadas habíamos aprendido que los datos eran la base de la información para generar conocimiento. A su vez, este era considerado la materia prima de la economía de la información y del conocimiento, motor del desarrollo de la nueva sociedad. En las organizaciones se hablaba de cómo gestionar la información mediante sistemas de información. La palabra que más se utilizaba hasta hace unos años era, efectivamente, *información*.

Pero en los últimos años, progresivamente, se ha empezado a hablar más de *datos* que de *información*. La evolución de las tecnologías ha generado grandes volúmenes de datos y, con ellos, la necesidad de gestionarlos para crear valor como paso previo a la generación de información. No obstante, la gestión de información y conocimiento en las organizaciones sigue existiendo. Estas continúan almacenando datos, documentos e información en sus sistemas de información, y continúan generando conocimiento. La diversidad de sistemas de información en una misma organización ha hecho necesario crear procesos y procedimientos para garantizar la gobernanza de la información. Por eso, este libro también incluye la vinculación entre todos estos conceptos: gestión de datos masivos, información, documentos, conocimiento y gobernanza de la información.

Además, en este contexto vemos que los grandes volúmenes de datos y la diversidad de formatos están impulsando el desarrollo de tecnologías y de sistemas de información denominados *big data*. Y todos los esfuerzos en gestión e inversión en tecnología se quedan en saco roto, si no se aplican técnicas de análisis de datos para extraer conocimiento para la acción. Este es el que crea ventaja competitiva en las organizaciones, y por tanto, valor. Por todo ello, este libro explica términos como Hadoop, MapReduce, *data mining*, *text mining* o NoSQL, vinculados con tecnología y análisis.

Por último, se constata que la complejidad de la gestión de datos, información y conocimiento requiere de profesionales especializados. Por ello, este libro también dedica un capítulo a los nuevos profesionales vinculados a la gestión de *big data*, haciendo un repaso a sus habilidades, competencias y conocimientos.

Todas estas cuestiones se presentan en el libro en cinco capítulos:

El primer capítulo expone el concepto de *big data* y las características de los grandes volúmenes de datos.

El siguiente explica cómo gestionar eficientemente esos datos, detallando las acciones involucradas en las cuatro fases de la cadena de valor de los macrodatos, los pasos para diseñar e implantar un proyecto, cómo garantizar la calidad de los datos y cómo unir la gestión del conocimiento, la inteligencia competitiva y la gestión de datos masivos para asegurar la gobernanza de la información.

El tercer capítulo presenta la aplicación del resultado del análisis de los datos masivos tanto en la gestión interna de la organización como en la generación de productos, servicios o mejoras, para generar conocimiento para la acción.

A continuación se muestran las herramientas existentes en la actualidad para gestionar macrodatos.

Para acabar, el último capítulo detalla las competencias y los conocimientos que deben tener los profesionales involucrados en la gestión de datos masivos.

Es así como, tras finalizar la lectura de todos estos capítulos, el lector, la lectora, adquirirá un conocimiento claro y básico de lo que entendemos en la actualidad por *big data*, y una sólida base para ampliar en su caso estos conocimientos con literatura más especializada.

Capítulo I

¿QUÉ ES *BIG DATA*?

Big data o datos masivos es un concepto abstracto que en los años recientes se ha convertido en una palabra de moda en diferentes ámbitos: negocios, ingeniería informática, información y documentación, o sistemas de información, entre otros.

Cuando acumulamos grandes volúmenes de datos, se nos plantea la necesidad de ver qué podemos hacer con ellos. Esto nos remite a la necesidad de gestionarlos con una finalidad organizativa, a disponer de tecnología y metodologías específicas. La misma gestión de datos nos lleva a generar información que sea relevante en el contexto de la organización, es decir, a generar conocimiento para la acción, y que sea aplicable por ejemplo a la toma de decisiones, al diseño de acciones para la organización o a la elaboración de planes estratégicos.

Por lo tanto, cuando hablamos de datos masivos, estamos hablando también de gestión de la información y de generación de conocimiento para la acción. Este campo científico es el que nos da las pautas metodológicas para gestionar grandes volúmenes de datos con el fin de crear valor mediante una serie de procesos y procedimientos. Pero, además, hemos de tener en cuenta que debemos contar con la tecnología, para capturar datos, procesarlos, analizarlos e interpretarlos de manera rápida y eficiente (Gómez y Conesa, 2015). Y justamente, hacer evolucionar la tecnología para responder a estas necesidades es uno de los retos que las bases de datos plantean a la ingeniería informática, a los sistemas informáticos. Por lo tanto, las metodologías que se trabajan en la gestión de la información y en el ámbito de la tecnología ofrecen un marco idóneo para estudiar qué son y cómo gestionar datos masivos.

Así pues, cuando definimos datos masivos tenemos que dar respuesta a una serie de interrogantes: ¿Para qué sirven? ¿Cómo gestionarlos? ¿Con qué?

Siguiendo estos interrogantes, tenemos varias definiciones de datos masivos, dependiendo del punto de vista del que se parta. Así, hay definiciones que tienen en cuenta: el significado de la misma palabra y sus implicaciones; el cambio de paradigma para analizar datos, originado por la existencia de grandes volúmenes de datos; las características de los datos masivos; o las tecnologías, técnicas y metodologías relacionadas con su procesamiento.

Estos aspectos forman parte no solo de las definiciones de datos masivos, como veremos más adelante, sino también de los temas que hemos de considerar cuando hablamos de *big data* o grandes volúmenes de datos. A pesar de la popularidad del término *big data* en nuestra lengua, a lo largo del libro utilizaremos indistintamente el término datos masivos, grandes volúmenes de datos o macrodatos. Este último término es la voz recogida por la Fundación del BBVA Español Urgente. En estos momentos, la Real Academia de la Lengua española aún no recoge un término para denominar *big data* en nuestra lengua.

A continuación tratamos con detalle diferentes definiciones y los aspectos que los expertos tienen en cuenta cuando se refieren a datos masivos.

1. Componentes de las definiciones

El término se suele definir predominantemente a partir de cómo hacer frente a su gestión y los cambios que supone con relación a la gestión tradicional de los datos y su aplicación. Veremos en detalle los siguientes aspectos: significado del término, cambio de paradigma, las características de los grandes volúmenes de datos y la tecnología asociada a la gestión de datos masivos.

Significado del término

Las mismas palabras datos masivos nos remiten a la acumulación de un considerable volumen de datos. Por lo tanto, el volumen es la primera característica de la definición, y marca indirectamente una de las novedades de los datos masivos, y una de las limitaciones de las aplicaciones que en las últimas décadas gestionaban datos. Aun así, por muy amplio que sea el volumen de datos, el conjunto de ellos por sí solo y sin procesar no nos aporta información. Por lo tanto, necesitamos incluir más aspectos en la definición. De aquí que el término *datos masivos* a menudo se defina desde otras perspectivas además del volumen, como, por ejemplo: ¿para qué sirve la acumulación de datos?, ¿qué hacemos con todos los datos?, ¿cómo almacenar y gestionar la amplia heterogeneidad del conjunto de datos?, ¿qué requerimientos debe tener la infraestructura de software y hardware?, ¿qué valor aporta a quien acumula datos? y ¿cómo generamos valor con la gestión de estos grandes volúmenes de datos?

Teniendo en cuenta estos aspectos, hay autores que definen *datos masivos* como un conjunto de datos que exceden la capacidad de procesamiento de las bases de datos convencionales dentro de un tiempo adecuado (Chen y otros, 2014; Padgavankar y Gupta, 2014). Esta definición ya nos hace pensar que es necesario contar con otros tipos de bases de datos y, por lo tanto, de tecnología. Pero hay más aspectos que tener en cuenta en la definición, y los veremos en los siguientes apartados.

Cambio de paradigma

Los datos masivos y los medios para su explotación han traído un cambio en la forma de extraer datos y analizarlos. Este cambio de paradigma es la base empleada por algunos autores para definir los datos masivos (Gómez y Conesa, 2015; Mayer-Schönberger y Cukier, 2013). El nuevo paradigma tiene en cuenta los siguientes aspectos:

- Nuevas maneras de hacer investigación y estudios de mercado, al prescindir del muestreo. En el momento de hacer estudios con gran volumen de datos, mediante las nuevas tecnologías que están surgiendo se pueden emplear todos los datos recogidos, y, por lo tanto, se prescinde del muestreo.
- Al no trabajar con un muestreo, se evitan los errores inherentes a este y se puede trabajar a nivel macro, viendo conexiones y agrupaciones que a primera vista pasan desapercibidas, y

extraer patrones.

- La causalidad de los hechos ya no es el objetivo del análisis de los datos, sino que lo que interesa es saber qué está pasando mediante el descubrimiento de perspectivas nuevas que han pasado desapercibidas a primera vista, gracias al análisis de las conexiones entre datos, pautas y correlaciones.

Características de los grandes volúmenes de datos

Los datos masivos también se han definido teniendo en cuenta otras características. Además de su volumen, en la definición se han considerado aspectos como la variedad de datos y la velocidad de procesamiento. Sucesivamente otros autores, cuando profundizan en su estudio han ido añadiendo más *V*. Por lo tanto, a las tres *V* iniciales (volumen, variedad y velocidad) se añadieron otras características vinculadas con calidad de los datos, tratamiento y gestión, o prestaciones de tecnología y software. Estas otras características son: veracidad, valor, visualización, verificación, variabilidad y viabilidad (Chen y otros, 2014; Padgavankar y Gupta, 2014).

A pesar de que la descripción de las características de los datos masivos a partir de las *V* es la más conocida, existen otras muchas. Todas ellas ejemplifican la proliferación de definiciones propias de una disciplina en formación (instrumentos, líneas de investigación, formación del corpus teórico y comunidad científica). A continuación vemos algunas de las características asociadas a los datos masivos.

a) Volumen

Cuando se habla de volumen de los datos masivos se hace referencia a dos aspectos. En primer lugar, al incremento exponencial de los datos, fruto del uso de las nuevas tecnologías y la facilidad de generar datos digitales, existentes tanto en el ámbito general como en ámbitos específicos: sectores industriales, sociales o públicos (por ejemplo, finanzas, servicios, manufacturas, ciencias de la vida, física, astronomía, canal Twitter de una empresa, datos de la actividad de negocio, datos del censo de población, registros médicos, impuestos, etc.). En segundo lugar, al hablar de volumen también se hace referencia a los retos que supone recoger, almacenar, procesar e integrar grandes volúmenes de datos procedentes de fuentes muy variadas y distribuidas.

A pesar de que no hay una definición explícita sobre lo grandes que tienen que ser los conjuntos de datos (datasets) para ser considerados datos masivos, sí que se suele considerar como referencia la limitación de los software tradicionales para capturar, procesar, almacenar y analizar datos. Por lo tanto, podemos considerar la limitación de almacenamiento de las bases de datos tradicionales y la diversidad de las fuentes de datos (estructuradas, semiestructuradas y no estructuradas) como la frontera para hablar de grandes volúmenes de datos.

Otro aspecto que se debe tener en cuenta son las unidades empleadas para medir la cantidad de información acumulada en los dispositivos de almacenamiento. Las unidades pueden estar basadas en dos sistemas: el binario o el decimal. Hay que tener en cuenta esta diferenciación por

las diferencias numéricas implícitas, porque puede dar lugar a confusión con consecuencias relevantes en las cantidades totales. Véanse las siguientes equivalencias entre el sistema numérico binario y el decimal.

Prefijo binario: **1 kibibyte = 1.024 bytes** = 2^{10} bytes

Prefijo del SI: **1 kilobyte = 1.000 bytes** = 10^3 bytes

Por un lado encontramos las unidades basadas en el sistema de numeración decimal, que son de base 10 y parten de la unidad *byte*. Toman el concepto de los prefijos del Sistema Internacional (SI) para nombrar a los múltiplos y submúltiplos de cualquier unidad del Sistema Internacional. Estas unidades son: kilobytes, megabytes, gigabytes, terabytes, petabytes, exabytes, zettabytes y yottabyte. Por otro lado, tenemos las unidades basadas en el sistema de numeración binario. Inicialmente se crearon múltiplos del byte pero con base binaria, por lo cual había confusiones. Las unidades basadas en el sistema binario están reguladas por los organismos International Electrotechnical Commission (IEC) e International Organization for Standardization (ISO) que han trabajado las unidades específicas para la ciencia de la información y tecnología. En 1999 se publicó una norma de la IEC (60027-2) para diferenciar la medición binaria de la decimal. Cogiendo los dos primeros dígitos del sistema decimal y añadiendo el sufijo -bi (de binario), la norma introdujo los siguientes prefijos: *kibi-*, *mebi-*, *gibi-*, *tibio-*, *pebi-* y *exbi-*, a los que en 2005 se añadieron *zebi-* y *yobi-*. Esta norma fue adaptada por la International Organization for Standardization como norma IEC 80000. La norma actual vigente es la 80000-13:2008, y está preparada y mantenida por el comité técnico 12 de la International Organization for Standardization (ISO/TC 12) y el comité técnico 25 de la International Electrotechnical Commission.Actualmente las unidades de información más empleadas son las basadas en el sistema de numeración decimal. Aun así, hay algunas iniciativas para emplear las unidades basadas en el sistema de numeración binario, como es el caso de GNU/Linux.

En definitiva, la gestión de grandes volúmenes de datos es una característica determinante para impulsar nueva investigación y el desarrollo de plataformas capaces de capturar, almacenar, analizar y visualizar estos grandes volúmenes de datos.

b) Variedad

Los datos son muy variados en cuanto a tipologías, formato y estructuras empleadas para su organización y presentación. Esto se debe a que proceden de fuentes muy variadas y además tienen que adaptarse a los requerimientos de los diferentes dispositivos que generan y emplean los datos (móviles, audio, vídeo, sistemas GPS, sensores de temperatura, medidores de viento, capital, sensores RFID, mensajes de Twitter, webs, internet de las cosas, etc.).

Teniendo en cuenta las estructuras en las que se organizan los datos, y las formas en las que se almacenan, podemos clasificar los datos según el nivel de estructuración: estructurados, no estructurados y semiestructurados (Gómez y Conesa, 2015).

Los datos **estructurados** se almacenan en campos de tablas de bases de datos relacionales

donde su longitud, denominación y formato han sido predefinidos. Por lo tanto, conocemos anticipadamente su organización, estructura, el tipo, su posición y las posibles relaciones entre ellos. La información se representa por datos elementales, no compuestos por otras estructuras. Las consultas a estas bases de datos se hacen mediante lenguaje SQL.

Ejemplos de bases de datos disponibles en las organizaciones son:

- Los ERP (*enterprise resource planning*), que recogen información en lo referente a la producción, distribución, inventario de compras y de productos fabricados, facturas y contabilidad.
- Los CRM (*customer relationship management*), que recogen datos de los clientes, desde la domiciliación hasta el historial de visitas, ventas, acciones de marketing, etc.
- Los SCM (*supply chain management*), que recogen los datos de cada una de las partes de la cadena de suministro, incluyendo proveedores y clientes, compras de materias primas para la producción, almacenamiento, gestión de existencias, producción y entrega de pedidos a clientes.

Los datos **no estructurados** se documentan en el formato en el que han sido creados, y por lo tanto no tienen una estructura predefinida, ni están almacenados en una tabla. La información no está representada por datos elementales, sino por una composición cohesionada de unidades estructurales de nivel superior. La interpretación y manipulación de estos datos es más compleja (Gómez y Conesa, 2015).

Ejemplos son documentos en PDF o en Word, imágenes, vídeo, audio, etc. Estos documentos están almacenados también en aplicaciones, como los ECM (*enterprise content management*), o los EDRMS (*electronic focuments and records management systems*). Estos son sistemas de gestión de contenidos que combinan tecnologías de sistemas de gestión de los *records* (documentos específicos generados por las organizaciones como evidencia de su actividad –que requieren un tratamiento específico–) y de los documentos (las otras tipologías de documentos que precisan otro tipo de tratamiento). Otros ejemplos son las webs, las publicaciones en redes sociales como Twitter o Facebook y los correos electrónicos.

Los datos **semiestructurados** son datos elementales, pero no tienen una estructura fija a pesar de que tienen algún tipo de estructura implícita o autodefinida. Los datos están encapsulados en ficheros semiestructurados. Es el caso de documentos escritos con lenguaje HTML, XML o SGML. Este tipo de datos presenta complejidad en el proceso de carga, porque se tienen que añadir tantas excepciones como posibilidades de variación existan. Por ejemplo, en el caso de un documento en XML, el orden de los datos no es importante, por lo tanto, cuando la base de datos cargue un documento, primero lo tendrá que interpretar y, si alguno de los campos no está estructurado, se produce pérdida de información.

c) Velocidad

El procesamiento de los datos masivos se debe hacer en el mínimo tiempo posible, y en

algunas ocasiones en tiempo real. Es un tema especialmente delicado cuando se deben tomar decisiones en contextos críticos; como, por ejemplo, en catástrofes naturales o pandemias, o cuando se efectúa el seguimiento de una campaña analizando comentarios de los actores a quienes esta va dirigida, para irla modificando a medida que se van analizando los comentarios que fluyen por las redes sociales.

La velocidad de gestión y análisis de los datos permite también ver con suficiente antelación hechos y acciones ante los cuales reaccionar con tiempo: clientes que se dan de baja, mal funcionamiento de un servicio o de un producto, abandono de apoyos, etc.

Esta característica está vinculada a las herramientas que deben gestionar los datos masivos, y se refiere a la velocidad de carga y la velocidad de procesamiento. Es decir, a qué velocidad se producirán, se procesarán y se podrá acceder a los datos, para satisfacer la necesidad expresada por el usuario. La velocidad también está vinculada a la visualización de los datos para facilitar el análisis y la extracción de patrones.

d) Veracidad

Dado que los datos masivos están vinculados a la extracción de información para tomar decisiones y crear conocimiento para la acción, es importante que sean fiables y que permanezcan siempre fieles a la realidad. La veracidad tiene que ver con aspectos como la certeza o incertidumbre y la exactitud o inexactitud de los datos.

Las personas que utilizan los datos para la toma de decisiones o el diseño de acciones necesitan confiar en los datos, en su exactitud y certeza, y estar seguros de que no hay datos erróneos.

Para asegurar la veracidad, es preciso aplicar una serie de medidas como proteger los datos de ataques malignos, y actualizar los datos, teniendo en cuenta su variabilidad. Estos aspectos están vinculados con el tratamiento de los datos, más concretamente con la limpieza de estos, que veremos más adelante, cuando tratemos la gestión de grandes volúmenes de datos.

e) Valor

Valor es otra de las características para definir los datos masivos, puesto que el análisis de los datos y la extracción de información crean conocimiento, que es fuente de innovación, competitividad, productividad y ocupación. Por eso, la gestión y el análisis de los datos masivos deben estar orientados a la creación de valor, como medio para justificar los esfuerzos invertidos.

f) Visualización

Esta característica tiene que ver con las prestaciones del software que gestiona los datos. Este debe facilitar que los grandes volúmenes de datos se presenten visualmente de manera práctica, dinámica, interactiva y comprensible. Además, a fin de extraer información de forma fácil, es relevante que la visualización vaya acompañada de contexto para facilitar el análisis. Por lo tanto, las plataformas que gestionen datos masivos también deberán tener en cuenta cómo presentan los datos.

El campo de investigación que explora soluciones de visualización se denomina *visual analytics* (VA) y ofrece soluciones para convertir el exceso de información en una oportunidad.

g) Verificación

Esta característica asociada a los datos masivos tiene relación con la seguridad de los datos y su integridad, especialmente cuando los datos tienen procedencia externa, como es el caso de los procedentes de la internet de las cosas (*internet of things* o IoT), o las que se almacenan en servidores en la nube. Por lo tanto, la verificación incluye las vías para asegurar la integridad de los datos mediante mensajes de autenticación, firmas digitales, certificados de terceros y otros sistemas (Liu y otros, 2015).

h) Variabilidad

Esta característica está relacionada con la velocidad con la que se producen los datos, ya que la generación de unos datos puede ocasionar la obsolescencia de otros. En consecuencia, esta característica de los datos masivos nos lleva a introducir procedimientos específicos para su gestión en términos de eficiencia, al mismo nivel que la veracidad.

i) Viabilidad

Las plataformas y el software para gestionar datos tienen capacidades y complejidades diferentes, y, por lo tanto, los presupuestos resultantes para comprar las infraestructuras tecnológicas necesarias pueden llegar a ser muy elevados. En consecuencia, cualquier proyecto que se defina para gestionar datos masivos en términos de eficiencia debe tener en cuenta su adaptación a las necesidades organizativas y las características de los datos. En segundo lugar, se debe observar el coste en cuanto a infraestructuras y herramientas necesarias para almacenar y analizar los datos masivos, garantizando su velocidad, veracidad, verificación y variabilidad. En cualquier caso, la característica de viabilidad nos recuerda que la inversión debe estar justificada por el valor que se extraiga de los datos, y que estos deben estar orientados al logro de los objetivos que la organización tenga fijados.

Tecnologías, técnicas y metodologías para procesar grandes volúmenes de datos

El término *datos masivos* también se ha empleado para referirse a las tecnologías y arquitecturas que gestionan grandes y variados volúmenes de datos, en términos de velocidad, captura, descubrimiento y análisis. Esto se debe a que los datos masivos promueven el crecimiento de infraestructuras y software, sin los cuales su gestión sería muy difícil, lenta y costosa, si no inviable o imposible.

Son diversas las definiciones que nos apuntan esta idea. Por ejemplo, Padgavankar y Gupta (2014) definen datos masivos como la gestión de nuevas bases de datos y diferentes enfoques analíticos para gestionar datos complejos y de gran volumen. Los mismos autores mencionan la necesidad de invertir en recursos humanos y en soluciones tecnológicas como plataformas de

gestión de bases de datos. Estos aspectos vinculados a la gestión de los datos masivos (nuevas tecnologías, y metodologías, perfiles profesionales y técnicas de análisis) han sido tomados como base para el desarrollo económico de las regiones (European Union, 2014) y la creación de nuevas áreas de ocupación (SAS, 2013).

En esta línea, la consultora y suministradora de tecnología en investigación de mercados IDC (International Data Corporation) define las tecnologías de datos masivos como una **nueva generación de tecnologías y arquitecturas** formuladas para extraer valor económico de grandes y variados volúmenes de datos, que facilitan capturar, descubrir y analizar datos a gran velocidad.

Capítulo II

¿CÓMO GESTIONAR EFICIENTEMENTE?

Los datos masivos y los conjuntos de datos tienen como características descriptivas, además de ser muy abundantes, su heterogeneidad, complejidad, desestructuración, falta de completitud, y su potencial de ser erróneos. Por lo tanto, cuando se diseñan procesos para gestionar datos, se deben tener en cuenta todos estos aspectos a fin de garantizar y preservar su calidad y la extracción de información útil y no errónea, que garantice la fiabilidad de los datos y, por lo tanto, del análisis resultante. Recordemos que la toma de decisiones se basa en los resultados del análisis, y de ser estos erróneos, la toma de decisiones puede tener resultados negativos. Además, la gestión debe tener en cuenta otros aspectos como privacidad, seguridad y protección.

Este capítulo presenta los aspectos a tener en cuenta cuando se gestionan los datos, pensando en la eficiencia de los procesos y la garantía de la calidad. A pesar de que en este capítulo se menciona alguna herramienta y método, en otro capítulo de este libro se presenta con más detalle la arquitectura tecnológica, las plataformas, herramientas y tecnologías que apoyan los diferentes ámbitos de la gestión de macrodatos.

La gestión de grandes volúmenes de datos se tiene que ver como un engranaje productivo, conformado por la entrada de una materia prima (los datos), cuya explotación y transformación producen unos servicios y/o productos de información y conocimiento. Por ello, este capítulo en primer lugar describe la cadena de valor del proceso productivo, con sus fases y subfases. En segundo lugar, orienta sobre cómo se puede diseñar, planificar e implantar un proyecto de gestión de datos masivos en una organización. En tercer lugar, aborda la calidad de los datos y cómo garantizarla. Por último, trata las relaciones de la cadena de valor de los datos con los procesos de gestión del conocimiento e inteligencia competitiva, que las organizaciones pueden tener implementados para generar conocimiento para la acción y ventaja competitiva.

1. Cadena de valor

Los datos son la materia prima que se debe procesar y transformar para generar valor, mediante una serie de procesos. Por eso los procesos de gestión de datos se presentan visualmente en una cadena de valor. Estos procesos se pueden englobar en tres: **explotación**, **almacenamiento** y **producción** (Chen y otros, 2014). El proceso de almacenamiento se debe acompañar de procesos propios de la gestión de la información como son la evaluación y el mantenimiento para garantizar la calidad de los datos almacenados para su recuperación. En la misma línea, Padgavankar y Gupta (2014) incluyen también el tratamiento de información

(*curation*).En cuanto al proceso de producción, este debe producir información y conocimiento para la acción, base de la creación de valor para la organización.

Esta cadena de producción se ha concretado en la cadena de valor de los datos masivos, que se compone de cuatro fases: generación, adquisición, almacenamiento y análisis de datos (Chen y otros, 2014; Hu y otros, 2014) (Ilustración 1).

Ilustración 1: Cadena de valor de los datos masivos (fuente: elaboración propia)

Por lo tanto, cuando hablamos de gestión de datos debemos pensar en los procesos de la gestión de la información, pero teniendo en cuenta los retos que plantean los diferentes formatos, tipos y origen de los datos masivos. Estos retos son diferentes de los que la gestión de información hasta ahora había planteado, y muchos de los cuales todavía necesitan investigación para resolverlos. Padgavankar y Gupta (2014) añaden la idea de compartir y transferir datos como una fase más de la cadena de valor, previa al análisis. Esto nos recuerda que los procesos propios de la gestión de conocimiento pueden contribuir a crear valor con los datos, sobre todo si están integrados en la fase de análisis para difundir la información y el conocimiento resultante.

Generación de datos

La cadena de valor de los datos masivos empieza con la generación de datos procedentes de una gran variedad de fuentes. Actualmente hay diferentes entornos donde se generan grandes cantidades de datos, y en muchos casos de manera automática: rastros que dejamos de las búsquedas que hacemos por internet, datos que generan los sensores y cámaras de vigilancia que se instalan en los edificios o naves industriales, sistemas de información existentes en las organizaciones, datos e información que generan las administraciones públicas o la investigación, internet de las cosas, etc. (Ilustración 2).

Ilustración 2: Fuentes y entornos de producción de datos en bruto (fuente: elaboración propia)

Internet de las cosas se basa en la capacidad que tienen los objetos cotidianos para conectarse a la red, ya sea para enviar información sobre su funcionamiento o sobre su entorno (mediante sensores integrados) o para recibir datos de otros dispositivos. La aplicación de esta filosofía aumentará de manera significativa la información del mundo que nos rodea. Esto permite digitalizar y distribuir información hasta ahora desconocida, cuyo análisis puede dar lugar a correlaciones, hasta ahora insospechadas (Gómez y Conesa, 2015). Estos datos se generan en muchos sectores y su explotación es fuente de oportunidades de negocio y minería de puestos de trabajo.

Obtención de datos

El segundo eslabón de la cadena es la obtención de datos de entre todos los existentes,

generados por diferentes tipos de objetos. La obtención se hace mediante varios métodos, técnicas y herramientas. En otro capítulo se describirán las herramientas y tecnologías que apoyan esta fase. Este capítulo se centra solo en los aspectos que se tienen en cuenta para gestionar datos.

Para obtener los datos, en primer lugar se deben definir los objetivos que se quieren cubrir e identificar las fuentes donde se generan los datos que se necesitan. Una vez realizado este paso previo, se deben evaluar los siguientes aspectos: a) la heterogeneidad de los datos, los cuales deben conservar su propia estructura, jerarquía y diversidad; b) la redundancia de datos, que multiplica la ocupación de espacio en los dispositivos de almacenamiento, con los costes asociados, y c) parámetros de calidad como la procedencia y validez de los datos, de los que algunos conjuntos de datos pueden carecer.

Debido a la complejidad del proceso de adquirir datos, en esta parte de la cadena de valor de los datos trabajamos con tres subprocesos: recogida, transmisión y preprocesamiento de datos (Chen y otros, 2014; Hu y otros, 2014). No hay un orden estricto entre los dos últimos procesos (transmisión y preprocesamiento) puesto que las dos operaciones pueden ocurrir indistintamente una antes que otra (Hu y otros, 2014) (Ilustración 3).

Ilustración 3: Subfases y procesos de la adquisición de datos (fuente: elaboración propia)

Cada una de estas subfases tiene su importancia y tienen que estar muy bien diseñadas para garantizar la calidad de los datos y la fiabilidad de los resultados del análisis efectuado en fases posteriores. A continuación se detallan estas subfases.

a) Recogida

En cuanto a la recogida de datos, esta subfase debe tener en cuenta la variedad de los datos atendiendo a su procedencia. Hay diferentes métodos de recogida, que se pueden agrupar en dos categorías: a) los que están basados en un enfoque *pull*, es decir, de extracción de información de manera proactiva por medio de un agente (por ejemplo, un robot que rastrea la web, denominado *crawler*), y b) los que están basados en enfoques push, es decir métodos que obtienen los datos mediante una distribución selectiva en una fuente concreta (Hu y otros, 2014). Tres métodos habituales de recogida son: el uso de sensores, ficheros log y web crawler (Chen y otros, 2014; Hu y otros, 2014).

- Los *ficheros log* son generados automáticamente por las aplicaciones y los aparatos digitales y graban datos de las actividades que se hacen en ellos. Se almacenan en unos formatos concretos para su posterior análisis. Por ejemplo, los servidores web almacenan el número de clics, las visitas y otras características de los usuarios.
- Los *sensores* recogen mediciones de temperatura, presión, vibraciones, caudal de un río, etc.
- La extracción de datos disponibles en red, como las páginas web, se hace mediante diferentes tecnologías, como el rastreo de páginas web, sistemas de segmentación de palabras y sistemas de indexación, entre otros métodos. Los *web crawler* descargan y almacenan las páginas web.

b) Transmisión

Los datos obtenidos se transfieren normalmente a un centro de datos para su posterior tratamiento y análisis. Este centro de datos está asociado a una arquitectura en red y un protocolo de transmisión. Las tecnologías que apoyan la arquitectura en red de este centro todavía están en construcción, y constituyen una interesante línea de investigación. Asimismo, los protocolos de red más importantes para la transmisión de datos no son aún del todo satisfactorios, y se está trabajando para mejorar el rendimiento de estos protocolos.

c) Preprocesamiento

En esta subfase se tratan los datos recogidos en bruto, puesto que al proceder de diferentes fuentes, presentan formatos diversos, duplicidades o en algunos casos errores y carencias, como hemos apuntado antes. Se debe tener en cuenta que el tratamiento de los datos se hace para asegurar su calidad y fiabilidad. A estos datos, en la fase de la cadena posterior se les aplicará el modelo de análisis seleccionado para generar conocimiento para la acción. Se deben aplicar las siguientes tareas: a) integración de datos procedentes de diferentes fuentes para reducir los gastos de almacenamiento, y para dar a los usuarios una visión uniforme de los datos; y b) evaluación de la consistencia de los datos, lo que conlleva limpiarlos para reducir el ruido y la redundancia y mejorar la precisión del análisis.

La investigación en el preprocesamiento de datos se focaliza en tres técnicas: a) integración, b) limpieza (*cleansing*) y c) eliminación de la redundancia.

- ***Integración***

La integración combina datos de diferentes fuentes y ofrece al usuario una visión unificada de estos (Hu y otros, 2014). La integración de datos se hace mediante herramientas de procesamiento de flujos y búsqueda (Chen y otros, 2014). Estas son: a) **ETL** integrado en un ***datawarehouse*** y b) **federación de datos**, a pesar de que están limitados atendiendo las elevadas necesidades de rendimiento de las aplicaciones de búsqueda o transmisión en tiempo real (*streaming*). Los métodos de integración de datos están mejor vinculados con herramientas de procesamiento de transmisión en tiempo real y de búsqueda.

El primer método, ETL integrado en *data warehouse*, incluye un proceso de extracción, transformación y carga, denominado ETL por sus siglas en inglés (*extract*, *transform* y *load*) (Chen y otros, 2014; Hu y otros, 2014). El método ETL consiste en los siguientes procesos:

- Para la **extracción** se conectan sistemas de fuentes, se seleccionan mediante los criterios y objetivos previamente establecidos, se recogen, se analizan y se procesan los datos necesarios.
- La **transformación** es la ejecución de una serie de reglas para que los datos extraídos se transformen en formatos normalizados que puedan ser tratados en las plataformas

tecnológicas.
- La **carga** es el procedimiento más complejo de los tres, y consiste en importar los datos extraídos y transformados a la infraestructura de almacenamiento seleccionada.

El método de **federación de datos** crea una base de datos virtual que, en lugar de datos, contiene información o metadatos sobre los datos actuales y su localización. Esta base de datos virtual puede ser interrogada y se le pueden agregar datos de fuentes diversas. Por lo tanto, la creación de metadatos es un elemento importante cuando se trabaja con grandes volúmenes de datos.

- ***Limpieza* (*cleansing*)**

La limpieza es un proceso clave para lograr la calidad de los datos. A tal fin tiene en cuenta la consistencia de los datos, su actualización, identifica datos imprecisos, incompletos o irracionales, y los modifica o borra para mejorar su calidad. Datos incompletos pueden responder a un valor indefinido o inexistente, porque cuando se ha definido el proceso de introducción de datos se ha decidido que determinados atributos no son obligatorios o tienen un formato libre.

Este proceso de limpieza incluye cinco procedimientos: 1) definir y determinar los tipos de errores, 2) buscarlos e identificarlos, 3) corregirlos, 4) documentar los ejemplos de error y los tipos de error, y 5) modificar los procedimientos de introducción de datos para reducir errores futuros (Chen y otros, 2014; Hu y otros, 2014). Estos procedimientos están acompañados de pautas de revisión, y tienen en cuenta los siguientes aspectos: formato, integridad, racionalidad y límite.

- ***Eliminación de redundancia***

La redundancia se refiere a la repetición de datos comunes en diferentes conjuntos de datos, de forma que puede incrementar gastos innecesarios de transmisión de datos y causar defectos en los sistemas de almacenamiento (desaprovechamiento del espacio de almacenamiento, inconsistencia de los datos, reducción de su veracidad y daño de estos).

Hay varios métodos y técnicas para reducir la redundancia, y es una línea de investigación actual. Entre estos métodos se pueden citar: detección de la redundancia, filtraje de datos y compresión de datos. Estos métodos pueden aplicarse a diferentes conjuntos de datos o entornos de aplicación. La selección de uno u otro método viene dada por varios aspectos: las propias características del conjunto de datos, el problema que debe resolverse, los requisitos de rendimiento y otros factores para escoger un esquema adecuado de preprocesamiento de datos (Hu y otros, 2014).

Almacenamiento de datos

Una vez hemos obtenido los datos, estos deben estar almacenados en una plataforma, en un formato adecuado para asegurar una rápida recuperación y posterior análisis y extracción de

valor. El sistema de almacenamiento debe asegurar la consistencia de los datos, en el caso de que haya varias copias de los mismos datos (en previsión de caídas de servidores) y reproducir los datos en varios servidores.

En el almacenamiento juega un papel importante definir cómo guardar los datos de cara a la eficiencia en la recuperación de los conjuntos de datos, cómo plantear las búsquedas y consultas de los datos almacenados, y cómo presentarlas a los usuarios. En este sentido, se deben explorar líneas de investigación en varios ámbitos: a) **lenguajes controlados**, para trabajar la precisión de la clasificación e indexación de todos los conjuntos de datos; b) **usabilidad** de los almacenes de datos, y c) **visualización** de los datos (Ilustración 4).

Ilustración 4: Almacenamiento de datos (fuente: elaboración propia)

Para esta fase se deben definir procedimientos claros para todas las tareas identificadas, a fin de que estas estén orientadas a la calidad de los datos. Algunos ejemplos de las tareas identificadas para almacenar los datos obtenidos y preprocesados son: recopilar y analizar metadatos, actualizar los metadatos, controlar y verificar la transformación de los datos, para ver si los datos se han cargado de la manera deseada en términos de calidad, detectar incorrecciones y encontrar el origen para resolver la causa, actualizar las fuentes de datos, revisar periódicamente las necesidades de información, y asegurar que los datos obtenidos satisfacen los requisitos de información de la organización, entre otros.

El almacenamiento se compone de infraestructura tecnológica (dispositivos y arquitectura de red) y de marco de gestión de datos para organizar la información de manera adecuada en aras de un procesamiento eficiente. La tecnología asociada a esta fase son los sistemas de almacenamiento masivo, los sistemas de almacenamiento distribuido y los mecanismos de almacenamiento (Chen y otros, 2014).

Tradicionalmente el software y el hardware para almacenar y gestionar datos y analizarlos están vinculados a los sistemas fundamentados en bases de datos relacionales. Este tipo de bases de datos surge en la década de los ochenta del siglo pasado y, aunque han ido evolucionando y aumentando sus prestaciones, presentan importantes limitaciones para gestionar grandes volúmenes de datos. Entre estas limitaciones destacan que solo gestionan datos estructurados, y que son sistemas basados en el lenguaje SQL (*structured query language*). Una parte de sus limitaciones quiso ser resuelta con un nuevo tipo de bases de datos. Por lo tanto, las bases de datos relacionales evolucionaron, y desde la década de los noventa las organizaciones tienen a su disposición los llamados almacenes de datos (*data warehouse*) o repositorios de datos.

Algunas de las grandes diferencias que presentan los *data warehouse* frente a las bases de datos relacionales son su orientación y la explotación de los datos. En concreto, los almacenes de datos están orientados a temas de interés y no a funcionalidades. Para ello, los *data warehouse* integran datos procedentes de varios sistemas informáticos para dar una visión global, común e integrada de los datos de la organización. De este modo los analistas ven los datos como si vinieran de una única fuente. Además, ofrecen información histórica y no volátil (solo de lectura para los usuarios

finales) para ver la evolución en temas de interés para la organización (Rius y otros, 2013).

Los almacenes de datos, tal como se concibieron en sus inicios, han ido evolucionando, para poder asimilar el aumento del volumen de datos, y hacer frente a los retos de los datos masivos, sobre todo los que no son estructurados. Aun así, actualmente los almacenes de datos presentan limitaciones, no solo por el volumen sino también en cuanto a los formatos desestructurados. Por ello el almacenamiento de datos masivos es una de las líneas de investigación actuales para dar respuesta a uno de los retos que los datos masivos presentan. Un ejemplo es la sustitución del lenguaje SQL por el lenguaje NoSQL, que se está convirtiendo en el núcleo central de las tecnologías para gestionar datos masivos. Actualmente ya están disponibles algunos sistemas de almacenamiento para grandes volúmenes de datos.

Análisis de datos

La cadena de valor de los datos masivos tiene como fase final el análisis de los datos. Este es clave para extraer el valor que los datos encierran con el fin de crear conocimiento para la acción. Desde siempre se ha dado valor a los datos y a la información para la toma de decisiones. Las organizaciones siempre han dispuesto de datos, aunque estuvieran plasmados en formato papel (como, por ejemplo, en libros de contabilidad). La disciplina de la inteligencia competitiva es un ejemplo de los procesos conducentes a extraer información estratégica y conocimiento para la acción, mediante técnicas de análisis. Los adelantos en la tecnología han permitido que el análisis cuente con el apoyo de herramientas cada vez más potentes. Los almacenes de datos (*data warehouse*) tienen como uno de sus objetivos facilitar el análisis de los datos. El análisis engloba un conjunto de procedimientos y modelos estadísticos para extraer información de un amplio conjunto de datos (Kune y otros, 2016). En este sentido hay una serie de métodos desarrollados, como son minería de datos, factores, correlaciones, regresiones test A/B, estadística, etc. (Chen y otros, 2014) (Ilustración 5).

Ilustración 5: Análisis de datos (Fuente: elaboración propia)

En la gestión de datos masivos, atendidas las altas inversiones en hardware y software, se da una relevancia especial al análisis de datos, sobre todo aquellos que no están estructurados. En efecto, Chen y otros (2014) indican que desde los inicios de la existencia de datos masivos ha habido un elevado interés para extraer información de los datos que aporte valor a las organizaciones. Aunque se da importancia a este tema, todavía hay camino por recorrer. El análisis de datos está siendo objeto de investigación, dado que las plataformas y las técnicas tradicionales presentan limitaciones para extraer información y gestionar la gran cantidad de datos (Tsai y otros, 2015). Encontramos tres grandes ámbitos de investigación: 1) Diseño de **tecnologías y software** que faciliten el análisis de datos según sus características: análisis de datos estructurados, análisis de texto, análisis de datos en web, análisis de datos multimedia, análisis de datos de redes y análisis de datos de móviles (Hu y otros, 2014); 2) Diseño de métodos

de **análisis** según el formato y la estructura de los datos, y 3) **Visualización** de la información en forma de gráficos y sobre todo la información resultante de los datos masivos para ayudar en el diseño de algoritmos y desarrollo de software (Huy otros, 2014). A continuación se describen algunos métodos de análisis.

a) Métodos de análisis

Chen y otros (2014) clasifican los métodos de análisis en tradicionales y específicos de los datos masivos. Así, el análisis de datos se puede hacer con minería de datos, análisis estadístico y visualización de datos, entre otros métodos. Aun así, la selección del método depende también del tipo de datos que se van a analizar (estructurados, desestructurados, web, etc.). Orientados a los nuevos formatos de datos, han surgido otros métodos de análisis, entre los que destacan la minería web (*web mining*), la minería de textos (*text mining*) y el análisis de redes sociales (*social networks analysis*). En consecuencia, actualmente hay nueva investigación en este ámbito. En este apartado se mencionan algunas de estas líneas (tabla 1).

Tabla 1: Métodos de extracción y análisis de datos (fuente: elaboración propia)

TIPO DE DATOS	EXTRACCIÓN	MÉTODOS
Estructurados	Detección de anomalías	Algoritmos
	Descubrimiento de estructuras mediante la explotación de características, tiempos y espacio	Minería de datos
Desestructurados: datos en texto	Datos de texto	Sistemas de minería de texto basados en: expresiones, procesamiento del lenguaje natural (modelos de temas, resumen de texto, clasificación, agrupación, minería de opinión, etc.)
Datos de web	Datos de texto, multimedia, vídeo, foros...	Minería de contenido de la web: análisis multimedia (texto, imágenes, audio, vídeo), minería de hipertexto
	Estructuras de los enlaces dentro de una web o entre varias webs	Minería de estructura web
	Análisis de *logs* almacenados en los servidores web y proxis, registros de los históricos de navegación, perfiles de usuarios, datos de registro, sesiones de usuario, preguntas de los usuarios, datos de *bookmark*, clics de ratón y barras de desplazamiento y cualesquiera otros datos generados en la interacción con la web.	Minería de uso de web

Datos multimedia	Videos, música, imágenes...	Resumen, anotación, indexación y recuperación, detección de acontecimientos, etc.
Redes sociales en línea	Datos masivos enlazables Datos de contenidos	Análisis de redes sociales Análisis de la estructura basada en enlaces Análisis basado en contenidos

- ***Análisis estadístico***

El análisis estadístico está basado en la teoría de la probabilidad que tiene en consideración la aleatoriedad y la incertidumbre. La estadística descriptiva y la estadística inferencial ofrecen los métodos para analizar los datos. La estadística descriptiva resume y caracteriza el conjunto de datos que queremos analizar. La inferencia permite extraer conclusiones de los datos sujetos a variaciones aleatorias. Las inferencias ayudan a generar modelos, predicciones, dar respuestas a preguntas (basadas en hipótesis), hacer estimaciones, pronósticos de futuras observaciones o correlaciones (descripciones de asociaciones). Para ello la estadística cuenta con varias técnicas.

Según el número de variables escogidas para analizar, se pueden hacer tres tipos de análisis, cada uno de los cuales cuenta con técnicas específicas:

a) El análisis univariante se utiliza para ver la distribución y dispersión o variabilidad interna de los datos. Las técnicas empleadas para calcular la distribución son el cálculo de frecuencias, la media, la moda y la mediana. Para calcular la dispersión se utiliza la desviación típica o la varianza.

b) El análisis bivariante se utiliza para estudiar el efecto de una variable sobre otra. Las técnicas empleadas son la comparación de medias, análisis de correlaciones, análisis de varianza y tablas de contingencia.

c) El análisis multivariante se emplea para analizar más de dos variables. Este análisis es más complejo que los anteriores y cuenta con un abanico de técnicas estadísticas y de algoritmos de cálculo más amplio. Las técnicas son: análisis de la varianza (para ver el efecto de dos factores sobre una variable técnica), análisis multivariante de la varianza, análisis discriminante, análisis de regresión lineal múltiple, análisis de regresión logística, análisis de covarianza, modelo lineal general, análisis factorial, análisis de conglomerados (o *clusters*).

- ***Minería de datos***

La minería de datos engloba un conjunto de metodologías, procesos de modelización y técnicas matemáticas para analizar datos provenientes de diferentes fuentes con el objetivo de extraer información previamente desconocida. Se aplica a conjuntos de datos estructurados por atributos o valores. No es válida para documentos de texto; para estos se debe emplear minería de texto.

La información se construye analizando las estructuras de los datos, de las cuales emergen

patrones de comportamiento y tendencias. Los patrones están basados en la observación del pasado, y las técnicas predictivas nos dan información de tendencias futuras. La información extraída apoya la toma de decisiones y tiene que estar alineada con los objetivos organizativos.

El proceso de minería de datos empieza en la fase de la cadena vista antes (obtención de datos), dado que se debe seleccionar el conjunto de datos adecuado a aquello que queremos analizar. Estos datos se tienen que evaluar para ver sus propiedades, frecuencias, dispersión, valores atípicos y la ausencia de datos. Si hace falta, estos datos se deben transformar, como hemos visto antes, en la fase previa de la cadena de valor. Cuando ya tenemos los datos evaluados, se tiene que aplicar la técnica de minería de datos, para lo cual se construye un modelo (predictivo, de clasificación o de segmentación).

Las técnicas de minería de datos están basadas en inteligencia artificial y estadística. Estas disciplinas facilitan la creación de algoritmos que permiten modelizar los datos. Los algoritmos pueden basarse en **clasificación supervisada y predictiva** o clasificación **no supervisada y descriptiva** (o de descubrimiento de conocimiento). Sus características son (Gironés, 2013a):

a) Los **algoritmos supervisados** tienen como objetivo la obtención de un modelo válido para predecir casos futuros a partir del aprendizaje de casos conocidos. A partir de un conjunto de objetos descritos por un vector de características y del que conocemos la clase a la que pertenece cada objeto, se construye un grupo de datos denominado de entrenamiento o de aprendizaje. Por lo tanto, parten de conocimiento existente.

b) Los **algoritmos no supervisados** tienen como objetivo obtener un modelo válido para clasificar objetos sobre la base de la similitud de sus características, pero sin partir de modelos predictivos. Se basan en un conjunto de objetos descritos por un conjunto de características, y a partir de una métrica que define la similitud entre objetos, se construye un modelo o regla general que clasificará todos los objetos. Por lo tanto, se descubre conocimiento.

Los algoritmos más representativos son:

a) **Redes neuronales**, para ver conexiones en una red, son una buena aproximación a problemas en los que el conocimiento es impreciso o variante en el tiempo. Se basan en clasificación supervisada.

b) **Regresión lineal**, para formar relaciones entre datos, a pesar de que es insuficiente en espacios multidimensionales donde intervienen más de dos variables. Se basan en clasificación supervisada.

c) **Árboles de decisión**, para hacer modelos de predicción, representan y categorizan una serie de condiciones, cuya visualización tiene forma de árbol y facilita la comprensión del modelo. Se basan en clasificación supervisada.

d) **Agrupamiento**, *clustering*, para ver agrupaciones de datos según criterios de distancia. El agrupamiento se hace sobre la base de jerarquías y partiendo de una fragmentación completa de los datos, estos se van agrupando. Esta técnica analiza datos que no tienen ninguna etiqueta o información añadida, por lo tanto, se tienen que descubrir grupos similares en los grupos de datos. Los datos que quedan más cercanos son los que tienen características comunes. Se basan en clasificación no supervisada.

e) **Segmentación**, para dividir grupos previamente existentes. Se basan en clasificación no supervisada.

f) **Reglas de asociación**, para encontrar relaciones entre combinaciones de valores en un conjunto de datos. Se basan en clasificación no supervisada.

- ***Minería web***

La minería web es una disciplina específica que desarrolla técnicas para extraer información y conocimiento de los enlaces, los contenidos de las páginas web y los *logs* de uso de los recursos de internet (Gironés, 2013a).

A pesar de que la minería web utiliza muchas técnicas de la minería de datos, las dos minerías presentan diferencias. La principal está en los procesos de captura de la información. La minería de datos tradicional utiliza los datos almacenados en un repositorio. En el caso de la minería web el proceso de captura de datos es una de las tareas más importantes y relevantes, al tener que rastrear sitios web. La heterogeneidad de la web, las estructuras de los enlaces y los datos no estructurados han hecho que la minería web desarrolle sus propias técnicas (Gironés, 2013a). Estas técnicas son:

- **Minería de la estructura web** (*web structure mining*). Esta técnica extrae información de la estructura web. Mediante el análisis de los enlaces puede estimar la relevancia de las páginas web, y también se pueden identificar comunidades de usuarios que comparten ámbitos de interés.

- **Minería del contenido web** (*web content mining*). Extrae patrones analizando el contenido de las páginas web. Por ejemplo, clasifica de forma automatizada páginas web en función de su contenido o extrae opiniones de los comentarios de los usuarios y descripciones de los productos que contiene.

- **Minería del uso de la web** (*web usage mining*). Tiene como objetivo extraer patrones de uso de los recursos de la red a partir de los logs que registran la actividad del usuario.

- ***Minería de texto***

La minería de texto es un conjunto de técnicas para identificar y extraer conocimiento de un corpus textual, que contiene datos no estructurados. Por eso difiere de la minería de datos, a pesar de que también se basa en algoritmos. La minería de texto es una aplicación de la lingüística computacional y del procesamiento de textos. Se fundamenta en los campos de investigación de la recuperación de la información, el aprendizaje automático y la lingüística computacional.

El potencial de la minería de texto está en la clasificación y la recuperación de los datos y la información contenida en los documentos. Más específicamente, la minería de texto: a) identifica hechos y datos puntuales –nombres de personas, organizaciones o acontecimientos– a partir del texto de los documentos (*feature extraction*); b) agrupa documentos similares (*clustering*), a partir de las similitudes que se establecen entre la terminología utilizada por los autores; c) determina el tema o los temas de los documentos mediante la categorización automática, por la que se asigna a

un documento una clase o un tema definidos con anterioridad; d) identifica conceptos tratados en los documentos y crea redes de conceptos, e) facilita el acceso a la información repartida entre los documentos del corpus, mediante la elaboración automática de resúmenes y la visualización de las relaciones entre los conceptos tratados en el corpus documental, y f) facilita la visualización y navegación de las colecciones de texto mediante una interfaz de usuario que muestre los datos en un formato que posibilite su interpretación y moverse con facilidad entre los diferentes textos analizados.

En cuanto a los procesos que se tienen que llevar a cabo, en la minería de texto también hay una fase de recuperación, preparación y valoración de la información, para eliminar el ruido existente en el texto. Es decir, las palabras superfluas, prescindibles y engañosas se deben eliminar. Esto quiere decir que se tendrán que eliminar las palabras y signos que pertenecen a las siguientes categorías (Gironés, 2013a):

- Conjunciones y preposiciones (*stopwords*), porque no tienen sentido propio, solo tienen la función de conectar palabras y frases. A tal fin se construye un diccionario propio de cada idioma, que es el que se carga en el software y sirve para eliminarlas del corpus.

- Palabras derivadas (*stemming* o lematización), referidas a plurales, conjugaciones de verbos, sufijos, prefijos, etc. Para ello se identificará la raíz de la palabra y se considerará la que mantiene mayor contenido.

- Signos de puntuación, mayúsculas y números excepto en casos especiales, como cuando se trate de fechas y horas.

- Objetos específicos de páginas web, en el caso de que el texto que se quiere analizar esté almacenado en webs. Ejemplos son el código HTML, o *tags* propias de la categorización del texto, como <body>, <metadata>, etc.

La minería de texto dispone de modelos **de representación de documentos** que permiten las aplicaciones de técnicas numéricas sobre ellos. El modelo vectorial propuesto por Salton (1971) permite representar los documentos a partir de un vector de pesos asociados a un conjunto de características seleccionadas del documento (Cobo y otros, 2009). La ponderación de las características seleccionadas de cada documento se realiza con diferentes estrategias, siendo la más habitual el llamado *esquema tf.idf*, donde el peso de una característica se obtiene como producto de dos factores:

factor tf: mide la frecuencia de aparición de la característica en el documento,

factor idf es la frecuencia inversa del documento y permite rebajar significativamente el valor de los pesos correspondientes a características con poco valor discriminante para aparecer en muchos documentos de la colección.

Detrás de la minería de texto e incorporados a sus herramientas tecnológicas, están los glosarios, tesauros, taxonomías y ontologías. Estos instrumentos facilitan establecer relaciones semánticas entre los términos y técnicas para la extracción de conocimiento. Ejemplos de estos vocabularios controlados son el tesauro multilingüe Eurovoc de la Unión Europea, NACE

(Nomenclature statistique des Activités économiques dans la Communauté Européenne) o la Clasificación Internacional de Patentes (CIP)(Wartena y García, 2015).

Actualmente se continúa investigando para explotar más las posibilidades que el lenguaje natural ofrece para extraer información de textos.

- ***Opinion mining y sentiment analysis***

En los últimos años la minería de texto ha facilitado también la exploración de las connotaciones del texto, el estudio de la subjetividad que rodea los estados personales o privados, las opiniones sobre aspectos de un producto, por ejemplo, las evaluaciones, las emociones y las especulaciones. Se reconoce como un ejemplo de estudio la interpretación del lenguaje orientado a la opinión en contraposición a la interpretación del lenguaje objetivo.

Aun así, valorar la opinión personal de forma automatizada, de manera que se concluya si es una opinión positiva o negativa es todo un reto, y un ámbito en el que se está trabajando, a partir de técnicas de lenguaje natural (NLP, *natural language processing*) (Gironés, 2013b).

- ***Análisis de redes sociales o social network analysis (SNA)***

El análisis de las redes sociales va más allá de analizar sitios como Facebook o Twitter. Por análisis de redes sociales entendemos el estudio de las interacciones y relaciones entre personas y organizaciones, llamados actores. La estructura conforma visualmente una red, donde los vértices son los actores y las líneas de unión entre los vértices son las relaciones entre ellos.

Aspectos que se tienen en cuenta son la **centralidad**, que analiza cómo se relaciona un actor con su entorno; la **proximidad**, que mide la distancia entre un actor y el resto de los actores de su entorno, y la **intermediación**, que mide el grado de influencia de un actor en las relaciones entre sus próximos. El análisis de estos aspectos permite ver el prestigio de un actor frente a terceros.

Ejemplos de ello son analizar agrupaciones de industrias o empresas de un mismo sector, o potencialidades para crear redes de innovación en un territorio.

- ***Gestión de la reputación (reputation management)***

La combinación de la minería de texto, la minería de opinión (*opinion mining*) y el análisis de las redes sociales junto con el procesamiento del lenguaje natural permite recuperar un conjunto de fuentes de información (como artículos, blogs, páginas web o comunidades virtuales en línea) para poder explorar la visión que hay en internet de una determinada organización.

- ***Social media analytics***

Cada vez más, las empresas están incorporando datos de las redes sociales en las aplicaciones

de *business intelligence* para analizar aspectos de su negocio. Las analíticas tienen que ser proactivas y orientadas a acciones futuras, por lo cual el análisis se efectúa en tiempo real. Bartrolí (2015) indica algunas de las métricas e indicadores a tener en cuenta:

- **Visibilidad y exposición**: cantidad de tráfico hacía la web, cantidad de visitas y páginas visitadas, cantidad de seguidores y suscriptores.
- **Sentimiento y notoriedad**: número de conversaciones sobre una marca en comparación al número de conversaciones relacionadas con la competencia.
- **Influencia**: capacidad de un usuario para modificar el comportamiento o la toma de decisiones de sus seguidores. Las medidas empleadas son variadas, entre ellas señalamos: número y calidad de los enlaces entrantes a su contenido, enlaces de Twitter que son reenviados, comentarios y «Me gusta» de Facebook, contenido compartido, etc.
- ***Engagement*** **o vinculación emocional**: cómo interactúa la gente con una empresa y su contenido. Indicadores que se pueden utilizar son: nuevos «Me gusta», número de veces que se ha compartido alguna información, menciones, comentarios, etc.
- **Popularidad**: en lo referente al número de gente que se suscribe al contenido de una empresa, para lo cual se pueden utilizar como indicadores: número de suscriptores por RSS o por correo electrónico, número de seguidores enTwitter, número de miembros en el grupo de LinkedIn o número de «Me gusta» en la página de Facebook, no en un contenido en concreto.

Las herramientas de medición son variadas, y los propios medios sociales ya ofrecen herramientas de análisis, monitorización y presentación de informes.

b) Visualización

El análisis visual de los datos es especialmente relevante cuando se trata de un estudio numérico, dado que facilita la comprensión del contenido informativo resultante.

Hay aplicaciones que ofrecen acceso a los datos de forma interactiva, simulaciones de tipos «What if» con la posibilidad de guardar los diferentes estudios hechos en forma de versiones, facilidad de diseño de informes con posibilidad de combinar gráficos con resúmenes numéricos de frecuencias, disponibilidad de columnas con operadores estadísticos, gráficos con movimiento de variables, funciones de optimización de objetivos, etc. (Gironés, 2013b).

Diseñar e implementar un proyecto de gestión de datos masivos

Los proyectos de gestión de datos masivos tienen un nivel de dificultad elevado, dado que involucran diferentes perfiles profesionales y diferentes áreas organizativas. Tienen que tener en consideración a toda la organización, sus datos y sus aplicaciones, y por ello deben estar en diálogo con los usuarios.

Antes de diseñar e implementar un proyecto de gestión de datos masivos en una organización, hace falta en primer lugar hacer una autoevaluación para conocer cuál es la situación en la que se encuentra la organización: si hay silos de información, qué sistemas informáticos hay implementados, conocer cuál es el objetivo del proyecto, es decir, qué expectativas tiene la

organización, qué se quiere conocer con el análisis de los datos, cuantificar el valor que el conocimiento obtenido puede aportar a la organización y qué limitaciones tiene para recoger, procesar, almacenar y analizar la información.

Además, los proyectos de gestión de datos masivos son proyectos muy tecnológicos, dado que deben tener en cuenta un gran número de aspectos, como, por ejemplo, arquitecturas funcionales, requisitos técnicos y de negocio, análisis de datos, diseño y construcción de metadatos, procesos de ETL y cubos OLAP, minería de datos, gestores de bases de datos y capas de integración, herramientas de visualización y navegación, políticas de seguridad, o gestión de infraestructuras, entre otros. Son proyectos basados en prototipos, en los que se hacen pruebas y se aprende de los errores (Rodríguez, 2012).

Hay diversas metodologías para gestionar y desarrollar proyectos en las que están involucradas infraestructuras tecnológicas. Una metodología es una guía que seguir durante la gestión de un proyecto. Se diferencia entre: a) metodología de gestión de proyectos (qué hacer, pero no cómo), b) metodología de desarrollo de proyectos (cómo crear un producto o servicio) y c) metodología de gestión de servicios (cómo operar eficientemente un servicio ya existente) (Santanach, 2013).

Dos ejemplos de metodologías de gestión de proyectos son: PMBOK (*Project Management Book Of Knowledge*) y PRINCE 2 (*Projects In Controled Environments 2*). En este apartado damos unos apuntes sobre la gestión de proyectos, tomando como base la metodología PMBOK por ser la base de la norma ISO 21500, *Guidance on project management*, reconocida internacionalmente.

La **metodología PMBOK** incluye directrices para gestionar un proyecto, normas, métodos, procesos y prácticas establecidas y reconocidas internacionalmente, y es válida para cualquier tipo de proyecto. Está basada en dos ejes principales: la gestión de proyectos orientada a lograr los objetivos del proyecto y las áreas de conocimiento. Se identifican cinco grupos de procesos de gestión de proyectos y nueve áreas de conocimiento o disciplinas incluidas en los procesos.

Los procesos para gestionar un proyecto según la metodología PMBOK son:

1) **Iniciación**: incluye la definición del proyecto o una nueva fase de un proyecto existente, que en detalle es desarrollar el acto de constitución del proyecto, identificar a los interesados y definir su alcance inicial.

2) **Planificación**: incluye establecer el alcance del proyecto (requisitos y estructura de los paquetes de trabajo), refinar los objetivos y definir las acciones necesarias para lograr dichos objetivos. En esta fase se tiene que incluir la planificación temporal de las actividades, la estimación de esfuerzos, la secuencialidad y la duración del trabajo, la distribución del trabajo y recursos necesarios, y la preparación del calendario definitivo. Otros aspectos que se deben tener en cuenta en la planificación de proyectos son: costes del proyecto, calidad y riesgos (identificar, analizar y diseñar respuestas).

3) **Ejecución**: hace referencia al día a día del proyecto, desde su inicio a su finalización, e incluye todos los procesos realizados para completar el plan de trabajo definido por la dirección del proyecto. En esta fase, la dirección del proyecto tiene que asegurar la calidad, gestionar las comunicaciones entre los miembros del equipo y las partes involucradas, gestionar los recursos humanos y técnicos y administrar compras y contactos.

4) **Seguimiento y control**: incluye los procesos necesarios para observar el progreso y el desarrollo del proyecto, recoger elementos y avisos sobre su desarrollo, analizarlos e identificar desviaciones o áreas en las que el plan de trabajo requiera cambios, y en consecuencia iniciar los cambios necesarios o la replanificación. Aspectos que deben tenerse en cuenta en los controles son alcance, costes, tiempos, riesgos, calendario y calidad.

5) **Cierre**: incluye aquellos procesos realizados para finalizar formalmente un proyecto o una fase de este. Estos son: asegurar la aceptación de los productos y la transición del proyecto y los productos al funcionamiento ordinario ya en producción de los nuevos sistemas, cerrar toda la documentación administrativa y los contratos, y documentar las lecciones aprendidas para las personas y la organización.

Las áreas de conocimiento o disciplinas tenidas en cuenta en la metodología PMBOK son:

- Gestión de la integración del proyecto
- Gestión del alcance del proyecto
- Gestión del tiempo del proyecto
- Gestión de costes
- Gestión de la calidad
- Gestión de recursos humanos
- Gestión de la comunicación
- Gestión de riesgos
- Gestión de las adquisiciones

Calidad de los datos

En la cadena de valor es importante velar por la calidad de los datos, para que los datos analizados y los resultados del análisis sean consistentes para tomar decisiones muy fundamentadas.

La calidad de los datos está fundamentada en una serie de propiedades: precisión, completitud, relevancia, validez de los datos, proveniencia, autenticidad, veracidad, exactitud, reputación y credibilidad. Para realizar una gestión de la calidad sistemática, la Organización Internacional de Normalización (ISO - International Organization for Standardization) ha desarrollado una norma específica para orientar la gestión de la calidad de los datos. Se trata de la ISO 8000 y las relacionadas.

Esta norma tiene en cuenta una serie de aspectos y tres roles genéricos.

Los aspectos que se tienen en cuenta son: a) las personas que deben participar en la gestión de los datos; b) los procesos que se responsabilizan de la gestión efectiva de los datos; y c) la mejora continua de los procesos dedicados a garantizar la calidad de los datos.

Los roles que se identifican son: a) gestor de datos, responsable de factores organizativos para gestionar la calidad de datos; b) administrador de datos, responsable de coordinar y supervisar el

trabajo de los técnicos de datos, alineados con las directrices del gestor de datos; y c) técnico de datos, responsable de los cambios de datos que se llevan a cabo, de la corrección de datos y la medición de la calidad.

Esta iniciativa, junto con los adelantos hechos en la calidad de la información constituyen una buena base para seguir construyendo un marco de gestión de la calidad de los datos.

Gestión del conocimiento, inteligencia competitiva y gestión de datos masivos: gobernanza de la información

Las actividades operativas (diarias), tácticas (a corto plazo) o estratégicas (a largo plazo) necesitan y generan datos e información en el seno de las organizaciones. La observación del entorno, para prever amenazas o detectar oportunidades, también precisa y genera datos e información. Todos estos datos y toda la información que encontramos en las organizaciones proceden de fuentes de información muy variadas, tanto internas como externas. Muchos de estos datos están producidos en entornos nuevos, como los canales que las organizaciones tienen en las redes sociales, para apoyar campañas de lanzamiento de productos o como herramienta de marketing para conocer tendencias de mercado, entre otros.

Independientemente de las fuentes de información, y tal como se ha visto antes, estos datos precisan de métodos de captura y almacenamiento diferentes de los sistemas de información habituales en las organizaciones. Está claro que una parte queda almacenada en los sistemas de información organizativos. La automatización y el tratamiento que se hace en los proyectos de gestión de datos masivos aseguran esta captación. Sin embargo, se debe recordar que hay datos, información y conocimiento que a pesar de estar dentro de la organización son difíciles de encontrar, porque no están almacenados en los sistemas de información. Por lo tanto, es preciso tener en cuenta los procesos y los procedimientos habituales para gestionar información y conocimiento en el seno de las organizaciones, para vincularlos con los procesos de gestión de datos masivos. De no hacerlo, se corre el peligro de fomentar o de continuar con islas informacionales en las organizaciones. Es decir, tener sistemas informáticos poco relacionados entre sí.

Los sistemas de información son un conjunto de elementos interrelacionados entre sí. Algunos de estos componentes son: procesos y procedimientos que deben estar vinculados a la gestión de la información, infraestructuras formadas por las tecnologías de la información y la comunicación, contenidos y personas, entre otros. Este apartado se ocupa de los procesos y procedimientos, como una vía para garantizar la gobernanza de la información y prevenir la creación de islas informacionales.

La gobernanza de la información es un conjunto de elementos que engloban dos dimensiones: gestión y cultura organizativa, y está orientada al control y al uso de la información desde una perspectiva holística, transversal a toda la organización. El control asegura la calidad y la relación de los datos y de la información existentes en el seno de la organización. Aspectos de detalle que tiene en cuenta la gobernanza de la información son: datos e información como materia prima, estructuras, procesos, reglas y procedimientos, gestión de riesgos, protección de datos personales,

protección de propiedad intelectual y sistemas informáticos para la creación, recogida, análisis, distribución, almacenamiento, valoración, uso y control de la información.

Se debe recordar que la gestión de la información y del conocimiento comporta unos procesos internos que se han agrupado en el ciclo de gestión de la información (Ilustración 6), el ciclo de gestión de la inteligencia competitiva (Ilustración 7) y el ciclo de la gestión del conocimiento (Ilustración 8).

Ilustración 6: Ciclo de gestión de la información (fuente: elaboración propia)

Ilustración 7: Ciclo de gestión de la inteligencia competitiva (fuente: elaboración propia)

Ilustración 8: Ciclo de gestión del conocimiento (fuente: elaboración propia)

A estos tres ciclos, también se puede añadir el ciclo de gestión de los documentos (Ilustración 9) que actúan como evidencias de las actividades de la organización. Estos documentos también contienen información. Además del valor primario de esta información –rendir cuentas de la actividad organizativa–, los documentos contienen un valor secundario que, insertado en los procesos de gestión de la información, contribuye a la creación de valor al que apuntan todos los ciclos.

Ilustración 9: Ciclo de gestión de documentos (fuente: elaboración propia)

Las similitudes entre los tres ciclos y la vinculación con el cuarto apuntan a diferentes facetas en lo que se ha denominado gestión de la información en genérico. A estos ciclos les tenemos que añadir la gestión de los macrodatos, que también queda integrada en el ciclo de gestión de la información, al mismo tiempo que lo hace evolucionar (Ilustración 10).

Ilustración 10: Vinculación de la gestión de datos masivos con la gestión de la información, del conocimiento y documentos (fuente: elaboración propia)

Vemos que la generación de datos no estaba incorporada de manera explícita en el ciclo de la gestión de la información. Aun así, es la causa primera de gestionar información. En el ciclo de gestión de datos masivos, la generación de datos forma parte del primer nivel de la cadena. El énfasis se debe a la relevancia que tiene la diversidad de orígenes de los datos y la diversidad de formatos y tipos, que ocasionan retos importantes para su gestión. Por lo tanto, esta fase se debe incorporar al nuevo ciclo de gestión de la información.

La fase de adquisición de los datos estaría vinculada con la fase de identificación y obtención de información. En esta fase es importante conocer las necesidades de los empleados de la organización en todos los niveles (operativos, tácticos y estratégicos). Los datos y la información obtenida se tienen que tratar y almacenar, siguiendo políticas integrales de gestión de la información. La clasificación y la indexación son parte de estas políticas, pero también tienen que

quedar cubiertos los procesos de gestión del conocimiento: identificar y compartir. La diseminación de información es una de las claves para generar y compartir conocimiento.

El análisis de la información recogida y la interpretación de los resultados del análisis es una fase especificada de manera explícita en el ciclo de la inteligencia competitiva. Como se ha mencionado antes, el objetivo del análisis es reducir el exceso de la información para facilitar su interpretación, de forma que la información se convierta en inteligencia para la acción. Con la interpretación, los decisores pueden comprender el entorno y predecir los cambios que pueden acontecer en este y que podrían impactar en la actividad de la organización. Los procesos de gestión del conocimiento tienen que incorporarse para crear conocimiento y capturarlo en un sistema de almacenamiento de la organización para ser compartido y aplicado en un futuro.

Capítulo III

GENERACIÓN DE CONOCIMIENTO PARA LA ACCIÓN

Los datos son la base para generar información como paso previo a la creación de conocimiento para la acción. Para ello, toda la gestión de los datos masivos se hace a lo largo de su cadena de valor, orientada a crear este conocimiento. Es al final de la cadena donde vemos la aplicación de los grandes volúmenes de datos. Este capítulo se focaliza en el final de esta cadena. Se puede estudiar la aplicación de los grandes volúmenes de datos tanto desde el punto de vista organizativo como desde el punto de vista sectorial. De manera general, podemos ver algunos ejemplos de la aplicación en el ámbito organizativo de la explotación de los macrodatos.

1. Gestión organizativa

El gran volumen de datos que las organizaciones generan en sus operaciones diarias es una base para crear valor. Estos datos están vinculados a varios aspectos organizativos, como por ejemplo el movimiento de los productos, la relación con clientes o suministradores. La mayor parte de las empresas explotan sobre todo los datos internos, aunque también pueden incorporar información del entorno, si tienen implementadas prácticas de inteligencia competitiva.

El origen del valor está en el análisis de los datos, y en que estos –procedentes de diferentes áreas y procesos de la organización– estén conectados. El análisis se puede hacer desde tres visiones.

La **visión perspectiva o descriptiva** mira los datos históricos, se identifican patrones y se observa qué se está haciendo y cómo. El resultado del análisis permite detectar aspectos de mejora, tanto dentro como fuera de la organización.

La **visión prospectiva o predictiva** analiza los datos para identificar patrones y encontrar relaciones entre los datos pero mirando al futuro, para averiguar qué podría suceder, que es más difícil de encontrar con el análisis descriptivo. El resultado del análisis informa de hacia dónde puede evolucionar el entorno y la propia organización.

Por último, la **visión prescriptiva** analiza los datos para conocer cuál es la mejor opción de entre un conjunto de circunstancias. En este sentido, el análisis evalúa y determina nuevas vías para actuar, qué objetivos de negocio por segmentos pueden ser más adecuados y hace un balance de todos los pros y los contras.

Con el análisis se puede observar la variación de diferentes parámetros e indicadores, como

por ejemplo los referentes a los rendimientos obtenidos, de forma que se puede generar conocimiento sobre los diferentes aspectos organizativos. Un cuadro de mando integral, definido adhoc para la organización, facilita medir la evolución de la actividad de una organización y los resultados en relación con los objetivos fijados. Al mismo tiempo da conocimiento interno de cómo se están ejecutando las tareas y empleando los recursos.

Este conocimiento generado es una base para los gestores de las organizaciones, en el aspecto operativo, táctico y estratégico. Estos gestores pueden innovar, diseñar planes estratégicos, diseñar acciones para la planificación interna o definir la relación con el entorno (clientes, actores del territorio o proveedores), crear nuevos modelos de negocio, o fundamentar la toma de decisiones para establecer condiciones comerciales con clientes, socios, políticas laborales con los empleados, planificación de la producción o compras de materias primas, entre otros.

Algunos de los aspectos en los que se puede aplicar la gestión de datos masivos para generar conocimiento y valor para las organizaciones son: cuadro de mando, conocimiento del cliente, producción y suministro, y contabilidad y finanzas.

Cuadro de mando

Esta técnica diseñada por Robert Kaplan y David Norton (Kaplan y Norton, 1997) contempla la evaluación de la visión y la estrategia organizativa desde varias perspectivas: la financiera, la del cliente, la de los procesos internos, el aprendizaje y el crecimiento. Para cada perspectiva se definen objetivos, indicadores, hitos y acciones. El seguimiento de estos apartados requiere mucha información. Por lo tanto, la gestión de datos masivos puede ayudar a crear y seguir el cuadro de mando, además de analizarlo de manera masiva una vez esté hecho.

Conocimiento del cliente

Según un estudio hecho por IBM en 2012, la mayor parte de las empresas orientan el análisis de los datos masivos a temas relacionados con el cliente, para conocer sus preferencias y su comportamiento. Esto se debe a que la adopción de la gestión de datos masivos en las empresas se está haciendo de manera progresiva, y se ha empezado por aplicarla en aquellos ámbitos que pueden proporcionar mayor valor a la organización.

Por otro lado, la segmentación del mercado facilita conocer mejor los productos y los servicios ofrecidos por las organizaciones –no solo los propios, sino también los de la competencia– y, por lo tanto, tomar decisiones sobre la orientación de los productos o servicios según las necesidades de los clientes, crear productos nuevos o mejorar los productos existentes. Con la segmentación del mercado y un mayor conocimiento de este, también se puede definir una política de precios variables según el análisis de datos en tiempo real del mercado. En cuanto a la operatividad, la extracción de conocimiento específico de los clientes también facilita orientar mejor las campañas publicitarias.

Además, para generar y capturar datos relevantes, las empresas intentan establecer canales de

colaboración con sus clientes. Un ejemplo es Ford Focus, que registra datos de los vehículos mientras son conducidos y cuando están aparcados (aceleraciones, presión de los neumáticos, etc.). Esta información la ve el conductor, pero también los ingenieros de Ford, que de este modo pueden conocer los hábitos de conducción de sus clientes, y con este conocimiento pueden diseñar mejoras en los productos que fabrican (Schrorck, 2012).

Producción y suministro

Los adelantos tecnológicos han desarrollado aparatos que graban datos masivos referentes a producción y cadena de suministro. Ejemplos de estos aparatos son sistemas sin cables, teléfonos inteligentes y GPS, entre otros. El análisis perspectivo o descriptivo de los datos referentes a los procesos internos da claves para innovarlos, para hacerlos más eficientes y más productivos. Esta innovación se puede hacer en varios ámbitos, como por ejemplo en la cadena de suministro de materias primas, relación con los proveedores, suministro de las ventas, recursos humanos dedicados, datos de la flota de transporte, entre otros aspectos. Un ejemplo es McLeod Russel India Limited, que eliminó el tiempo de inactividad de los sistemas de comercio del té haciendo un seguimiento de las cosechas, la producción y el marketing.

Otro ejemplo son los datos internos referentes a la flota de vehículos o aviones, registrados por sensores. Estos datos permiten extraer información muy variada para reducir costes y mejorar servicios. Más concretamente, esta información ayuda a prever averías, planificar acciones preventivas, como por ejemplo planificar el calendario de recambios, o bien planificar rutas alternativas. La compañía de transporte UPS ofrece un ejemplo de cómo los datos recogidos en los sensores incorporados en los vehículos permiten diseñar rutas que minimicen las paradas del vehículo debido a semáforos, de forma que el consumo de combustible sea inferior, y que además reduzcan el tiempo de entrega. Air France y KLM también utilizan los datos de sus aviones recogidos durante los vuelos para anticiparse a las averías. DHL es otro ejemplo de cómo explotar los datos disponibles almacenados por su flota, para planificar rutas u ofrecer más servicios según datos grabados de los servicios ya realizados.

Contabilidad y finanzas

La auditoría de las finanzas ha cambiado también fruto de la integración de las TIC en procesos operativos. El comercio electrónico es un ejemplo. La forma de recoger evidencias por parte de los auditores ha cambiado. Los datos masivos también pueden ayudar a mejorar la eficiencia de los procedimientos de auditoría. Aplicando la técnica de minería de datos, los auditores pueden analizar datos externos (datos de censo, *social media*, noticias, entre otras fuentes), para evaluar los riesgos de negocio o de fraude, pueden efectuar controles internos, y analizar transacciones de clientes. Así pues, hay autores que ven en la gestión de los macrodatos un potencial para ayudar a los auditores a llevar a cabo sus tareas, mejorándolas en términos de eficiencia. Aun así, todavía hay escaso consenso sobre el papel que pueden desempeñar los datos

masivos en la auditoría, por lo que es un campo que se debe seguir explorando.

2. Aplicaciones sectoriales

Hay muchas vías por las cuales los datos masivos pueden crear valor en diferentes sectores. Ejemplos son la racionalización de gastos en el ámbito sanitario, gracias al historial clínico compartido por los diferentes médicos que atienden a un paciente, o la detección de señales para luchar contra el terrorismo o para prever acciones que afectan a la seguridad en actos públicos.

Manyika y otros (2011), en un estudio hecho en Estados Unidos, señalan que hay sectores que tienen más potencial que otros para sacar provecho de los macrodatos. Los sectores más posicionados son aquellos que fabrican productos electrónicos y ordenadores, y los vinculados al ámbito de la información. Otros sectores, como el de la administración pública, con un buen potencial para sacar provecho de la gestión de datos masivos, tienen barreras para su explotación. Una de ellas es la de cambiar su mentalidad para estar orientados a la disponibilidad de los datos para explotarlos.

Un estudio más detallado de las iniciativas existentes permitirá conocer con más profundidad las aplicaciones que la gestión de los macrodatos tiene en la vida diaria de las organizaciones para obtener beneficios.

Salud y sanidad

En este sector son diversos los beneficios que se pueden obtener si se aplica la gestión de grandes volúmenes de datos. El más conocido es la optimización de tratamientos médicos y las pruebas médicas que se hacen a los pacientes. Un ejemplo es el estudio que hizo Healthcare Alliance que constata cómo con el análisis de datos masivos mejoraron los resultados de los pacientes y se redujo el gasto en asistencia sanitaria.

Otras aplicaciones se ven en el sector farmacéutico, con la gestión de epidemias o pandemias, o en la creación de un sistema de vigilancia de la salud en todo un país.

Turismo

Es otro de los ámbitos donde la gestión de los datos masivos ayuda a los diferentes actores involucrados en el sector a diseñar acciones y planes estratégicos «más inteligentes». Se encuentran ejemplos tanto desde el punto de vista de las ciudades que quieren promoverse como sector turístico como desde el punto de vista de agencias que quieren construir destinos turísticos atractivos para sus clientes. Tenemos que recordar que las agencias de viajes se tienen que reinventar, dada la autonomía que proporciona internet para organizarse uno mismo el viaje.

Las huellas digitales que dejan los visitantes desde diferentes dispositivos, como tarjetas de créditos o telefonía móvil, constituyen un volumen considerable de datos que, combinado con sistemas de información geográfica, permiten conocer, además del número real de visitantes, datos suficientes como para obtener los perfiles de los turistas que visitan una localidad, sus preferencias, procedencia, necesidades, etc. Estos datos permiten extraer información, con la que diferentes actores locales vinculados con el turismo pueden diseñar productos y servicios adaptados a estos perfiles. Ejemplos son: las compañías áreas, que pueden fluctuar sus precios de vuelos según destino y franja horaria, los hoteles, que pueden definir sus estrategias de reserva, y los restaurantes, que pueden redactar sus menús y platos recomendados, entre otras posibilidades.

Sector financiero

Este sector tiene un gran volumen de datos procedentes de canales diversos, y con formatos estructurados y desestructurados, que precisan ser integrados, para explotar su potencial y hacer frente a las incertidumbres del entorno. Igual que en otros sectores, los datos ayudan a segmentar los clientes para diseñar de forma automática campañas de marketing, o diseñar acciones de fidelización o seguimiento de la competencia, para diferenciarse de esta y obtener más clientes. Ejemplos de otras aplicaciones de los datos masivos son: la gestión de los riesgos implícitos en la concesión de créditos, la gestión de los fraudes en los medios de pago o blanqueo de capitales, financiación del terrorismo, etc. o la detección de señales del entorno para prever riesgos y turbulencias en ese entorno, entre otros.

Además de esta finalidad, la explotación de las redes sociales, el conocer la reputación y el capital social en línea de la propia entidad ayudan a pensar acciones para recuperar la reputación del sector.

Bibliotecas

Este sector desde hace muchos años explota datos cuantitativos generados en su gestión diaria, como por ejemplo el número de usuarios o los títulos más solicitados. Aun así, la gestión de datos masivos permite incorporar más fuentes y extraer más conocimiento, si aplicamos técnicas como la minería de datos, minería de texto o minería web. Por ejemplo, se pueden ver las temáticas de más interés entre los usuarios para gestionar las colecciones, las épocas del año con más consultas o los perfiles demográficos de los usuarios por zonas geográficas, entre otros ejemplos. En el caso de las bibliotecas universitarias, se pueden hacer mapas de conocimiento, con técnicas bibliométricas, a partir de los artículos publicados por los investigadores de la universidad.

Seguros

La gestión de datos masivos ayuda a las compañías aseguradoras a detectar el fraude en las

reclamaciones. Las técnicas clásicas hasta ahora, como las acciones legales y la vigilancia con detectives, son costosas porque requieren mucho tiempo y dinero. El análisis predictivo y la segmentación de riesgos ayudan a identificar patrones para detectar el fraude.

Un ejemplo es la experiencia de la empresa Santam, que desarrolló una solución de análisis avanzado que capta datos procedentes de las reclamaciones presentadas, valora cada reclamación confrontándola con factores de riesgos identificados y divide las reclamaciones en cinco categorías de riesgo, separando las reclamaciones que parecen fraudulentas y de riesgo más alto de los casos de bajo riesgo. De este modo, además de ahorrar dinero con el fraude, reduce el tiempo de procesamiento de las reclamaciones de bajo riesgo.

Administración pública

El sector público –gobiernos y administraciones públicas– también genera grandes cantidades de datos. Por tanto, el sector público también puede explotar datos tanto para su gestión interna, en términos de mejora de la productividad y eficacia para optimizar el dinero de los contribuyentes, como en la gestión del territorio, para generar ventaja competitiva con relación a otros territorios. Un ejemplo claro del uso de los datos es la explotación de los metadatos existentes en los sistemas informáticos referentes a demografía, infraestructuras del territorio o recursos naturales, entre otros. La gestión de estos datos orientados a generar valor en el territorio se puede hacer siguiendo la metodología señalada por la inteligencia territorial (Wartena y García, 2015).

3. Reutilización de la información (sector infomediario)

Los datos y la información que genera la administración pública deben ser abiertos y accesibles a la ciudadanía (individuos o empresas), según la legislación (Ley 18/2015), y directivas europeas (Directiva 2013/37/UE). Por lo tanto, el valor de esta información se multiplica y puede generar amplia diversidad de conocimiento para la acción, dependiendo de quién, cómo y para qué se traten los datos. Estos datos son materia prima de un sector económico emergente, el de la reutilización de la información del sector público, también llamado sector infomediario, que genera productos de información a partir de los datos abiertos.

En el marco de la legislación y las directivas europeas han surgido portales de datos abiertos, donde las administraciones públicas publican sus datos y documentos. La diversidad de fuentes y formatos y el gran volumen de datos deben ser gestionados teniendo en cuenta los procesos de la cadena de valor de los datos masivos. Estos datos constituyen la materia prima para empresas del sector infomediario. La aplicación de diferentes técnicas de análisis facilita la extracción de información para la elaboración de productos de información sobre temas del ámbito de la

cultura o del turismo, información geográfica, meteorológica, entre otros (ASEDIE, 2016).

Capítulo IV

HERRAMIENTAS PARA GESTIONAR MACRODATOS

La gestión de grandes volúmenes de datos requiere de hardware y software específicos con características diferentes de los existentes hasta ahora. Se precisan infraestructuras y herramientas que puedan alcanzar los diferentes aspectos de gestión de datos, como son rapidez de procesamiento, almacenamiento a gran escala y redes potentes y rápidas, a precios más bajos. De esta forma se pueden preservar y utilizar grandes volúmenes de datos a una velocidad más rápida y económica (Kune y otros, 2016; Demchenko y otros, 2014).

Tres aspectos apuntan la necesidad de evolución de los sistemas informáticos que hasta ahora tienen las organizaciones: formatos de datos desestructurados procedentes de fuentes de información hasta ahora no tenidas en consideración (variedad), capacidad de almacenamiento (volumen) y capacidad de procesamiento (velocidad). Estos tres aspectos son los que se utilizan para definir los datos masivos (las tres *V* iniciales que se tomaron en cuenta para las primeras definiciones de *big data* o de los macrodatos). Por eso, los almacenes de datos (*data warehouse*) existentes hasta hace poco, a los que ahora se denomina tradicionales, tienen que incorporar nuevas prestaciones.

Generalmente, las nuevas tecnologías surgidas o que han evolucionado en los últimos años se relacionan con los nuevos retos de la gestión de datos: diversidad, reducción, integración y limpieza, indexación e interrogación, análisis y minería.

Todas las fases de la cadena de valor de la gestión de macrodatos requieren alguna tecnología. Las clasificaciones de estas tecnologías se hacen teniendo en cuenta en qué fase incide la tecnología. Así se habla de tecnologías de almacenamiento y transporte de datos, tecnologías para analizar grandes volúmenes de datos, tecnologías de visualización y tecnologías orientadas a los servicios ofrecidos en la nube (*cloud computing*).

Cada una de estas áreas está desarrollando productos y genera servicios. El ejemplo más representativo y conocido es el *cloud computing* o tecnologías orientadas a servicios que son aplicaciones de software alojadas en redes públicas o privadas basadas en modelos de suministro para establecer las tarifas de precios.

Este capítulo trata estos aspectos y presenta algunos conceptos vinculados con la ingeniería informática, necesarios para entender las herramientas y tecnologías que gestionan datos masivos. Estas aclaraciones conceptuales ayudan a entender y diferenciar términos que encontraremos en diferentes publicaciones cuando queremos profundizar sobre los temas tratados. En segundo lugar, se presenta de manera general la arquitectura en la que se fundamenta la gestión de los datos masivos, y algunas de las herramientas que se vinculan a las diferentes fases de la cadena de valor de los macrodatos. Por último, se exponen las diferentes aplicaciones existentes en el

mercado, y que se deben tener en cuenta cuando se implanten tecnologías para gestionar datos masivos.

1. Conceptos

En el ámbito de la ingeniería informática se habla de arquitecturas, plataformas, herramientas, tecnologías y sistemas de información para referirse a las herramientas que hay detrás de la gestión de datos (información o documentos) o de la creación de estos. Este apartado presenta algunos de estos conceptos, para facilitar la comprensión de los conceptos expuestos en este capítulo, y facilitar la lectura de la literatura publicada sobre el tema.

Arquitectura

En el ámbito de la ingeniería informática a menudo se habla de arquitectura, tanto aplicada al software como al hardware. Cuando se aplica en el software, hace referencia al conjunto de componentes lógicos e intangibles que forman parte de un programa o un sistema, sus funciones y relaciones para dar respuesta a un problema específico, en un contexto concreto, y los principios que guían su diseño y evolución (Bass y otros, 2012; ISO/IEC/IEEE 42010:2011 System and software engineering – Architecture description y estándar IEEE 1471:2000). Cuando se relaciona con el hardware hace referencia a los componentes físicos y tangibles de un ordenador, de una tecnología, equipo electrónico, informático, periféricos (disco duro, CD-ROM, etc.), etc.

La complejidad de la arquitectura del software y la diversidad de prácticas para su desarrollo impulsaron la realización de un estándar por parte del IEEE (Institute of Electrical and Electronics Engineers) (IEEE 1471:2000). Este estándar ha sido sustituido por otro, el ISO/IEC/IEEE 42010, que, tomando la base del estándar IEEE 1471, describe los sistemas y productos software y define los requisitos que deben cumplir las descripciones que se hagan de arquitecturas empresariales, de sistemas o de software. Por lo tanto, el principal objetivo es normalizar y homogeneizar las prácticas para describir arquitecturas. Este estándar ofrece un glosario común y un marco conceptual que facilitan cómo redactar, comunicar y revisar las descripciones y especificaciones de la arquitectura y sus requisitos empleando un lenguaje concreto. Es decir, puntos de vista, marcos y lenguajes de introducción ayudan a codificar convenciones y prácticas comunes de la descripción de las arquitecturas.

Plataformas / software para el procesamiento o almacenamiento de información

En informática una plataforma tecnológica es una base para hacer funcionar unos módulos concretos de software o de hardware compatibles. En el diseño de la plataforma también se tiene en cuenta el sistema operativo, lenguaje de programación e interfaz de usuario.

Sistemas de información

Un sistema de información es un conjunto de elementos interrelacionados entre sí, algunos de los cuales son recursos informáticos, que tienen la función de asegurar la transformación y disponibilidad de la información a partir de los procedimientos que rigen el sistema. Un sistema de información recopila, procesa, almacena y difunde información para cumplir un objetivo específico.

Los sistemas de información se clasifican dependiendo de a qué nivel organizativo apoyen. Así tenemos sistemas operacionales que apoyan las transacciones básicas de negocio y sistemas de apoyo a las decisiones. Estos últimos han ido evolucionando hacía los sistemas de *business intelligence* o inteligencia de negocio, que tienen como núcleo los denominados almacenes de datos o *data warehouse* que apoyan también el nivel táctico y estratégico. Todos estos sistemas de información están basados en bases de datos relacionales que solo gestionan datos estructurados. Con la introducción de los grandes volúmenes de datos, y de los datos desestructurados, estos sistemas de información han llegado a su límite, y por tanto necesitan cambios. Estos se están produciendo en todos los ámbitos: software, hardware e infraestructuras de telecomunicaciones.

Los datos masivos precisan tecnologías específicas para procesar a una alta velocidad grandes cantidades de datos variados. Los sistemas de datos masivos necesitan almacenar y gestionar conjuntos de datos heterogéneos y masivos, además de garantizar su función y rendimiento en términos de rápida recuperación, escalabilidad y protección de la privacidad (Hu y otros, 2014).

En el ámbito del hardware también se observan limitaciones para gestionar datos masivos. Las tecnologías y el hardware más avanzados tienen dificultades para procesar un gran volumen de datos empleando un solo ordenador. Por eso se apunta al *cloud system* como una vía para procesar datos en bruto para obtener datos refinados.

Ecosistema de los datos masivos

La producción y gestión de datos masivos involucra un amplio abanico de elementos tanto tecnológicos como organizativos y empresariales. De aquí que se empiece a hablar de ecosistema de macrodatos (Demchenko y otros, 2016), para referirse a todos los componentes complejos relacionados con los datos –modelos, infraestructuras, tecnologías, protocolos y arquitecturas– empleados en el ciclo de vida de los datos para almacenar, procesar, visualizar y entregar resultados.

2. Arquitecturas para los macrodatos

La mayoría de la literatura sobre datos masivos se focaliza en un componente tecnológico o en una solución vinculada a un único problema de los muchos que plantea la gestión de datos masivos.

Actualmente hay muchas herramientas en el mercado y, por lo tanto, es preciso conocer las prestaciones de cada una de ellas así como las necesidades de información de la organización para generar valor y ventaja competitiva.

La diversidad de prestaciones está vinculada con las necesidades de cada una de las fases de la cadena de valor de la gestión de los datos masivos (Ilustración 11).

Ilustración 11: Tecnologías involucradas en cada fase de la cadena de valor de los datos masivos (fuente: elaboración propia)

Generación de datos

Los datos están generados por diferentes hardware y software: sensores (de temperatura o de movimiento), redes sociales, correo electrónico, sistemas de información organizativos, páginas web o bases de datos de organizaciones que empleamos como fuentes de información. Exceptuando los sistemas de información disponibles en la organización, muchos datos provienen de dispositivos y tecnologías que quedan fuera del control de la organización.

Los datos generados quedan almacenados en las mismas infraestructuras que los producen. Periódicamente los datos en bruto deben ser extraídos de estas infraestructuras y traspasados a los sistemas de almacenamiento definidos por la organización. Estos datos tienen que ser tratados y además preservar siempre su contexto y origen para preservar la calidad de los datos.

Adquisición de datos

En esta fase de la cadena, la organización obtiene datos de diferentes procedencias y en diferentes formatos, de acuerdo con sus necesidades de información. Estos datos se deben tratar tanto en cuanto a formatos como a contenido, de manera que ofrezcan una visión única de los datos. Para ello se requiere una serie de procesos que garanticen la calidad de los datos, y eso se efectúa mediante su trazabilidad: fecha y hora exacta en la que el dato ha sido extraído, momento en que se produjo la transformación, y el instante en que se cargó desde la fuente (el entorno de origen) hasta el destino.

Además, en esta fase, cuando se evalúan las herramientas que efectuarán los procesos de preparación, interpretación e integración de datos, se debe tener en cuenta la velocidad de carga y la velocidad de procesamiento. La primera se refiere a los procesos de extracción, transformación y carga de datos (ETL) en el *data warehouse* o plataforma que se encargará de obtener los datos (Gómez y Conesa, 2015).

A tal fin, las herramientas ETL (extract, transform and load) que se escojan para esta fase

deben disponer de un motor de base de datos potente y una plataforma de integración que incorpore: funcionalidades de calidad de datos; funciones de gestión de datos maestros o MDM (*master data management*), para asegurar la uniformidad, precisión, administración, semántica, consistencia y responsabilidad de los datos; capacidad para la gestión de metadatos; posibilidad de gestionar la trazabilidad en los datos, y herramientas de diseño con capacidades de reingeniería inversa, para modelar y construir.

Almacenamiento

El almacenamiento de los datos masivos se hace en sistemas de archivos, bases de datos y modelos de programación, para los que hay diversidad de software en el mercado.

a) Sistemas de archivo

Como todos los sistemas de almacenamiento, los destinados a gestionar grandes volúmenes de datos tienen que almacenar, organizar, nombrar, compartir y proteger los archivos. Aun así, dado que los datos y los sistemas de almacenamiento están repartidos en diferentes lugares, debido a su volumen y diversidad de producción, la gestión de los ficheros de datos masivos debe cumplir toda una serie de requisitos. Estos son: rendimiento de la lectura y escritura, acceso a datos simultáneos, creación de sistemas de archivos según demanda y técnicas eficientes para sincronizar archivos (Kune y otros, 2016). Por lo tanto, en el momento de diseñar sistemas de archivos se deben tener en cuenta los siguientes aspectos (Kune y otros, 2016):

- **Acceso distribuido y transparencia en la localización**. Los usuarios no son conscientes de que los ficheros están distribuidos, por lo tanto, deben disponer de directorios unificados, de forma que los usuarios puedan acceder del mismo modo que lo hacen a los archivos locales. Esto implica que haya consistencia entre los nombres de los ficheros locales y los remotos.
- **Gestión de fallos**: los programas de las aplicaciones y los usuarios deben seguir operando incluso cuando el sistema tenga componentes que fallan. Esto se consigue con algunos niveles de reproducción y redundancia.
- **Heterogeneidad**: El sistema de archivos se debe componer de diversidad de hardware y plataformas de sistemas operativos.
- **Distribución muy definida de los datos**: para optimizar el rendimiento, es preferible ubicar objetos individuales cerca de los procesos que los emplearán.
- **Tolerancia a la partición de la red**: Toda la red o ciertos segmentos de esta pueden ser inaccesibles a un usuario en ciertos periodos (por ejemplo, un portátil se desconecta de una operación). El sistema de archivos debe ser tolerante para gestionar la situación y aplicar mecanismos de sincronización apropiados.

b) Tecnologías de bases de datos

El almacenamiento y la recuperación de datos e información se efectúan en bases de datos. Para gestionar datos masivos, sobre todo los desestructurados, las bases de datos empleadas son diferentes de las bases de datos relacionales, dado que estas no pueden dar respuesta a los retos que los datos masivos plantean, en cuanto a categorización y volúmenes (Chen y otros, 2014). Las nuevas bases de datos tienen dos características relevantes: 1) no seguir el esquema entidad-relación, y por tanto carecer de estructura de datos prefijada de tablas y relaciones; y 2) no utilizar el lenguaje SQL. Este último aspecto da nombre a este tipo de bases de datos, NoSQL (*not only SQL*).

c) Modelos de programación

Los macrodatos normalmente se almacenan en centenares o miles de servidores comerciales, que operan con modelos de programación paralelos para procesar los datos. Los modelos paralelos tradicionales, como MPI (*message passing interface*) y OpenMP (*open multi-processing*), pueden ser inadecuados para operar con los programas paralelos a gran escala. Por eso, han surgido otros modelos de programación paralelos para mejorar el rendimiento de NoSQL y reducir el vacío de rendimientos de las bases de datos relacionales. Estos modelos son clave para la fase posterior de análisis de datos masivos.

Algunos de estos modelos de programación son: MapReduce, Dyrad, Ajo-Pairs (especial para biometría, bioinformática y minería de datos) y Pregel (para procesar grafos de grandes medidas) (Chen y otros, 2014).

Análisis de datos y visualización

La gestión de esta fase de la cadena de valor también opera con tecnologías que, como en las otras fases, también deben tener en cuenta la perspectiva del volumen de datos, velocidad de generación y procesamiento y la variedad de los datos (Tsai y otros, 2015). Otro aspecto que debe tenerse en cuenta es el tipo de análisis que se requiere hacer para responder a las necesidades de información identificadas previamente (análisis descriptivo, predictivo o prescriptivo) (Hu y otros, 2014). Para analizar los datos se precisan métodos de análisis, arquitectura de análisis en tiempos reales y *offline*, y software para explotar y analizar los datos masivos (Chen y otros, 2014).

a) Métodos de análisis

Entre los métodos de análisis tradicionales destacan la visualización de datos, el análisis por clústeres, análisis de factores, análisis de regresiones, análisis estadístico y algoritmos de minería de datos. Para el análisis de estos datos, las herramientas como *data warehouse* ya incluyen software que facilita las operaciones asociadas en estas tecnologías. Las herramientas más empleadas para cálculos estadísticos son SPSS y Excel.

Para la gestión de los datos masivos, sobre todo para las que tienen datos desestructurados y semidesestructurados han surgido nuevos métodos de análisis con el objetivo de extraer rápidamente información clave. Estos métodos son:

- *Bloom filter*: almacena valores de datos de encriptación (*hash*).
- *Hashing*: transforma los datos en valores numéricos de longitud fija, o en valor de indexación.
- *Index*: mejora la velocidad de las acciones de insertar, borrar, modificar e interrogar.
- *Triel*: se utiliza para recuperar rápidamente información y hacer estadísticas de frecuencia de palabras.
- *Parallel computing*: es el uso en paralelo de varios recursos informáticos para completar las tareas computacionales. Descompone un problema y asigna las partes a varios procesos separados, para ser completados de manera individual.

b) Arquitectura de análisis

Esta arquitectura tiene en cuenta las características de los datos masivos de velocidad, variabilidad, volumen y valor. Las arquitecturas se construyen considerando los siguientes aspectos (Chen y otros, 2014):

- La **presión de tiempo**, por lo cual el análisis se hará en tiempo real u *offline*. El análisis se hará en tiempo real cuando los datos cambian constantemente y además se necesita analizarlos con rapidez. El análisis *offline* se hará cuando los datos se analizan sin requisitos de tiempos. A tal fin se pueden utilizar arquitecturas que reduzcan el coste de conversión del formato de los datos y mejorar la eficiencia de adquisición de datos.

- El nivel de **volumen de memoria** que se requiere para hacer el análisis. Desde este punto de vista tenemos tres tipos de arquitectura: a) análisis de nivel de memoria, cuando el volumen total de datos es pequeño y no sobrepasa el máximo de memoria del clúster; b) análisis de inteligencia de negocio, que se utiliza cuando la escala de datos sobrepasa el nivel de memoria, pero puede ser importante en el entorno de análisis de la inteligencia de negocio; y c) análisis masivo, cuando el volumen de datos sobrepasa la capacidad de productos y bases de datos relacionales tradicionales. La mayoría de los datos que se analizan son para hacer un análisis *offline*.

- La **complejidad** de los diferentes datos que deben analizarse, que influye en el tipo de algoritmo que seleccionamos para hacer el análisis.

El análisis visual de la información extraída es una potente herramienta para comparar modelos y conjuntos de datos que facilita la interpretación y la toma de decisiones. Por eso hay herramientas que han trabajado mucho este aspecto.

c) Criterios de selección

La diversidad de herramientas de análisis y visualización existente es amplia, por lo que es preciso tener claro cuáles son las necesidades de la organización para seleccionar la más adecuada. Dependiendo del tipo de análisis y de qué información se quiere obtener y cómo se quiere visualizar, se deberán escoger unas herramientas u otras, dado que no son iguales las que hacen análisis predictivo que las que pueden hacer análisis en tiempo real. Los estudios comparativos de

la consultora Gartner y sus resúmenes plasmados en los denominados cuadros mágicos son una útil fuente de información que puede servir de orientación en la selección del software.

Hay diversos tipos de aplicaciones para mostrar el resultado del análisis, entre las cuales destacan: *scorecards* o cuadros de mando, informes predefinidos, informes a medida, consultas (*queries*) o cubos OLAP (*online analytic processing*), alertas, análisis estadístico, pronóstico, modelado predictivo o minería de datos, optimización y minería de procesos.

3. Plataformas, tecnologías y aplicaciones más empleadas

Son muchas las herramientas al alcance de las organizaciones que quieren iniciar un proyecto de gestión de macrodatos. Además se deben considerar los aspectos más relevantes en la selección de herramientas, y se debe alinear la selección a las necesidades reales de la organización. Este apartado introduce algunos conceptos relacionados con las herramientas que se están utilizando más en la configuración de las plataformas de datos masivos.

Hadoop

Es un software de código abierto para el almacenamiento[1], el procesamiento y el análisis de grandes volúmenes de datos (desestructurados, estructurados y semiestructurados). Está disponible bajo licencia Apache 2.0[2], compatible con otras licencias de código abierto. Los conjuntos de datos están distribuidos en grupos de ordenadores que utilizan modelos sencillos de programación. Este software está diseñado para pasar de servidores individuales a miles de máquinas (Camargo y otros, 2015). Una característica importante es su escalabilidad, por lo que puede crecer agregando módulos: Hadoop Distributed File System (HDFS) (sistema de archivos distribuidos Hadoop) y HadoopMapReduce.

Esta plataforma es la base de una buena parte de las soluciones que se comercializan en el mercado. Además, se puede implementar sobre hardware a un coste relativamente bajo.

Hadoop Distributed File System (HDFS)[3]

Es un sistema de archivos basado en una arquitectura maestra-esclavo. Es escalable, tolera los fallos (mal funcionamiento del hardware o del software), y cuenta con una arquitectura distribuida. Los archivos están distribuidos en varias máquinas para su procesamiento, a pesar de que parece que trabaje en un solo archivo.

MapReduce[4]

Es un modelo de programación creado por aplicaciones que deben procesar en un hardware

grandes cantidades de datos de forma paralela, y puede ser ejecutado en varios lenguajes de programación, como Java, Ruby, Python y C++. Cuando los datos entran para ser procesados, estos se dividen en grupos para su procesamiento de manera distribuida, en paralelo, en diferente hardware para después combinar el resultado.

Hadoop Eco System

Para incrementar la eficiencia y funcionalidad de Hadoop se han desarrollado otras tecnologías especialmente diseñadas para gestionar datos masivos con Hadoop y MapReduce. Al conjunto de estas tecnologías se denomina Hadoop Eco System, y son Apache PIG[5], Apache HBase[6], Apache Hive[7], Apache Sqoop[8], Apache Flume[9] y Apache Zookeeper[10].

Capítulo V

COMPETENCIAS PARA LA GESTIÓN DE DATOS MASIVOS

En los últimos años, se señala al sector de la gestión de los datos masivos como un ámbito con grandes expectativas de ocupación. En efecto, en el mercado laboral en los últimos años ha habido un incremento considerable de demanda de profesionales para implementar y desarrollar proyectos de gestión de datos masivos.

Los propios procesos asociados a la cadena de valor de gestión de los datos masivos requieren profesionales especializados en cuatro ámbitos: a) desarrollo e implementación de soluciones de gestión de datos masivos; b) desarrollo de las actividades para gestionar los datos en cuanto a obtención, tratamiento, almacenamiento de los datos y mantenimiento de la calidad de estos; c) administración y mantenimiento de la arquitectura de los sistemas informáticos; y d) extracción de información y creación de conocimiento para obtener los beneficios de los datos mediante el análisis.

Dado que la gestión de datos masivos alcanza diferentes aspectos y tareas, las competencias asociadas a las ofertas de trabajo también tendrán que ser diversas. A esto se debe añadir que la explotación de datos masivos es de unas dimensiones tales que involucra a más de un profesional. Esta diversidad conlleva que la gestión de datos masivos requiera de conocimientos de ámbitos diferentes, no solo informáticos. Por lo tanto, está claro que los equipos de profesionales involucrados en la gestión de los macrodatos están caracterizados por la multidisciplinariedad.

En consecuencia, las terminologías surgidas para definir a los profesionales de la gestión de datos masivos son diversas, y carecen de consenso en el mercado laboral y entre los expertos. Este hecho es un indicio de que el gestor de datos masivos es un perfil en construcción (Ilustración 12). Aun así, el proyecto EDISON en Europa está trabajando en la definición de los perfiles, las competencias y los conocimientos asociados a la gestión de datos masivos (Demchenko y Belloum, 2016).

Ilustración 12: Perfiles asociados a la gestión de datos masivos (fuente: elaboración propia)

En este capítulo presentamos dos temas. El primero describe las competencias y los conocimientos que se requieren para llevar a cabo las diferentes actividades de la cadena de valor de los datos masivos. El segundo apartado ofrece un abanico de perfiles identificados en el ámbito de los datos masivos.

1. Habilidades y conocimientos

La gestión de datos masivos requiere de conocimientos multidisciplinares y habilidades diversas: técnicas, investigativas, analíticas, interpretativas, comunicativas para presentar los resultados de los análisis y creativas. Por otro lado, además de estas habilidades son precisos conocimientos del ámbito de la gestión y la administración de empresas, ingeniería, informática, estadística, matemáticas y gestión de la información y documentación (Demchenko y Belloum, 2016; SAS, 2013a y 2013b) e incluso de humanidades.

Desde el punto de vista de la gestión y la administración de empresas es necesario conocer las funciones y las actividades del negocio, es decir, los procesos de negocio. Desde el punto de vista de la ingeniería informática se deben conocer varios tipos de software, lenguajes de programación vinculados a bases de datos y metodologías de desarrollo de software (SAS, 2013a y 2013b). Es esta perspectiva de las arquitecturas y aplicaciones informáticas la que predomina, y, en consecuencia, en el mercado laboral encontramos una fuerte tendencia a identificar la gestión de datos masivos con perfiles tecnológicos. Aun así, si tomamos como base la cadena de valor de la gestión de datos masivos, veremos que otros perfiles son también relevantes, y no emergen en el perfil demandado en el mercado laboral (Ilustración 12). Así, por ejemplo, en la gestión de los datos es necesario poseer conocimiento sobre legislación que afecta a la protección de datos, aspectos éticos de la cesión de datos, y conocimiento sobre gestión de datos para preservar su calidad y facilitar su recuperación en sistemas orientados al usuario que velen por la facilidad de acceso y usabilidad de los datos (almacenamiento, mantenimiento, reutilización y extracción de conocimiento). Estos temas se trabajan en el grado en Información y Documentación.

Competencias y conocimientos relacionados con las infraestructuras que gestionan los datos masivos

De manera transversal, todas las fases de la cadena de valor precisan de software y hardware para generar, captar, almacenar y analizar información. Por lo tanto, será necesario construir y desarrollar esta infraestructura. Esto implica conocer metodologías de desarrollo de software, lenguajes de programación y bases de datos.

En cuanto a las **metodologías de desarrollo de un software**, a pesar de que las actividades son las mismas, hay diferencias respecto a cómo y cuándo se tienen que llevar a cabo, y esto da lugar a métodos diferentes. Por ejemplo, las ***metodologías ágiles*** son un conjunto de métodos de desarrollo de software iterativos que permiten cambiar la idea inicial a medida que avanza el proyecto. La obtención de información empírica facilita el descubrimiento de soluciones. Un método iterativo organiza el desarrollo en una serie de iteraciones, cada una de las cuales es un miniproyecto contenido que amplía el resultado final de la iteración anterior. Se asegura así que al final de la iteración haya un resultado utilizable.

- **Scrum** es un método ágil de desarrollo de software, iterativo e incremental, que sigue el principio de generar solo los artefactos que aportan valor importante, minimizando el conjunto

de prácticas y artefactos (documentos, modelos, programas, etc. que se generan como resultado del trabajo del ingeniero), tareas y roles.

- **TDD** (*test-driven development*) o desarrollo guiado por pruebas (DGP) es un método de desarrollo de software guiado por pruebas. Los requisitos se trasladan a pruebas. Primero se hace un test automático de prueba, se ejecuta y se ve qué falla, para después hacer el código necesario para que los tests se hagan de manera correcta, limpiando previamente las partes de código que no funcionan.

- **SOA** (*service oriented architecture*) o arquitectura orientada a servicios es una técnica para poner en práctica los métodos que se hayan elegido para desarrollar un proyecto de software. Una de estas técnicas es la reutilización, y SOA es un ejemplo. SOA tiene como objetivo incrementar el nivel de granularidad de la unidad de abstracción. Se reutiliza el servicio completo en vez de reutilizar una parte del código del sistema que se quiere desarrollar.

- **BDD** (*behavior-driven development*) es otra técnica para desarrollar software, guiada por el comportamiento (DGC), surgida a partir del desarrollo guiado por pruebas (DGP). El desarrollo está guiado por el comportamiento y combina las técnicas generales y los principios de DGP.

- Con relación a los lenguajes de programación, **OO** (*object oriented*) u orientación a objetos es un lenguaje de programación que permite escribir programas en términos del más alto nivel de abstracción del que la máquina puede ejecutar por sí misma. La orientación a objetos propone hacer una abstracción del problema que se está intentando resolver en lugar de hacerla la máquina que finalmente ejecutará el programa. Este enfoque resulta un paradigma muy útil no solo para la programación sino también para el análisis.

- Por último, **OOP** (*object –oriented programming*) o programación orientada a objetos es un tipo de programación basada en dos técnicas que facilitan la reutilización: la ocultación de información y la abstracción. La ocultación de información consiste en esconder los detalles sobre la estructura interna de un módulo, para definir los aspectos de los objetos que son visibles públicamente (y, por lo tanto, utilizables por los reutilizadores) y los que no. La abstracción consiste en identificar los aspectos relevantes de un problema para concentrarse solo en estos.

Específicamente, Demchenko y Belloum (2016) identifican las siguientes competencias, asociadas al desarrollo de software e infraestructuras (hardware):

- Usar principios de ingeniería para investigar, diseñar o desarrollar estructuras, instrumentos, máquinas, experimentos, procesos, sistemas teorías o tecnologías.
- Desarrollar herramientas de análisis de datos especializados para apoyar a la toma de decisiones ejecutivas.
- Diseñar, construir y gestionar bases datos relacionales y no relacionales.
- Desarrollar y aplicar soluciones informáticas a los problemas relacionados con el ámbito de conocimiento empleando plataformas de análisis de datos de amplio rango.
- Desarrollar soluciones para el acceso seguro y fiable a los datos.
- Desarrollar algoritmos para analizar varias fuentes de datos.
- Hacer prototipo de aplicaciones de análisis de datos nuevos.

Estas competencias tienen asociado un conjunto de ámbitos de conocimientos, entre los que destacan: inteligencia artificial, aprendizaje automático, estadística, lenguajes de programación y metodologías de desarrollo de software.

Competencias y conocimientos vinculados al ciclo de vida de los datos

Para seleccionar los datos que pueden satisfacer las necesidades de negocio, tratarlos y mantenerlos según el ciclo de vida de los datos, se requieren las competencias denominadas de gestión de datos, custodia (*data curation*) y preservación (Demchenko y Belloum, 2016). Las competencias relacionadas son:

- Desarrollar e implementar la estrategia de los datos.
- Desarrollar modelos de datos incluyendo metadatos.
- Integrar diferentes fuentes de datos y suministrarlas para análisis futuros.
- Desarrollar y mantener un repositorio de análisis de datos históricos.
- Recoger y gestionar diferentes fuentes de datos.
- Visualizar datos variables y complejos, incluyendo los procedentes de un dominio específico.

Los ámbitos de conocimiento asociados a estas competencias son: planificación de gestión de datos, indexación y clasificación de datos, creación de metadatos, diseño y mantenimiento de procesos para garantizar la calidad de datos (precisión, completitud, relevancia, validez, proveniencia, autenticidad, veracidad, exactitud, reputación y credibilidad) y preservación de los datos. Estos ámbitos de conocimiento se asocian con el perfil denominado archivero y bibliotecario digital.

Conocimientos vinculados con el análisis de los datos y la extracción de conocimiento para la acción

Para analizar datos Demchenko y Belloum (2016) identifican las siguientes competencias:

- Usar apropiadamente los métodos estadísticos aplicados a los datos para extraer conocimiento.
- Usar el análisis predictivo para analizar datos masivos y descubrir nuevas relaciones.
- Buscar y analizar conjuntos de datos complejos, combinar diferentes fuentes y tipos de datos para mejorar el análisis.
- Desarrollar un análisis especializado para facilitar la toma de decisiones ágil.
- Saber aplicar el aprendizaje automático (*machine learning*).
- Saber analizar el negocio.

Los conocimientos asociados al análisis son minería de datos, minería de texto, minería web, análisis de redes sociales, estadística, procesos de negocio, visualización y conocimientos de bases de datos a nivel de usuario.

Competencias y conocimientos multidisciplinares

Para extraer el valor de los datos hay que conocer el ámbito de la organización donde se trabaja, el sector, y de manera específica sus objetivos y su mercado. Las competencias identificadas son:

- Comprender el ámbito de negocio y suministrar información, trasladar problemas de negocio desestructurados a un marco matemático abstracto.
- Usar datos para mejorar servicios ya existentes o desarrollar nuevos servicios.
- Participar de manera estratégica y táctica en decisiones financieras que impactan en la gestión y a las organizaciones.
- Recomendar e implementar objetivos estratégicos relacionados con el negocio y alternativas.
- Dar servicios de apoyo científico, técnico y analítico a los roles organizativos.
- Analizar fuentes de datos múltiples con finalidades comerciales.
- Analizar los datos de los clientes para identificar y optimizar acciones de relaciones con los clientes.

Competencias relacionadas con métodos de investigación o gestión de procesos de negocio

Demchenko y Belloum (2016) identifican dos ámbitos donde se aplicarán los resultados del análisis de datos: el ámbito de investigación y el ámbito de negocio. Por eso prefieren hablar de investigación o de procesos de negocio. Sea en uno o en otro ámbito, las competencias identificadas son:

- Crear nuevo conocimiento y capacidades empleando técnicas de métodos científicos (hipótesis, test y evaluación); revisión crítica o métodos de investigación y desarrollo del ámbito de la ingeniería.
- Estudiar sistemáticamente para lograr conocimiento o comprensión de los aspectos fundamentales de fenómenos y hechos observables, y descubrir nuevos enfoques para lograr nuevos objetivos.
- Llevar a cabo trabajo creativo, empleando de manera sistemática la investigación o experimentación, para descubrir o revisar conocimiento de la realidad, y emplear este conocimiento para idear nuevas aplicaciones.
- Aplicar ingenio a problemas complejos, y desarrollar ideas innovadoras.
- Saber trasladar estrategias hacia planes de acción y seguirlos hasta su finalización.
- Influenciar el desarrollo de objetivos organizativos.

Cuadro resumen de competencias

A continuación presentamos un cuadro resumen (Tabla 2) de las competencias antes expuestas, y que aparecen en el proyecto EDISON de identificación del profesional de la gestión de datos masivos (Demchenko y Belloum, 2016):

Tabla 2: Competencias de los profesionales responsables de la gestión de macrodatos (fuente: elaboración propia)

GRUPOS DE COMPETENCIAS	**DETALLE**
Análisis de datos	- Usar apropiadamente los métodos estadísticos aplicados a los datos para extraer conocimiento. - Usar el análisis predictivo para analizar datos masivos y descubrir nuevas relaciones. - Buscar y analizar conjuntos de datos complejos, combinar diferentes fuentes y tipos de datos para mejorar el análisis. - Desarrollar un análisis especializado para facilitar la toma de decisiones ágil. - Saber aplicar el aprendizaje automático (*machine learning*). - Saber analizar el negocio.
Desarrollo de software e infraestructura (hardware)	- Usar principios de ingeniería para investigar, diseñar o desarrollar estructuras, instrumentos, máquinas, experimentos, procesos, sistemas teorías o tecnologías. - Desarrollar herramientas de análisis de datos especializados para apoyar a la toma de decisiones ejecutivas. - Diseñar, construir y gestionar bases de datos relacionales y no relacionales. - Desarrollar y aplicar soluciones informáticas a los problemas relacionados con el ámbito de conocimiento empleando plataformas de análisis de datos de amplio rango. - Desarrollar soluciones para el acceso seguro y fiable a los datos. - Desarrollar algoritmos para analizar varias fuentes de datos. - Hacer prototipo de aplicaciones de análisis de datos nuevos.
Competencias y conocimiento de temas científicos	- Comprender el ámbito de negocio y suministrar información. - Trasladar problemas de negocio desestructurados a un marco matemático abstracto. - Usar datos para mejorar servicios ya existentes o desarrollar nuevos servicios. - Participar de manera estratégica y táctica en decisiones financieras que impactan en la gestión y a las organizaciones. - Recomendar e implementar objetivos estratégicos relacionados con el negocio y alternativas. - Dar servicios de apoyo científico, técnico y analítico a los roles organizativos. - Analizar fuentes de datos múltiples con finalidades comerciales. - Analizar los datos de los clientes para identificar y optimizar acciones de relaciones con los clientes.
	- Desarrollar e implementar la estrategia de los datos. - Desarrollar modelos de datos incluyendo metadatos.

Gestión de datos, custodia (*data curation*) y preservación	- Integrar diferentes fuentes de datos y suministrarlas para análisis futuros. - Desarrollar y mantener un repositorio de análisis de datos históricos. - Recoger y gestionar diferentes fuentes de datos. - Visualizar datos variables y complejos, incluyendo los procedentes de un dominio específico.
Métodos de investigación	- Crear nuevo conocimiento y capacidades empleando técnicas de métodos científicos (hipótesis, test y evaluación); revisión crítica o métodos de investigación y desarrollo del ámbito de la ingeniería. - Estudiar sistemáticamente para lograr conocimiento o comprensión de los aspectos fundamentales de fenómenos y hechos observables, y descubrir nuevos enfoques para lograr nuevos objetivos. - Llevar a cabo trabajo creativo, empleando de manera sistemática la investigación o experimentación, para descubrir o revisar el conocimiento de la realidad, y emplear este conocimiento para idear nuevas aplicaciones. - Aplicar ingenio a problemas complejos y desarrollar ideas innovadoras. - Saber trasladar estrategias hacia planes de acción y seguirlos hasta su finalización. - Influenciar en el desarrollo de objetivos organizativos.

2. Perfiles

Son muchas las terminologías que han ido surgiendo a lo largo de los últimos años para denominara los profesionales encargados de la gestión de datos masivos. Al margen de las terminologías existentes, y tomando como base tanto la cadena de valor antes descrita como los profesionales existentes, se pueden asociar tres perfiles profesionales a la gestión de datos masivos:

1) ingenieros informáticos, en cuanto a construcción y desarrollo de bases de datos y plataformas que generen, capturen, procesen y almacenen los datos, la información creada y los resultados de los análisis;

2) ingenieros de telecomunicaciones, que implementen y velen por la transmisión de los datos;

3) profesionales de la información, que velen por la calidad del ciclo de vida de los datos y del ciclo de vida de la información y del conocimiento, analicen los datos y faciliten a los directivos de las organizaciones la generación de conocimiento para la acción, mediante la creación de productos de información como informes o resúmenes. Las tareas de estos profesionales son: identificación de necesidades, captura de datos e información –según las necesidades detectadas–, correcto almacenamiento, clasificación e indexación –para su rápida recuperación– y análisis y creación de productos de información. Estos profesionales de la información adquieren su formación básica en el grado en Información y Documentación.

En el ámbito de la informática, en cuanto a los profesionales que desarrollan las estructuras y

arquitecturas para la gestión, se han identificado los siguientes perfiles: desarrolladores, arquitectos, analistas, administradores, gestores de proyectos de datos masivos, diseñadores y científicos de datos (Tabla 3). Con relación al científico de datos hay autores que los dotan de más competencias que extrapolan el ámbito de la ingeniería informática. Por eso, lo tratamos en un subapartado específico.

Tabla 3: Resumen de perfiles de ingeniería informática y conocimientos asociados (fuente: elaboración propia a partir de SAS, 2013)

PERFILES	**DETALLES**	**COMPETENCIAS**	
		TÉCNICAS	**METODOLÓGICAS**
Desarrollador de	*Business intelligence* Web Software Negocio Analista Aplicaciones Bases de datos *Front-End*	NoSQL Java SQL Javascript MySQL Linux Oracle Hadoop HTML	TDD (*test driven development*) CSS Metodologías ágiles
Arquitecto de	Soluciones Datos *Business intelligence*	Oracle Java SQL Hadoop SQL Server	*Data modelling* ETL *Enterprise architecture* *Open source* Análisis
Analista de	negocio datos inteligencia de negocio apoyo	Oracle (BI EE e Informes) SQL Java	*Data modelling* ETL Análisis Análisis de datos
Administrador de	Sistemas (Linux o Unix) Bases de datos (Oracle, SQL, Teradata, MySQL)	Linux MySQL Puppet Hadoop Oracle	Gestión de la configuración Recuperación de desastres *Recovery* *Clustering* ETL
Gestor de proyectos	Oracle *Technical Project* Managers *Business intelligence*	Oracle (BI EE, EBS R12) Netezza, Business Objects Hyperion	ETL Desarrollo de software ágil PRINCE2 Gestión de *stakeholders*
Diseñador		Oracle (especialmente BIEE) SQL Netezza SQL Server MySQL	ETL *Datamodelling* Analytics CSS *Unit testing* *Data integration* *Data mining* *Business intelligence*

		UNIX	*Data warehouse* *Big data* Migración Middleware
Científico de datos		Hadoop Java NoSQL C++	Estadística Análisis Matemáticas Análisis de datos Inteligencia artificial Minería de datos

Al margen de esto, hemos de tener en cuenta que los profesionales de dirección y administración de empresa serían los ejecutores de los resultados del análisis y responsables últimos de obtener el valor que la gestión de datos hace posible.

En nuestro mercado laboral han surgido también diferentes términos para denominar los diferentes trabajos asociados a la gestión de datos masivos. Podemos ver como ejemplo la clasificación hecha por Gallego (2015):

- higienista de datos (separa el ruido de los datos irrelevantes de aquellos que son de interés; sería equivalente al curador de datos);
- explorador de datos (selecciona los datos que sirven por un proyecto concreto);
- arquitecto de soluciones de negocio (estructura los datos para que puedan ser analizados);
- científico de datos (recopila los datos y los analiza para crear modelos de predicción);
- experto en campañas (realiza acciones de marketing con los datos).

Data curator

Algunos de los perfiles antes mencionados han sido etiquetados por otros autores como *data curator* o curador de datos y *content curator* o curador de contenidos. Son traducciones calcadas del inglés que a menudo sirven para denominar a la persona que se encarga de recoger datos o información relevante sobre un tema de interés. A esta actividad se tendrían que añadir también las actividades de preservar, mantener, archivar y depositar los datos para mantenerlos seguros, intactos y accesibles para su reutilización. Este aspecto de reutilización es relevante sobre todo para las organizaciones responsables de cumplir con la ley de reutilización de la información pública.

Chief data officer (CDO)

Este perfil ha surgido en el ámbito de la gestión de los macrodatos, pero es la transformación de lo que en los últimos decenios se venía denominando *chief data information* (CIO), responsable de la gestión de la información en las organizaciones, asociado sobre todo a los sistemas informáticos.

En la misma línea, el CDO es una figura dentro de la organización asociada a un directivo

responsable de los datos de la organización y de definir la estrategia de información. En consecuencia, es el responsable de diseñar, implementar y supervisar la gobernanza de la información guardada en la organización en la que es responsable.

Científico de datos

El científico de datos es un término que está surgiendo con fuerza en los últimos años. Alcanza casi toda la cadena de valor, pero se focaliza sobre todo en el análisis de datos. Aparte del conocimiento y las competencias vinculadas al ámbito de la ingeniería informática y la gestión de la información, Davenport y Patil (2012) le añaden la competencia de la curiosidad para hacer descubrimientos a partir del gran volumen de datos.

El proyecto EDISON apuesta por el término *científico de datos* para denominar el ámbito profesional de los datos masivos. Este proyecto tiene como objetivo definir un marco de competencias que sirva, por un lado, a las universidades para definir los currículos de la oferta formativa específica para *big data* y, por otro, a los empleadores para definir el conjunto de competencias que requieren para su sector de actividad. Como hemos visto antes, el científico de datos tiene asociados cinco grupos de competencias: análisis de datos; ingeniería de la ciencia de los datos; conocimiento del ámbito de la ciencia de los datos; gestión de datos y métodos científicos o de gestión de procesos de negocio (dependiendo del ámbito en el que se aplica la investigación de los datos) (Demchenko y Belloum, 2016).

Capítulo VI

EPÍLOGO

Big data es un término que remite a un conjunto de acciones e instrumentos cuyo objetivo es extraer valor de grandes volúmenes de datos. El valor se concreta en productos, servicios, decisiones, planes, estrategias, innovación..., en definitiva, en conocimiento para la acción y en ventaja competitiva.

El término, frecuentemente utilizado en nuestra lengua en inglés, carece a fecha de hoy de una entrada en el diccionario de la Real Academia de la Lengua Española. Por ello, se están utilizando distintos términos como *macrodatos* y *grandes volúmenes de datos*. En todo caso, la terminología gira entorno a los datos.

Esto contrasta con el predominio en décadas pasadas del uso del término *información* o el de *conocimiento*, con el fin de reforzar el valor del resultado de la gestión de datos: la información y el conocimiento. El énfasis actual en los datos remarca el valor de estos como materia prima de nuevos sectores económicos, como el de la reutilización de la información. No obstante, se debe tener en cuenta que siempre se trata de gestionar datos e información para crear conocimiento. Por tanto, en cierta manera, se puede decir que no estamos ante nada nuevo.

A pesar de la novedad del término, en efecto los datos, la información y el conocimiento siempre han sido gestionados. La toma de decisiones basada en datos y en información ha sido un pilar de la gestión en las organizaciones desde hace milenios. Es la manera como la humanidad ha ido avanzando. Ante ello cabe preguntarse cuál es la novedad del concepto *big data*.

Son varios los aspectos novedosos. En primer lugar, el mismo término lo indica, el amplio volumen de datos que bases de datos tradicionales no podían procesar. En segundo lugar, la gran variedad de formatos de los datos, debido a que cada vez se generan más datos desde lugares y aparatos muy diversos (como sensores, internet, plataformas tecnológicas, por poner algunos ejemplos). En tercer lugar, la velocidad de procesamiento de datos, ya que se procesan para su análisis en tiempo real. Es decir, aunque siempre se han procesado y analizado datos, el desarrollo tecnológico facilita introducir e integrar datos que hasta hace poco era difícil o imposible integrar. Por último, como novedad de *big data*, tenemos todo el conjunto de software y hardware que se está desarrollando para cumplir las expectativas que los datos masivos han planteado, y que las bases de datos tradicionales como los *data warehouse* primero y las soluciones de *business intelligence* después o los gestores de contenidos no podían gestionar.

Hay aspectos menos novedosos que ya eran tenidos en cuenta en la gestión de datos e información tradicional, pero que ahora cobran más protagonismo, por dos razones. En primer lugar, para garantizar la calidad de datos y su fiabilidad. En segundo lugar, para asegurar la relevancia y pertinencia en la recuperación de datos e información.Para ello, la gobernanza de la información es clave, y por tanto se deben integrar las actividades tenidas en consideración en el ciclo de la información, en el del conocimiento, en el de la inteligencia competitiva y en la cadena

de valor de los datos masivos. Algunas de estas actividades son indización, catalogación y clasificación para garantizar la procedencia de los datos cuando se recuperan, y la preservación para garantizar la propia recuperación a lo largo del tiempo.

Por último, al observar los *big data*, se debe tener en cuenta la evolución que están obligados a hacer ciertos profesionales vinculados con la gestión de los datos, la información y el conocimiento. Tanto las novedades de los datos masivos como la continuación de tareas clásicas adaptadas a los requerimientos de *big data* conllevan el desarrollo de competencias y habilidades que se están denominando con distintos términos, siendo el más prevalente el de *científico de datos*. No obstante, bajo esta denominación también se pueden ver diversas especializaciones como el gestor de infraestructuras de datos, el archivero de datos, el bibliotecario digital o el curador de datos, por citar algunas.

En definitiva, las ingentes cantidades de datos que genera nuestra sociedad y su correcta gestión plantea retos en distintas disciplinas, siendo las principales la ingeniería informática, la biblioteconomía, la archivística y la gestión documental. La resolución de estos retos forma parte de la historia de estas disciplinas que aún está por venir mientras se continúen gestionando datos, información y conocimiento.

Bibliografía

Asociación Multisectorial de la Información (ASEDIE) (2016). *Informe Sector Infomediario 2016*. <http://www.asedie.es/informes.html>.

Bartrolí Muñoz, Àlex (2015). «Social Media Analytics». En: Àlex Bartrolí Muñoz; Núria Braulio Gil; Josep Curto Díaz y otros. *Nuevas tendencias tecnológicas en BI*. Barcelona: Oberta UOC Publishing.

Bass, Len; Clements, Paul; Kazman, Rick (2012). «Software Architecture in Practice». *SEI Series in Software Engineering*. Addison-Wesley.

Chen, Min; Mao, Shiwen; Liu, Yunhao (2014). «Big Data: A Survey». *Mobile New Application* (núm. 19, págs. 171-209).

Davenport, Thomas H.; Patil, D.J. (2012). «Data Scientist: The Sexiest Job of the 21st Century». *Harvard Business Review* (octubre). <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>.

Demchenko, Yuri; Belloum, Adam (2016). *EDISON Discussion Document. Data Science Competence Framework (CF-DS): Approach and First Working Version.* <http://edison-project.eu/sites/edison-project.eu/files/attached_files/node-29/edison-cf-ds-draft-cc-v06.pdf>.

European Commission Communication from the Commission to the European Parliament, the Council, the European Economics and Social Committee and the Committee of the Regions (2014). *Towards a thriving data-driven economy*. Bruselas: COM(214) 442 final. <http://europa.eu/rapid/press-release_MEMO-14-455_en.htm>.

Gallego, José Antonio (2015). *Claves para poder formar un buen equipo de big data.* <http://www.centrodeinnovacionbbva.com/noticias/claves-para-poder-formar-un-buen-equipo-de-big-data>.

Gironés Roig, Jordi (2013a). *Data mining*. Barcelona: Fundació Universitat Oberta de Catalunya.

Gironés Roig, Jordi (2013b). *Business Analytics*. Barcelona: Fundació Universitat Oberta de Catalunya.

Gómez García, José Luís; Conesa i Caralt, Jordi (2015). *Introducción al big data*. Barcelona: Oberta UOC Publishing.

Hu, Han; Wen, Yonggang; Chua, Tag-Seng y otros (2014). «Toward Scalable Systems for Big Data Analytics». *IEEE Access* (núm. 2, págs. 652-687).

Kune, Raghavendra; Konugurthi, Pramod Kumar; Agarwal, Arun y otros (2016). «The anatomy of big data computing». *Software practice and experience* (núm. 46, págs. 79-105).

Liu, Chang; Yang, Chi; Zhang, Xuyun y otros (2015). «External integrity verification for outsourced big data in cloud and IoT: A big picture». *Future Generation Computer Systems* (núm. 49, págs. 58-67).

Manyika, James; Chui, Michael; Brown, Brady otros (2011). *Big data: the next frontier for innovation, competition and productivity*. McKinsey Global Institute.

Mayer-Schönberger, Viktor; Cukier, Kenneth (2013). *Big data: la revolución de los datos masivos*. Madrid: Turner.

Padgavankar, M.H.; Gupta, S.R. (2014). «Big Data Storage and Challenges». *International Journal of Computer Science and Information Technologies* (vol. 5, núm. 2, págs. 2218-2223).

Rius Gavídia, Àngels; Serra Vizern, Montserrat; Curto Díaz, Josep (2013). «Introducción al almacenamiento de datos» En: Àngels Rius Gavídia; Montserrat Serra Vizern; Albert Abelló Gamazo y otros. *Data warehouse*. Barcelona: Oberta UOC Publishing.

Rodríguez, José Ramón (2012). «Características de los proyectos de inteligencia de negocio». En: Pere Mariné Jové;José Ramón Rodríguez; Isabel Guitart Hormigo. *Mecanismos de apoyo a la gestión de proyectos de business intelligence*. Barcelona: Fundació Universitat Oberta de Catalunya.

Santanach Casals, Daniel (2013). «Gestión de proyectos tecnológicos». En: Carlota Bustelo Ruesta; Daniel Santanach Casals. *Selección de herramientas e implementación*. Barcelona: Eureca Media.

SAS (2013a). *Big Data Analytics. An assessment of demand for labour and skills, 2012-2017*. e-skills UK.

<http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=6243>.

SAS (2013b). *Big Data Analytics. Adoption and Employment Trends, 2012-2017*. e-skillsUK. <http://www.sas.com/offices/europe/uk/downloads/bigdata/eskills/eskills.pdf>.

Tsai, Chun-Wei; Lai, Ching-Feng; Chao, Hang-Chieh y otros (2015). «Big Data analytics: a survey». *Journal of Big Data* (vol. 2, núm. 21, págs. 1-32).

Wartena, Christian; García-Alsina, Montserrat (2015). «Keyword Extraction from Company Websites for the Development of Regional Knowledge maps». En:A. Fred; J. L. G. Dietz; K. Liu y otros (eds.) (2015). *Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 5th International Joint Conference, IC3K 2013, Vilamoura, Portugal (19-22 de septiembrede 2013). Revised Selected Papers. Series: Communications in Computer and Information Science (vol. 454, págs. 96-111). Berlín / Heidelberg: Springer-Verlag.

Notas

[1] http://hadoop.apache.org/

[2] Para conocer más sobre la licencia Apache: http://www.apache.org/licenses/.

[3] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[4] https://hadoop.apache.org/docs/stable/hadoop-mapreduce-clien t/hadoop-mapreduce-client-core/MapReduceTutorial.html

[5] http://pig.apache.org/

[6] http://hbase.apache.org/

[7] http://hive.apache.org/

[8] http://sqoop.apache.org/

[9] https://flume.apache.org/

[10] http://zookeeper.apache.org/

Generamos inmensas cantidades de datos cuya gestión, de forma adecuada, genera valor y ventaja competitiva. Las actividades y tecnologías para gestionar este gran volumen de datos se denominan *big data*, y sus profesionales *científicos de datos*, aunque procedan de distintas disciplinas (ingeniería informática, biblioteconomía, archivística o gestión documental).

Este libro presenta los conceptos fundamentales y detalla cómo proceder en la gestión y creación de valor. Para ello, parte de la gobernanza de la información e integra las actividades de cuatro ciclos de gestión: el de la información, el del conocimiento, el de la inteligencia competitiva y el de los grandes volúmenes de datos.

Montserrat García-Alsina
Doctora en Sociedad de la Información y el Conocimiento, profesora de la Universitat Oberta de Catalunya e investigadora del grupo de investigación KIMO (Knowledge and Information Management in Organizations) de la misma universidad.

EDITORIAL UOC