CAPÍTULO 5
BIG DATA Y OPEN DATA: EL PODER DE LOS DATOS "GRANDES Y ABIERTOS"

El año 2016 fue el despliegue universal de Big Data y 2017 se espera sea el de la llegada comercial de la tendencia, marcos de trabajo y herramientas de analítica de datos a un gran número de organizaciones y empresas. El término ha sido ya considerado por grandes pensadores, economistas, políticos... como «el nuevo petróleo»

Big Data —grandes datos, grandes volúmenes de datos o *macrodatos*, como recomienda utilizar la Fundación Fundéu— supone la confluencia de una multitud de tendencias tecnológicas que venían madurando desde la primera década del siglo XXI, que se consolidan durante los años 2011 a 2013 cuando irrumpieron con gran fuerza en organizaciones y empresas en particular, y en la sociedad en general: movilidad, redes sociales, aumento de la banda ancha y reducción del costo de la conexión a Internet, medios sociales (en particular las redes sociales), Internet de las cosas, geolocalización, y de modo muy significativo la computación en la nube (*cloud computing*).

Los grandes datos o grandes volúmenes de datos han ido creciendo de modo espectacular. Durante 2011 se crearon 1.8 zettabytes de datos (1 billón de gigabytes) según la consultora IDC, y esta cifra se duplica cada dos años. Un dato significativo: Walmart, la gran cadena de almacenes de los Estados Unidos, ya hace unos años poseía bases de datos con una capacidad de 10 petabytes, y procesaba más de un millón de transacciones cada hora. Los Big Data están brotando por todas partes y su uso adecuado proporcionará una gran ventaja competitiva a las

organizaciones y empresas; en cambio, su ignorancia creará grandes riesgos en ellas y no las hará competitivas. Para ser competitivas en el siglo actual, como señala Franks (2012): "Es imperativo que las organizaciones persigan agresivamente la captura y análisis de estas nuevas fuentes de datos para alcanzar los conocimientos y oportunidades que ellas ofrecen".

Los profesionales de desarrollo de Big Data y de análisis de datos tienen mucho trabajo por delante y serán las profesiones relacionadas con estos temas las más demandadas en los años sucesivos.

En este capítulo introduciremos al lector en el concepto de Big Data, así como en las diferentes formas en que una organización puede hacer uso de ellos para obtener un mayor rendimiento en su toma de decisiones. No sólo se hará hincapié en el concepto y sus definiciones más aceptadas, sino que estudiaremos las oportunidades que trae consigo su adopción, y los riesgos de su no adopción, dado el gran cambio social que se prevé producirá el enorme volumen de datos que se irán creando y difundiendo.

Hoy día los datos proceden de numerosas fuentes, desde los que proceden de videojuegos hasta las innumerables cantidades de datos de operaciones en los grandes almacenes, en los bancos, la administración pública, los sensores, los teléfonos inteligentes, etc. Todos estos datos procedentes de fuentes tradicionales han ido constituyendo los grandes volúmenes de datos, y crecen de modo exponencial; así, las bases de datos de organizaciones y empresas han ido creciendo y pasando de volúmenes de datos de terabytes a petabytes.

Sin embargo, son los *datos de la Web* los que hoy día configuran el trozo más grande del "pastel" que es Big Data, ya que probablemente es la fuente de datos más utilizada y reconocida en la actualidad, y aún lo seguirá siendo en las próximas décadas. Pero, hay muchas otras fuentes que añaden ingentes cantidades de datos, que aumentan día con día las grandes cantidades de volúmenes de datos. Algunos de los orígenes más usuales de éstos son:

- Datos de la Web.
- Datos de los medios sociales (redes sociales, blogs, wikis).
- Datos de Internet de las cosas.
- Datos de interconexión entre máquinas, M2M (Internet de las cosas).
- Datos industriales de organizaciones y empresas, procedentes de múltiples sectores.
- Datos de la industria del automóvil.
- Datos de redes de telecomunicaciones.

- Datos de medios de comunicación (prensa, radio, televisión, cine…).
- Datos procedentes de sensores en los más diferentes campos de la industria y la agricultura.
- Datos de videojuegos en locales recreativos, casinos, lugares de ocio…
- Datos procedentes de posiciones geográficas y de telemetría: geolocalización.
- Datos procedentes de chips NFC, RFID, códigos QR y Bidi, en aplicaciones de comercio electrónico.
- Datos procedentes de servicios de telefonía móvil (celular) inteligente: texto, datos, audio, video, fotografía.
- Datos procedentes de redes inteligentes (*smart grids)*.
- Datos personales, datos de texto….
- Otros.

Una tendencia clara que se observa a diario es que las tecnologías fundamentales, que contienen y transportan datos, conducen a múltiples fuentes de grandes datos en las industrias más diferentes. Y a la inversa, diferentes industrias pueden aprovecharse de numerosas fuentes de datos.

DEFINICIÓN DE BIG DATA

No existe unanimidad en la definición de Big Data, aunque sí hay cierto consenso en la fuerza disruptiva que suponen los grandes volúmenes de datos y la necesidad de su captura, almacenamiento y análisis. Son numerosos los artículos (*white papers*), informes y estudios relativos al tema de Big Data en los últimos años, y en este libro seleccionamos las definiciones de instituciones relevantes y con mayor impacto mediático y profesional. En general, existen diferentes aspectos en los que casi todas las definiciones están de acuerdo y con conceptos consistentes para capturar la esencia de Big Data: crecimiento exponencial de la creación de grandes volúmenes de datos, origen o fuentes de datos y la necesidad de su captura, almacenamiento y análisis para conseguir el mayor beneficio para organizaciones y empresas, junto con las oportunidades que ofrecen y los riesgos que conlleva su no adopción.

La primera definición que daremos es la de Adrian Merv, vicepresidente de la consultora Gartner, quien en la revista *Teradata Magazine*, del primer trimestre de 2011, definió este término como: “Big Data excede el alcance de los entornos de *hardware* de uso común y herramientas de *software* para capturar, gestionar y

procesar los datos dentro de un tiempo transcurrido tolerable para su población de usuarios"[1].

Otra definición muy significativa es la del McKinsey Global Institute [2], que, en un informe muy reconocido y referenciado, de mayo de 2011, definió el término así: "Big Data se refiere a los conjuntos de datos cuyo tamaño está más allá de las capacidades de las herramientas comunes de software de bases de datos para capturar, almacenar, gestionar y analizar". Esta definición es, según McKinsey, intencionadamente subjetiva e incorpora una definición cambiante, "en movimiento", qué "de grande" necesita ser un conjunto de datos para ser considerado Big Data; es decir, no se lo define en términos de ser mayor que un número dado de terabytes (en cualquier forma, es frecuente asociar el término Big Data a terabytes y petabytes). Suponemos, dice McKinsey, que a medida que la tecnología avanza en el tiempo, el tamaño de los conjuntos de datos que se definen con esta expresión también crecerá. De igual modo, McKinsey destaca que la definición puede variar para cada sector, dependiendo de cuáles sean los tipos de herramientas de software normalmente disponibles, y cuáles los tamaños comunes de los conjuntos de datos en ese sector o industria. Teniendo presente estas consideraciones, los Big Data en muchos sectores hoy día variarán desde decenas de terabytes a petabytes y ya casi exabytes, camino de zettabytes.

Otra fuente de referencia es la consultora tecnológica IDC[3], que apoyándose en sus propios estudios, considera que: "Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad, con el objeto de extraer valor económico de ellos".

La empresa multinacional de auditoría Deloitte lo define como: "El término que se aplica a conjuntos de datos cuyo volumen supera la capacidad de las herramientas informáticas (computación) de uso común, para capturar, gestionar y procesar datos en un lapso de tiempo razonable. Los volúmenes de Big Data varían constantemente, y en la actualidad oscilan entre algunas decenas de terabytes hasta muchos petabytes para un conjunto de datos individual" [4].

Otra definición muy acreditada por venir de la mano de la consultora Gartner es: "Big Data son los grandes conjuntos de datos que tiene tres características principales: volumen (cantidad), velocidad (velocidad de creación y utilización) y variedad (tipos de fuentes de datos no estructurados, tales como la interacción social, video, audio, cualquier cosa que se pueda clasificar en una base de datos)"[5]. Estos factores, naturalmente, conducen a una complejidad extra de los Big Data. En síntesis "'Big Data' es un conjunto de datos tan grandes como diversos que rompen las infraestructuras de TI tradicionales"[6].

Gartner considera que la esencia importante de Big Data no es tanto el tema numérico, sino todo lo que se puede hacer si se aprovecha el potencial y se descubren nuevas oportunidades de los grandes volúmenes de datos.

En suma, la definición de Big Data puede variar según las características de las empresas. Para unas empresas prima el *volumen*; para otras, la *velocidad*; para otras, la *variabilidad* de las fuentes. Las empresas con mucho volumen o *volumetría* van a estar interesadas en capturar la información, guardarla, actualizarla e incorporarla en sus procesos de negocio; pero hay empresas que, aunque tengan mucho volumen, no necesitan almacenar, sino trabajar en tiempo real y a gran velocidad. Otras, por el contrario, pueden estar interesadas en gestionar diferentes tipos de datos.

Un ejemplo clásico son los sistemas de recomendación: sistemas que en tiempo real capturan información de lo que está haciendo el usuario en la Web, lo combina con la información histórica de ventas, lanzando en tiempo real las recomendaciones. Otras empresas tienen otro tipo de retos como fuentes heterogéneas, y lo que necesitan es combinarlas. La captura es más compleja, ya que hay que combinar en un mismo sitio y analizarla.

TIPOS DE DATOS

Los Big Data son diferentes de las fuentes de datos tradicionales que almacenan datos estructurados en las bases de datos relacionales. Es frecuente dividir las categorías de datos en dos grandes tipos: *estructurados* (datos tradicionales) y *no estructurados* o *sin estructura* (datos Big Data). Sin embargo, las nuevas herramientas de manipulación de Big Data han originado unas nuevas categorías dentro de los tipos de datos no estructurados: *datos semiestructurados* y *datos no estructurados* propiamente dichos.

DATOS ESTRUCTURADOS

La mayoría de las fuentes de datos tradicionales son datos estructurados, datos con formato o esquema fijo que poseen campos fijos. En estas fuentes, los datos vienen en un formato bien definido que se especifica en detalle, y que conforma las bases de datos relacionales. Son, fundamentalmente, los datos de las bases de datos relacionales, las hojas de cálculo y los archivos,

Los datos comunes almacenados en bases de datos, registrados en campos con un nombre específico y con unas relaciones entre ellos, se almacenan en filas y columnas y son fáciles de introducir, almacenar y analizar. Proporcionan la mayor parte de la información actual de la empresa, como datos de los sistemas de registro, transacciones comerciales, censos de población, ventas, clientes, finanzas, etc. Estos tipos de datos se localizan en un campo fijo de un registro o archivo específico y sus contenidos se incluyen en bases de datos relacionales, en hojas de cálculo, normalmente. Los datos se organizan en torno a un modelo de datos.

Un modelo de datos —ciclo de vida del dato o cadena de valor del dato— contiene: los tipos de datos empresariales que su empresa va a registrar, el modo de almacenamiento, el proceso y el modo de acceso a dichos datos. Los datos estructurados normalmente se almacenan en bases de datos relacionales y hojas de cálculo, en filas y columnas, con los campos explicitados en ellas. Así, los campos de datos de una base de datos estándar de clientes de una empresa, incluirán nombres, dirección, número de teléfono de contacto, dirección de correo electrónico, etc., o en el caso de ser la base de datos de empleados, incluirá también edad, profesión, etcétera.

Los campos deberán ser definidos con el tipo de datos que va a contener: datos numéricos o de texto, con indicación expresa de su tipo de información; por ejemplo, el campo dirección ha de ser tipo texto, y el campo número de teléfono, tipo numérico (o también texto si se desea admitir el signo + como código de salida internacional en lugar del clásico 00 que también admiten las operadoras de telefonía). También se pueden adoptar otras convenciones como incluir menús desplegables que limitan las opciones de los datos que se pueden introducir en un campo y asegurar coherencia de entrada:

Titulación	Ciudad
Sr.	Madrid
Sra.	Granada
Dr.	Medellín
Dra.	Ciudad de México
Ing.	Lima
Licenciado (Graduado)	Santo Domingo

Los datos estructurados se componen de piezas de información que se conocen de antemano, vienen en un formato especificado y se producen en un orden, también, especificado. Estos formatos facilitan el trabajo con dichos datos. Formatos comunes son: fecha de nacimiento (DD, MM, AA); documento nacional de identidad o pasaporte (por ejemplo, 8 dígitos y una letra); número de la cuenta corriente en un banco (20 dígitos), etcétera.

La gestión y búsqueda de los datos estructurados en las bases de datos relacionales se realizan con el lenguaje de programación estándar SQL —lenguaje creado por IBM en la década de los setenta— y que todavía sigue en vigor y soporta a la mayoría de las bases de datos establecidas en organizaciones y empresas.

Sin embargo, las bases de datos relacionales tienen un gran inconveniente en la era de los grandes volúmenes de datos: la escasa facilidad que tiene para manejar datos no estructurados.

Datos con formato o esquema fijo que poseen campos fijos. Son los datos de las bases de datos relacionales, las hojas de cálculo y los archivos, fundamentalmente.

DATOS NO ESTRUCTURADOS

Los **datos no estructurados (sin estructurar)** son datos sin tipos predefinidos. Se almacenan como "documentos" u "objetos" sin estructura uniforme, y se tiene poco o ningún control sobre ellos. Tienen un formato que no puede ser gestionado (indexado) fácilmente en tablas de bases de datos relacionados. Datos no estructurados son:

- Video, audio, imágenes, fotografías
- Datos de texto (archivos de texto o documentos, tales como Word, PowerPoint, PDF...)
- Documentos multimedia
- Correos electrónicos
- Textos de mensajería (SMS, mensajes de WhatsApp, Line, Telegram, Viber, Snapchat...)
- Publicaciones en redes sociales

Por ejemplo, las imágenes se clasifican por su resolución en pixeles. Ejemplos comunes de datos que no tienen campos fijos son: audio, video, fotografías, documentos impresos, cartas, hojas electrónicas, imágenes digitales, formularios especiales, mensajes de correo electrónico y de texto, formatos de texto libre como correos electrónicos, mensajes instantáneos SMS, artículos, libros, mensajes de mensajería instantánea tipo WhatsApp, Line, Hangouts, Telegram, Snapchat, Skype, Messenger Facebook, Google Allo, WeChat... Al menos, 80% de la información de las organizaciones no reside en las bases de datos relacionales o archivos de datos, sino que se encuentran esparcidos a lo largo y ancho de la organización; todos estos datos se conocen como datos no estructurados.

Sin duda, los datos más difíciles de dominar por los analistas son los datos no estructurados, pero su continuo crecimiento ha provocado el nacimiento de herramientas para su manipulación como es el caso de MapReduce, Hadoop o bases de datos NoSQL.

Ejemplos comunes de datos que no tienen campos fijos: audio, video, fotografías, o formatos de texto libre como correos electrónicos, mensajes instantáneos SMS, artículos, libros, mensajes de mensajería instantánea tipo WhatsApp, Snapchat, Viber...

DATOS SEMIESTRUCTURADOS

Los datos semiestructurados tienen propiedades de datos estructurados y no estructurados, y pueden tener algún tipo específico de estructura que se puede utilizar en un análisis de datos, pero no contienen la estructura de un modelo de datos. Asimismo, poseen un flujo lógico y un formato que puede ser definido, pero no es fácil su comprensión por el usuario. Son datos que no tienen formatos fijos, pero sí etiquetas y otros marcadores que permiten separar los elementos dato. Ejemplos de datos semiestructurados son:

- Documentos XML de páginas web
- Contenidos de blogs y redes sociales
- *Software* de tratamiento de textos
- Lenguajes de marca de hipertexto extensibles

La lectura de datos semiestructurados requiere el uso de reglas complejas que determinan cómo proceder después de la lectura de cada pieza de información. Ejemplos típicos de datos semiestructurados son:

- Los registros *Web log*s de las conexiones a Internet. Un *Web log* se compone de diferentes piezas de información, cada una de las cuales sirve para un propósito específico. Ejemplos comunes son el texto de etiquetas de lenguajes XML y HTML.
- El software de tratamiento de textos que incluyen metadatos que pueden contener nombre del autor, ISBN del libro, fecha de edición, fecha de compra, etc.; sin embargo, su contenido está sin estructurar.
- Publicaciones, entradas de Facebook o LinkedIn, que se pueden clasificar por autor, información, longitud de texto, opiniones de seguidores, etc., pero su contenido normalmente no está estructurado.

Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos dato. Ejemplos típicos son el texto de etiquetas de XML y HTML.

CARACTERÍSTICAS DE BIG DATA

Cada día creamos 2.5 *quintillones* de bytes de datos, de manera que 90% de los datos del mundo actual se han creado en los últimos dos años[7]. Estos datos proceden de todos los sitios: sensores utilizados para recoger información del clima, entradas (*posts*) en sitios de medios sociales, imágenes digitales, fotografías y videos, registros de transacciones comerciales y señales GPS de teléfonos celulares, por citar unas pocas referencias. Estos datos son, según IBM, Big Data.

Big Data, al igual que la nube (*cloud*), abarca diversas tecnologías. Los datos de entrada a los sistemas de Big Data pueden proceder de redes sociales, *log*s, registros de servidores Web, sensores de flujos de tráfico, imágenes de satélites, flujos de audio y de radio, transacciones bancarias, MP3 de música, contenido de páginas Web, escaneado de documentos de la administración, caminos o rutas GPS, telemetría de automóviles, datos de mercados financieros. ¿Todos estos datos son realmente los mismos?

En 2001, Dougn Laney —analista de META Group, hoy Gartner[8]— definía el crecimiento constante de datos como una oportunidad y un reto para investigar el volumen, velocidad y variedad. Posteriormente, Mark Beyer[9], vicepresidente de Gartner, presentó un informe sobre la emergencia de Big Data y sus características principales: volumen, velocidad y variedad.

IBM planteó —como también hizo Gartner— que Big Data abarca tres grandes dimensiones, conocidas como el "Modelo de las tres V" (3 V o V^3): *volumen*, *velocidad* y *variedad* (*variety*). Existe un gran número de puntos de vista para visualizar y comprender la naturaleza de los datos y las plataformas de software disponibles para su explotación; la mayoría incluirá una de estas tres propiedades V en mayor o menor medida. Sin embargo, algunas fuentes, como es el caso de IBM, cuando tratan las características de los Big Data también consideran una cuarta característica que es la *veracidad*, y que analizaremos también para dar un enfoque más global de la definición y características de los Big Data. Otras fuentes notables añaden una quinta característica, *valor*, y llegan añadir hasta 7 u 8V, como veremos más adelante.

VOLUMEN

Las empresas amasan grandes volúmenes de datos, desde terabytes hasta petabytes. Las cantidades que hoy nos parecen enormes, en pocos años serán normales. Estamos pasando de la era del petabyte a la era del exabyte, y para el periodo 2015 a 2020 se espera que entremos en la era del zettabyte. IBM da el dato de 12 terabytes para referirse a lo que crea Twitter cada día sólo en el análisis de productos para conseguir mejoras en la eficacia.

En el año 2000 se almacenaron en el mundo 800.000 petabytes. Se espera que en el año 2020 se alcancen los 40 zettabytes (ZB). Solo Twitter genera más de 9 terabytes (TB) de datos cada día. Facebook, 10 TB, y algunas empresas generan terabytes de datos cada hora de cada día del año. Las organizaciones se enfrentan a volúmenes masivos de datos, y las que no conocen cómo gestionar estos datos están abrumadas por ello. Sin embargo, la tecnología existe, con la plataforma tecnológica adecuada para analizar casi todos los datos (o al menos la mayoría de ellos, mediante la identificación idónea) con el objetivo de conseguir una mejor comprensión de sus negocios, sus clientes y el *marketplace*. IBM plantea que el

volumen de datos disponible en las organizaciones hoy día está en ascenso, mientras que el porcentaje de datos que se analiza está en disminución.

La característica volumen es la más popular y reconocida dentro del término Big Data, aunque al día de hoy no es la más significativa. Se necesita almacenar la información y para ello se requiere una base de datos capaz de almacenar y gestionar las enormes cantidades de datos. Los tamaños de los archivos son muy grandes y cada segundo/minuto se generan grandes volúmenes de datos que crecen de modo exponencial, en la expresión que IDC denomina "El universo digital de datos". El volumen de datos es una de las propiedades más destacadas en cualquier definición de Big Data, pero existen otras propiedades de igual o mayor importancia en la actualidad.

VELOCIDAD

La importancia de la velocidad de los datos o el aumento creciente de los flujos de datos en las organizaciones, junto con la frecuencia de las actualizaciones de las grandes bases de datos, son características importantes a tener en cuenta. Esto requiere que su procesamiento y posterior análisis, normalmente ha de hacerse en tiempo real para mejorar la toma de decisiones sobre la base de la información generada. A veces, cinco minutos es demasiado tarde en la toma de decisiones; los procesos sensibles al tiempo —como pueden ser los casos de fraude— obligan a actuar rápidamente. Imaginemos los millones de escrutinios de los datos de un banco con el objetivo de detectar un fraude potencial o el análisis de millones de llamadas telefónicas para tratar de predecir el comportamiento de los clientes y evitar que se cambien de compañía.

La importancia de la velocidad de los datos se une a las características de volumen y variedad, de modo que la idea de velocidad no se asocia a la tarea de crecimiento de los depósitos o almacenes de datos, sino que se aplica la definición al concepto de los datos en movimiento, es decir, la velocidad a la cual fluyen los datos. Las empresas están tratando cada día con mayor intensidad, petabytes de datos en lugar de terabytes, y el incremento en fuentes de todo tipo como sensores, chips RFID, chips NFC, datos de geolocalización y otros flujos de información que conducen a flujos continuos de datos, imposibles de manipular por sistemas tradicionales.

VARIEDAD

Las fuentes de datos son de cualquier tipo. Los datos pueden ser estructurados y no estructurados (texto, datos de sensores, audio, video, flujos de clics, archivos *log*s), y cuando se analizan juntos se requieren nuevas técnicas. Imaginemos el registro en vivo de imágenes de las cámaras de video de un estadio de fútbol, o las de vigilancia de calles y edificios.

En los sistemas de Big Data las fuentes de datos son diversas y no suelen ser estructuras relacionales comunes. Los datos de imágenes de redes sociales pueden venir de una fuente de sensores y no suelen estar preparados para su integración en una aplicación.

En el caso de la Web, la realidad de los datos es confusa. Distintos navegadores envían datos diferentes; los usuarios pueden ocultar información y usar diversas versiones de software, ya sea para comunicarse entre ellos, realizar compras o para leer un periódico digital. No obstante, los riesgos que conlleva la no adopción de las tendencias de Big Data son grandes, ya que:

- La voluminosa cantidad de información puede llevar a una confusión que impida ver las oportunidades y amenazas dentro de nuestro negocio y fuera de él, y perder así competitividad.
- La velocidad y flujo constante de datos en tiempo real puede afectar a las ventas y a la atención al cliente.
- La variedad y complejidad de datos y fuentes puede llevar a la vulneración de determinadas normativas de seguridad y privacidad de datos.

El volumen asociado con los Big Data conduce a nuevos retos para los centros de datos que intentan lidiar con su variedad. Con la explosión de sensores y dispositivos inteligentes, así como las tecnologías de colaboración sociales, los datos en la empresa se han convertido en muy complejos, ya que no sólo incluyen los datos relacionales tradicionales, sino también priman en bruto datos semiestructurados y no estructurados procedentes de páginas Web, archivos de registros Web (*Web log*), incluyendo datos de los flujos de clics, índices de búsqueda, foros de medios sociales, correo electrónico, documentos, datos de sensores de sistemas activos y pasivos, entre otros.

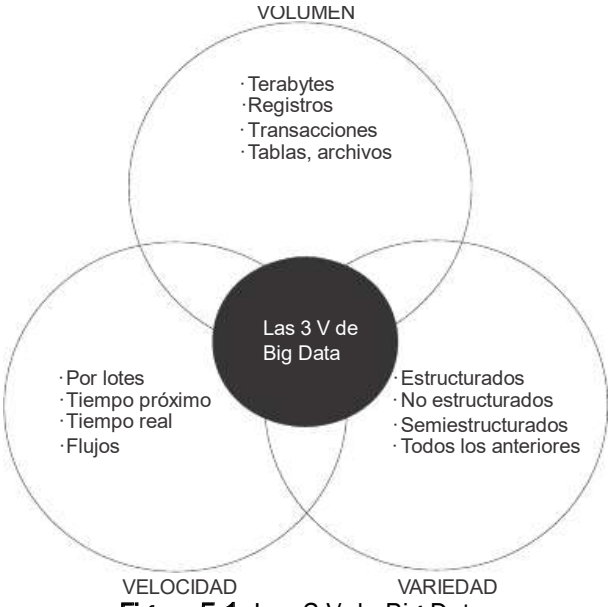

Figura 5.1. Las 3 V de Big Data.
Fuente: Philip Russom: "Big Data Analytics", en Teradata, Fourth Quarter 2011. Disponible en: <http://tdwi.org/blogs/philip-russom>.

Dicho en forma sencilla, la *variedad* representa todos los tipos de datos y supone un desplazamiento fundamental en el análisis de requisitos, desde los datos estructurados tradicionales hasta la inclusión de los datos en bruto, semiestructurados y no estructurados, como parte del proceso fundamental de la toma de decisiones. Las plataformas de analítica tradicionales no pueden manejar esta variedad. Sin embargo, el éxito de una organización dependerá de su capacidad para resaltar el conocimiento de los diferentes tipos de datos disponibles en ella, que incluirá tanto los datos tradicionales como los no tradicionales[10]. Por citar unos ejemplos, el video y las imágenes no se almacenan fácil ni eficazmente en una base de datos relacional, y mucha información de sucesos de la vida diaria como los datos climáticos, cambian dinámicamente. Por todas estas razones, las empresas deben capitalizar las oportunidades de los grandes datos, y tener la capacidad de analizar todos los tipos de datos, tanto relacionales como no relacionales: texto, datos de sensores, audio, video, transaccionales.

EL MODELO DE LAS 5 V

IBM añadió una cuarta V y posteriormente una quinta V[11]. De igual forma, Bernard Marr, uno de los grandes expertos mundiales en Big Data, publicó el artículo: "*Big*

data: the 5 vs everyone must know"[12] donde añade dos nuevas propiedades: **Veracidad** y **Valor**, que luego incluiría en su libro de Big Data, publicado en 2015.

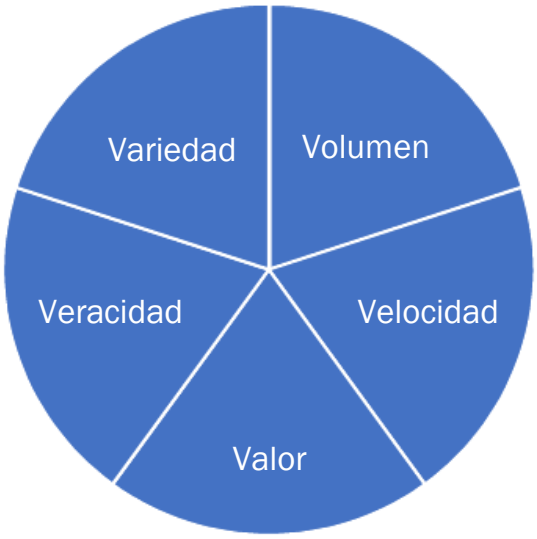

Figura 5.2. Las 5 V de Big Data.

VERACIDAD

En su definición de Big Data, al comentar la característica de veracidad, IBM proporciona un dato estremecedor: "Uno de cada tres líderes de negocio (directivos) no se fía de las informaciones que utilizan para tomar decisiones (lo denomina incertidumbre)." ¿Cómo puede, entonces, actuar con esta información si no se fía de ella? El establecimiento de la veracidad o fiabilidad (*truth*) de Big Data supone un gran reto a medida que la variedad y las fuentes de datos crecen.

El ya citado gurú de Big Data, Bernard Marr, publicó un artículo en LinkedIn y posteriormente en su libro donde insiste en esta característica, en que considera que se refiere al desorden o confiabilidad en muchos tipos de Big Data cuya calidad y precisión son menos controlables (citaba el caso de los mensajes de *Twitter* con sus etiquetas "hash", abreviaturas, tipos y leguaje coloquial, así como la fiabilidad y precisión del contenido). Pese a ello, Marr considera que las tecnologías y la analítica de Big Data permiten trabajar con estos tipos de datos, los cuales compensan, en ocasiones, la falta de calidad y precisión.

VALOR

Existe una quinta característica que también se suele considerar y es muy importante: el valor. Las organizaciones estudian obtener información de los grandes datos de una manera rentable y eficiente. Aquí es donde las tecnologías de código abierto tales como Apache Hadoop se han vuelto muy populares. Hadoop es un software que procesa grandes volúmenes de datos a través de un *cluster* de centenares o incluso millares de computadores de un modo muy económico.

De hecho, Marr insiste: es muy importante tener acceso a Big Data, pero a menos que se convierta en valor, será inútil. Es preciso asumir costos y beneficios y el valor será una característica vital.

IBM también ha considerado su quinta V de Valor. Considera que la capacidad de conseguir mayor valor a través del conocimiento (*insights*) de la analítica le proporciona una gran importancia a esta propiedad.

EL MODELO DE LAS 7 V

A las cinco características anteriores se están uniendo según algunos modelos de Big Data: las nuevas e importantes de *visualización* y *viabilidad*.

VISUALIZACIÓN

Es el modo en que los datos se presentan para encontrar patrones y claves que permitan la obtención de resultados para una toma de decisión eficiente. Las iniciativas de Big Data requieren herramientas de visualización de datos óptimas. Estas herramientas permiten a los usuarios finales realizar búsquedas y acceder a la información rápidamente y en muchos casos en tiempo real. Es una gran ventaja para los clientes quienes se muestran satisfechos cuando tienen el control de la información en el mismo momento que ésta se produce. La visualización es una parte muy importante ya que ayuda a las organizaciones a responder a preguntas de interés para el desarrollo del negocio.

VIABILIDAD

Esta propiedad se refiere a la capacidad que tienen las empresas de generar un uso eficaz del gran volumen de datos que manejan. Esta característica también adopta la forma de la variabilidad para referirse también a que el gran volumen de datos está cambiando constantemente (este es el caso de los asistentes virtuales como Siri o el computador cognitivo Watson que reúnen datos para el procesamiento del lenguaje).

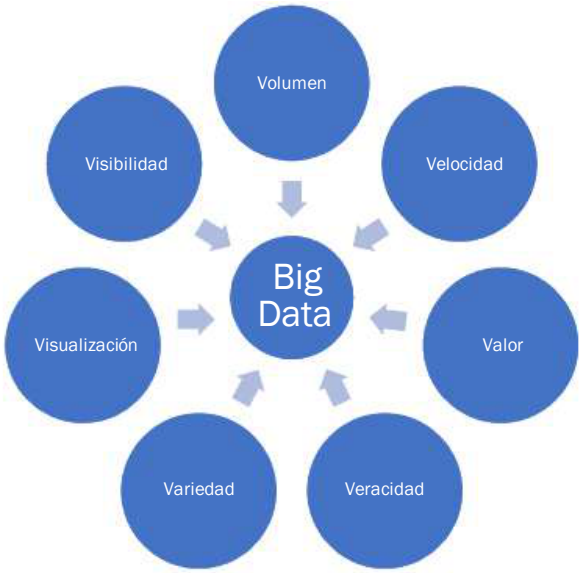

Figura 5.3. Modelo de Big Data de las 7 v

EL TAMAÑO DE LOS BIG DATA

La megatendencia de los Big Data no está directamente relacionada con la cantidad específica de datos. Recordemos que hace una década los almacenes de datos (*data warehouse*) de las grandes empresas, cuando tenían de 1 a 10 terabytes se consideraban enormes. Hoy se puede comprar en cualquier gran almacén, unidades de disco de 1 a 5 terabytes por precios inferiores a 100 euros (Soares, 2012), y muchos almacenes de datos de empresas han roto la barrera del petabyte. En la feria CES 2017 de Las Vegas, se han presentado modelos de pendrives de 2 terabytes, con lo que la reducción de tamaño se ha hecho muy considerable para los dispositivos de almacenamiento.

En este contexto, la pregunta lógica es: ¿cuál es la parte más importante de Big Data, la parte *big* o la parte *data*? O de manera más específica: ¿ambas partes? o ¿ninguna? Para muchos expertos, el tema de debate es cuánto supone *big* (grandes volúmenes) dado que el tema *data* es el soporte fundamental de la tendencia.

Recordemos que según IDC, el universo digital de datos se dobla cada dos años, y que más del 70% de los datos creados se generarán por los consumidores, y por encima del 20% por las empresas. IDC13 predice que el universo digital se multiplicará por un factor de 44 para llegar a 40 zettabytes en 2020.

Tal vez una respuesta más ajustada a la situación actual es que ni la parte big ni la parte data son los aspectos más importantes de Big Data. Es necesario resaltar lo que hacen las organizaciones con los grandes datos; es lo más

importante. El análisis de los grandes datos que realice su organización combinado con las acciones que se tomen para mejorar su negocio es lo realmente importante.

En resumen, el valor de Big Data es tanto *big* como *data*, y su indicador final dependerá del análisis de los datos, cómo se realizará y cómo mejorará el negocio.

BREVE RESEÑA HISTÓRICA DE BIG DATA

La historia del término Big Data se puede dividir en dos etapas. Primero, con el nacimiento y expansión del concepto en el campo científico y de negocios, restringido su uso a su conceptualización como tal en la jerga técnica y académica; este periodo se puede datar entre 1984 y 2007. Segundo, con la difusión del término ya con criterio tecnológico y económico, que produce beneficios a las organizaciones y empresas, que comienzan a estudiar la tecnología, a desarrollar herramientas para el análisis de los grandes volúmenes de datos o aquellas otras que comienzan a utilizar estas herramientas para sacarles un rendimiento en las empresas y negocios; este periodo se puede considerar que inicia en el 2008.

El profesor Francis X. Diebold[14] —en un trabajo de investigación realizado sobre el origen e implantación del término Big Data, y publicado en el lejano noviembre de 2012 analizaba el término desde su aparición en escritos académicos y de negocios, y desde su perspectiva de economista-estadístico. Según Diebold, el uso académico del término Big Data se remonta a Tilly, en 1984, y en el lado no académico cita una primera reseña, publicada en 1987, relativa a una técnica de programación denominada *small code, big data*. En 1989, y por último en 1993, se habla de *Big Data applications*.

Por último, Diebold menciona un trabajo de Laney (2001)[15] titulado *Three V´s of Big Data (Volume, Variety and Velocity*), donde se conceptualiza el significado del término y el fenómeno de Big Data. Las conclusiones de la investigación de Diebold (él también interviene como uno de los primeros científicos, en este caso en el área de la estadística y la econometría, que utiliza el término en el año 2000) es que comienza a ser utilizado en dos grandes disciplinas: Ciencias de la Computación (Informática) y Estadística/Econometría, y que nació a mediados de los años noventa, en Silicon Graphics Inc. (SGI), en la persona de John Mashey; y posteriormente en 1998, Weiss y Indurkbya, en computación; y Diebold (2000), en estadística/econometría, y Douglas Laney (META Group, hoy Gartner). En resumen, concluye Diebold que el término se puede atribuir razonablemente a Marsey, Weiss e Indurkhya, Diebold y Laney.

EL ORIGEN MODERNO DE BIG DATA

En 2008, Steve Lohr[16], de *The New York Times*, publicó que, de acuerdo con diferentes científicos de computación y directivos de la industria, el término Big Data fue calando en ambientes tecnológicos y comenzó a generar ingresos económicos. Estamos totalmente de acuerdo con Lohr, ya que también de modo ininterrumpido he seguido los avatares de Big Data.

Pero, sin duda, es el artículo que *Wired* [17] publicó en junio del mismo año, el detonante de la explosión de los Big Data; así también lo considera Lohr.

Wired publica un artículo en que se presentaban las oportunidades e implicaciones del diluvio de datos moderno; declaraba en aquel entonces que vivíamos en la era del petabyte; sin embargo, el petabyte era una unidad de medida de datos almacenados en soportes digitales, pero ya era necesario pensar en términos de exabytes, zettabytes y yottabytes. El estudio de investigación de *Wired*, que así recogía el artículo, tenía una introducción en la que planteaba los siguientes argumentos:

> *Existen sensores en todas partes, almacenamiento infinito, nubes de procesadores. Nuestra capacidad para capturar, almacenar (Warehouse) y comprender las cantidades masivas de datos está cambiando la ciencia, la medicina, los negocios y la tecnología. A medida que crece nuestra colección de hechos y figuras, se tendrá la oportunidad de encontrar respuestas a preguntas fundamentales, debido a que la era de los big data no es sólo más: más es diferente (Because in the era of big data, more isn't just more, more is different).*

En ese mismo número, Chris Anderson[18], su director editor, publicó otro artículo en que cuestiona el hecho de que el diluvio de datos podía dejar obsoleto el método científico. En el artículo planteaba que hacía diez años, los *crawlers* de los motores de búsqueda hacían una única base de datos. Ahora Google y compañías similares están tratando el *corpus* masivo de datos como un laboratorio de la condición humana. Ellos son los hijos de la era del petabyte. La era del petabyte es diferente porque "más es diferente". Los kilobytes se almacenaban en discos flexibles; los megabytes se almacenaban en discos duros. Los terabytes se almacenaron en *arrays* de discos. Los petabytes se almacenan en la nube. A medida que nos movemos en paralelo a la progresión anterior, nos desplazamos de la analogía de las carpetas (*folders*) a la analogía de los gabinetes de archivos, y de ahí a la analogía de la biblioteca (*library),* y en la era de los petabytes a la analogía de las organizaciones en la nube.

Lohr (2012), en el artículo antes citado, considera que a finales de 2008 se produjo el espaldarazo del mundo científico, ya que los Big Data fueron adoptados por un grupo de investigadores muy reconocidos del ámbito de la computación agrupados en torno a la prestigiosa Computing Community Consortium, un grupo que colabora con el National Science Foundation (NSF) de los Estados Unidos, y la

Computing Research Association, también de los Estados Unidos, que a su vez representa a investigadores académicos y corporativos. Este consorcio publicó un influyente artículo (*white paper*): "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society"[19].

Otra noticia destacada que comenta Lohr es el hecho de que IBM en 2008 adoptó también Big Data en su marketing, especialmente después de que el término comenzara a tener gran resonancia entre sus clientes. Posteriormente en 2011, IBM introdujo en Twitter, #IBMbigdata, y en enero de 2012 publicó su primer libro electrónico sobre tecnologías de Big Data (*Understanding Big Data*).

Desde un punto de vista popular que demuestra la penetración del término, ya no sólo en los negocios, en el campo académico y en la investigación, sino en la sociedad en general y en la vida diaria, es que la tira cómica del genial Dilbert de Scott Adams recogía en sus viñetas de julio de 2012, la incorporación del Big Data. En una viñeta, Dilbert comenta: *It comes from everywhere it know all* (proviene de todas partes, lo sabe todo), para concluir: *according to the book of Wikipedia, its name is Big Data* (según el libro de Wikipedia, su nombre es 'Big Data').

Big data es el corazón de la ciencia y de los negocios modernos. Los primeros grupos de científicos centrados en sus evidencias, publicaron en agosto de 2012 un dosier especial "Big Data Special Issue" en la revista *Significance,* publicación conjunta de la American Statistical Association y la Royal Statistical Society[20].

FUENTES DE DATOS

El gran volumen de datos procede de numerosas fuentes, especialmente de las nuevas fuentes como medios sociales (*social media*) y los sensores de máquinas (máquina a máquina, M2M, e Internet de las cosas). La oportunidad de expandir el conocimiento incrustado en ellos, por combinación de esa inmensidad de datos con los datos tradicionales existentes en las organizaciones está acelerando su potencialidad; además, gracias a la nube (*cloud*), a esa enorme cantidad de información se puede acceder de modo *ubicuo*, en cualquier lugar, en cualquier momento y, prácticamente, desde cualquier dispositivo inteligente.

Los directivos y ejecutivos de las compañías se pueden volver más creativos a medida que extraen mayor rendimiento de las nuevas fuentes de datos externas y las integran con los datos internos procedentes de las bases de datos relacionales y heredadas (*legacy*) de las propias compañías. Los medios sociales están generando terabytes de información de datos no estructurados como conversaciones, fotos, video, documentos de texto de todo tipo, a los que hay que añadir los flujos de datos que fluyen de sensores, de la monitorización de procesos, fuentes externas de todo tipo, desde datos demográficos hasta catastrales, historial y predicciones de datos del tiempo climático, entre otros.

Las fuentes de datos que alimentan los Big Data no paran de crecer pero, como reconocía tempranamente el estudio de McKinsey (2011: 19)[21], citando fuentes oficiales de estadística de los Estados Unidos, numerosas empresas de todos los sectores tenían almacenadas, ya en 2009, al menos 100 terabytes, y muchas podían llegar a tener más de 1 petabyte. Algunos datos ilustrativos por sectores eran en esas fechas: fabricación discreta, 966 petabytes; banca, 619 petabytes; gobierno, 858 petabytes; comunicación y medios, 715 petabytes. Es decir, además de las nuevas fuentes de datos que comentamos en el capítulo, numerosas empresas de todo tipo tienen almacenados petabytes de datos, que se convierten en fuentes de datos tradicionales que son responsables, a su vez, de los grandes volúmenes de datos actuales.

El origen de los datos que alimentan los Big Data procederán de numerosas fuentes tanto tradicionales como nuevas, que iremos desglosando a continuación, y aunque los datos no estructurados constituirán los porcentajes más elevados que deberán gestionar las organizaciones —al menos del 80% al 90%, según los estudios que se consulten—, no podemos dejar a un lado la inmensidad de datos estructurados presentes en organizaciones y empresas, y que en numerosísimas ocasiones están aflorando datos que permanecían ocultos, y esta creciente avalancha de datos de innumerables fuentes está comenzando a tener gran fuerza y potencialidad a la hora de la toma de decisiones.

TIPOS DE FUENTES DE BIG DATA

Las fuentes de datos origen de los Big Data pueden ser clasificadas en diferentes categorías, cada una de las cuales contiene a su vez un buen número de fuentes diversas que recolectan, almacenan, procesan y analizan. IBM clasifica las fuentes de datos, según (Soares, 2012), como muestra la Figura 5.4. Esta taxonomía de fuentes de datos es una de las más referenciadas en la década actual, como las categorías globales de fuentes de datos manejadas por Big Data, pese a que han ido surgiendo nuevas fuentes de datos que se irán comentando a lo largo del libro pero que se pueden insertar en alguna de estas cinco grandes categorías.

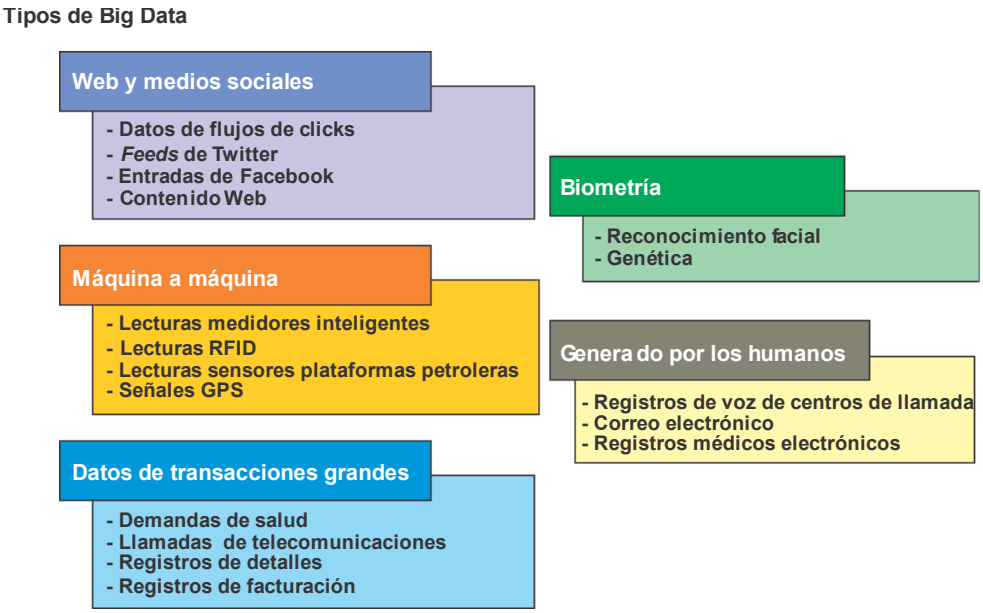

Figura 5.4. Fuentes de datos de Big Data. Fuente: Soares (2012)[22] (adaptada).

Web y social media (medios sociales)

Incluye contenido Web e información que es obtenida de los medios sociales como Facebook, Twitter, LinkedIn, Pinterest, Instagram; blogs como Technorati, de periódicos y televisiones; *wikis* como MediaWiki, Wikipedia; marcadores sociales como Del.icio.us, Stumbleupon; agregadores de contenidos como Digg, Meneame. En esta categoría los datos se capturan, almacenan o distribuyen teniendo presente las características siguientes: los datos incluyen datos procedentes de los flujos de clics, *tuits*, *retuits* o entradas en general (*feeds*) de Twitter, Tumblr, entradas (*posting*) de Facebook y sistemas de gestión de contenidos Web diversos tales como YouTube, Flickr, Picasa; o sitios de almacenamiento de información como Dropbox, Box.com, One Drive...

Los datos de la Web y de los medios sociales se analizan con herramientas de analítica Web y analítica social mediante el uso de métricas y de indicadores KPI.

Máquina-a-Máquina (M2M) /Internet de las cosas

M2M se refiere a las tecnologías que permiten conectarse a otros diferentes dispositivos entre sí. M2M utiliza dispositivos como sensores, medidores que capturan datos de señales particulares (humedad, velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad), contadores inteligentes (medición de consumo de electricidad en hogares, oficina, industria...). Estos datos se transmiten a través de redes cableadas, inalámbricas, móviles, satélites... a otros dispositivos, aplicaciones... que traducen estos datos

en información significativa. Entre los dispositivos que se emplean para capturar datos de esta categoría podemos considerar chips o etiquetas RFID, chips NFC, medidores inteligentes (de temperaturas, de electricidad, presión), sensores, dispositivos GPS que ocasionan la generación de datos mediante la lectura de los medidores, lecturas de los chips RFID y NFC, lectura de los sensores, señales GPS, señales de GIS, etcétera.

La comunicación M2M ha originado el conocido *Internet de las cosas o de los objetos* (véase Capítulo 6) , que representa a los miles de millones de objetos que se comunican entre sí y que pueden acceder si es necesario a Internet.

Transacciones de grandes datos

Son los grandes datos transaccionales procedentes de operaciones normales de transacciones de todo tipo. Incluye registros de facturación, en telecomunicaciones y registros detallados de las llamadas (CDR, Call Detail Record), con contenidos de información, sobre origen, destino, duración, y otros, como los datos de los teléfonos móviles inteligentes y tabletas. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados. Los datos generados procederán de registros de llamada de centros de llamada, departamentos de facturación, reclamaciones de las personas, presentación de documentos, etcétera.

Biometría

La biometría o reconocimiento biométrico[23] se refiere a la identificación automática de una persona basada en sus características anatómicas o trazos personales, tales como información procedente del cuerpo humano y actividad física (huellas digitales, reconocimiento facial, escaneo de la retina, genética...). Los datos anatómicos se crean a partir del aspecto físico de una persona, incluyendo huellas digitales, iris, escaneo de la retina, reconocimiento facial, genética, ADN, reconocimiento de voz, incluso olor corporal. Los datos de comportamiento incluyen análisis de pulsaciones y escritura a mano. Los avances tecnológicos han incrementado considerablemente los datos biométricos disponibles.

Algunas aplicaciones de interés son: en el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación; en el área de negocios y de comercio electrónico, los datos biométricos se pueden combinar con datos procedentes de medios sociales, lo que hace aumentar el volumen de datos contenidos en los datos biométricos.

Los datos generados por la biometría se pueden agrupar en dos grandes categorías: genética y reconocimiento facial.

Datos generados por las personas

Las personas generan enormes y diversas cantidades de datos como son la información que guarda un centro de llamadas telefónicas (*call center*) al establecerlas, notas de voz, correos electrónicos, documentos electrónicos, estudios y registros médicos electrónicos, recetas médicas, documentos papel, faxes. El problema que acompaña a los documentos generados por las personas es que pueden contener información sensible que necesita, normalmente, quedar oculta, enmascarada o cifrada de alguna forma para conservar la privacidad. Por eso, estos datos necesitan ser protegidos por las leyes nacionales o supranacionales (como es el caso de la Unión Europea o el Mercosur) relativas a protección de datos y privacidad.

Una característica importante en el caso de los datos procedentes de los seres humanos es la trazabilidad, huellas o "rastro digital" que dejamos las personas al navegar u utilizar Internet y los diferentes sitios que visitamos, tales como páginas de periódicos, revistas o blogs, el uso de las redes sociales, etc. Esta **huella** o **traza digital** identifica el «rastro» y, en consecuencia, nuestra identidad digital, que se puede utilizar para definir nuestro perfil o patrón de comportamiento (fotos, videos que subimos a la nube, contenidos más demandados, mensajes o *post* (entradas o artículos) que publicamos en Facebook o Twitter, búsquedas que realizamos, clic en "me gusta", etiquetas en las redes sociales, mensajes que publicamos, búsquedas que realizamos, etc.).

DATIFICACIÓN

«La huella digital que dejan la mayoría de las actividades humanas e informáticas se puede recoger y analizar para proporcionar información sobre diversos asuntos, desde la salud hasta el crimen o el rendimiento empresarial. Las pocas actividades que no dejan una huella digital actualmente pronto lo harán» Marr (2015: 56). Bernard Marr denomina a este proceso de captura de datos útiles como ***datificación*** y considera que existen muchas formas de datos útiles. Algunas de las formas son recientes como las publicaciones en redes sociales y otras existen desde hace mucho tiempo, como es el caso del registro de conversaciones, pero la carencia de capacidad de almacenamiento suficiente o un modo de analizar estas grabaciones para guardarlas han limitado su utilidad. Los grandes centros de datos de la nube y de las organizaciones que permiten y facilitan esta tarea de almacenamiento está cambiando el proceso de *datificació*n. Así, recuerda Marr que los datos se extraen de numerosas fuentes, tales como: nuestras actividades, nuestras conversaciones, fotos y videos, sensores y, últimamente, sobre todo, internet de las cosas.

DATOS EN ORGANIZACIONES Y EMPRESAS

Los datos que manejan las organizaciones y empresas se agrupan en dos grandes categorías: datos internos y datos externos. Todos ellos a su vez, como ya se ha comentado anteriormente pueden ser datos estructurados, no estructurados o semiestructurados.

DATOS INTERNOS

«Los datos internos representan todo aquello a lo que su empresa tiene o podría tener acceso en la actualidad, incluyendo los datos personales o privados que recoge la empresa y pertenecen a ésta, cuyo acceso está controlado por usted» (Marr 2015: 55). Algunos ejemplos citados por Marr son:

- Comentarios de los clientes
- Datos de ventas.
- Datos de las encuestas a los empleados o los clientes
- Datos en video de circuitos cerrados de televisión
- Datos transaccionales
- Datos de registros de los clientes
- Datos de control de existencias
- Datos de RR. HH

Los datos internos de una empresa están constituidos por todos aquellos datos que recoge y pertenecen a ella y cuyo control está gestionada por ella misma, aunque la recolección, almacenamiento, proceso y utilización de dichos datos se realiza por sus empleados; el uso cada día más frecuentes de los robots conversacionales (*chatbots*) hace que una gran parte de la empresa puede automatizarse, sobre todo los relacionados con la gestión y administración de los datos de los clientes en sus comunicaciones con la empresa.

DATOS EXTERNOS

«Los datos externos de una organización son una variedad infinita de información que existe fuera de la misma. Los datos externos pueden ser públicos y privados. Los públicos son datos que todas las personas pueden obtener, tanto reuniéndolos gratuitamente como pagando a un tercero para conseguirlos, o haciendo que un tercero los recoja por usted. Los privados normalmente son aquellos que necesitaría conseguir y por los que tendría que pagar a otra empresa o a un tercero,

proveedor de datos.» (Marr 2015:55-66). Algunos ejemplos de datos externos citados por Marr (2015: 56) son:

- Datos meteorológicos.
- Datos oficiales como los censales.
- Datos de Twitter.
- Datos de perfiles en redes sociales.
- Google Trends o Google Maps,
- [Datos de Facebook, Instagram, Pinterest, Amazon, Microsoft, otras aplicaciones de Google]

ARQUITECTURA DE BIG DATA

La gestión (administración) de grandes volúmenes de datos requiere de una arquitectura específica que se compone de una serie de capas o etapas que manejan los datos, desde su captura de las diferentes fuentes de datos hasta su etapa final de visualización de los resultados obtenidos. Las cuatro capas más consideradas en el proceso de tratamiento de Big Data son:

- Recolección (ingesta) de datos
- Almacenamiento
- Análisis de datos
- Visualización de datos (resultados)

Previa a la recolección de datos se requiere una etapa previa de **identificación de las fuentes de datos** y que es muy importante en la decisión de la arquitectura ya que implica identificar las diferentes fuentes de datos y su clasificación en función de su naturaleza y tipos, como vimos en apartados anteriores. Los aspectos a considerar en la identificación de las fuentes de datos son:

- Identificar las fuentes internas y externas.
- Calcular la cantidad de datos detectada (a ingerir) de cada fuente de datos.
- Identificar los mecanismos de obtención de datos (*push* o *pull)*.
- Determinar el tipo de fuente de datos (archivos, bases de datos, datos web...)

- Determinar el tipo de datos: estructurado, no estructurado o semiestructurado.

Figura. 5.5. Arquitectura de Big Data

RECOLECCIÓN DE DATOS

Es la etapa de obtención (ingesta) de datos. Se ha convertido en una etapa de gran interés en el proceso de Big Data ya que existen numerosos datos públicos que se producen en enormes cantidades, numerosos dispositivos desperdigados por todo el planeta que emiten, procesan y recogen información de las más diversas actividades (posicionamiento de individuos y vehículos, niveles de contaminación, temperaturas...); de igual forma, infinidad de dispositivos móviles que también emiten y capturan datos, etcétera.

Existen dos métodos de recolección de datos que se utilizarán según los casos:

- *Por lotes (batch).* Se conecta cada cierto tiempo a las fuentes de información (sistemas de archivos o bases de datos).
- *Tiempo real (streaming).* Este tipo de recolección de información trata directamente con la fuente de información de manera continua y en forma que la información se obtiene en tiempo real, cada vez que se necesita.

Los sistemas de información actuales, y sobre todo, los específicos de tratamiento de Big Data, pueden obtener la información de las dos formas anteriores.

ALMACENAMIENTO DE DATOS

Los sistemas de almacenamiento tradicionales se han tenido que adaptar a las grandes cantidades de datos que se generan, así como a la velocidad a la que se producen. Por esta razón las bases de datos tradicionales (relacionales) no se adaptan a estas necesidades y se requieren nuevos sistemas de almacenamiento. Entre otras características, los nuevos sistemas de almacenamiento deben ser **escalables,** precisamente debido a los grandes volúmenes de datos que precisan las compañías y de cualquier tipo, los cuales tienen que procesarse de acuerdo a las necesidades de las mismas (aumentando o disminuyendo sus capacidades). Este tipo de almacenamiento escalable tendrá que ser más transparente y

eficiente dado que debe permitir la ampliación o reducción requerida y, por consiguiente, las tecnologías utilizadas se han de adaptar a esta característica.

Los sistemas de almacenamiento de Big Data más utilizados son: **Hadoop y Spark** —sistema por excelencia de archivos distribuidos—, bases de datos **NoSQL (MongoDB, HBase, Cassandra**...) y las bases de datos **en memoria** (SAP Hana). Los sistemas de almacenamiento actuales deben permitir la integración de los datos de estos sistemas con los datos tradicionales almacenados en las bases de datos relacionales. La integración de todos los tipos de datos será el gran éxito de los sistemas de almacenamiento y, por consiguiente, integrar datos procedentes de los almacenes de datos (*data warehouses*, *data marts)*, sistemas de datos operacionales y los citados sistemas distribuidos como las bases de datos **NoSQL** o HDFS de **Hadoop.** Precisamente una de las grandes ventajas de HDFS de **Hadoop** es su capacidad de almacenamiento escalable (aumenta o reduce su cantidad y capacidad de almacenamiento según requiera el usuario).

Otra de las características que debe tener el almacenamiento de datos actual está relacionada con los sistemas de análisis ya citados: **síncrono** (tiempo real) y con optimización a baja latencia, y **asíncrono** (los datos se capturan, registran y analizan por lotes).

PROCESAMIENTO Y ANÁLISIS DE DATOS

Esta etapa suele considerarse como una o dos etapas según la metodología utilizada. Una vez que se tienen almacenados los datos, se han de convertir en conocimiento (valor) mediante el procesamiento y análisis de toda la información almacenada. Se trata de ser capaz de procesar datos en un tiempo razonable y alejarse de los estudios tradicionales de mercados estáticos. Los tipos de procesamiento son: *Batch* (por lotes), *Streaming* (en tiempo real) e *híbrido*.

En el procesamiento por lotes se recolecta la entrada para un intervalo especificado de tiempo y las transformaciones se ejecutan de un modo planificado; la carga de datos históricos es una operación típica por lotes. Las tecnologías emergentes más utilizadas son MapReduce, Hive y Pig.

El procesamiento en tiempo real implica la ejecución de las transformaciones de datos cuando éstos se recogen; las tecnologías más utilizadas son Spark, además de los componentes de Hadoop.

El procesamiento conduce al análisis de datos. Las soluciones tradicionales de análisis de datos suelen ser predefinidas y lentas, lo cual, ante un incremento del volumen de los datos y una variedad en su origen, proporcionan una información limitada ya que sólo pueden analizar terabytes de datos estructurados y, actualmente, se almacenan y manejan petabytes y exabytes de datos. Las soluciones más idóneas son básicamente específicas para Big Data que ofrecen unas técnicas de analítica más ágiles y proactivas de este tipo de información.

El análisis de datos almacenados utiliza modelos, algoritmos y herramientas adecuadas para proporcionar visibilidad sobre los datos para que puedan ser consultados en la capa de visualización o capa de consumo.

Esta etapa es decisiva y en la actualidad el análisis de Big Data se realiza por profesionales especializados de administradores de bases de datos y científicos de datos.

VISUALIZACIÓN DE DATOS

Los resultados del análisis de datos es la etapa de consumo que debe permitir su visualización para una correcta toma de decisiones. Esta capa de Big Data muestra el producto del almacenamiento y procesamiento de la información cuyo resultado es la producción de conocimiento. En la actualidad existe un gran número de herramientas de visualización de datos que proporcionan una gran eficiencia a las compañías.

Las herramientas de visualización permiten a los usuarios finales hacer búsquedas y acceder a la información rápidamente, en algunos casos en tiempo real, de modo que los usuarios puedan tener el control de la información en el mismo momento en que se produce. La enorme cantidad de herramientas de visualización de datos se agrupa por categorías: gráficos, mapas, *cartogramas*, tablas, infografías, nubes de palabras. En el capítulo 11 "Analítica de datos (Big Data & Analytics) trataremos detalladamente las herramientas gráficas más recomendadas cuya selección es un criterio importante en la hora del tratamiento de grandes volúmenes de datos. Hay una gran oferta de herramientas de visualización gratuitas y de pago; una selección de herramientas muy empleadas es: Tableau, Canva, Google Fusion Tables, QlickViewm CartoDB, D3.js, etc. El organismo oficial español Red.es publicó un estudio de herramientas de visualización de datos, disponible en su página web: *Recopilación de herramientas de procesamiento y visualización de datos.*[24]

OPEN DATA. EL MOVIMIENTO DE LOS DATOS ABIERTOS

Una variante muy importante de Big Data es la estrategia *Open Data* (datos abiertos) o apertura de datos. La estrategia *Open Data*, que históricamente nació en 2009 en Washington (ciudad pionera en este movimiento, **`data.gov`**), se refiere a la posibilidad de que el ciudadano acceda a los datos del Gobierno que antes sólo eran analizados en el interior de las administraciones públicas.

Open Data consiste en una iniciativa para poner a disposición de las personas y empresas residentes en el país, datos de carácter público.

Aunque la iniciativa de *Open Data* nació en los Estados Unidos, hoy día forma parte de la Agenda Digital Europea, donde numerosos países (entre ellos España)

han promovido iniciativas de datos abiertos, así como en América Latina, donde se desplegaron y promovieron también pronto, iniciativas nacionales de Open Data en países tales como Perú, México, Argentina y Colombia, La tendencia Big Data puede proporcionar una gran ventaja competitiva a las empresas y grandes beneficios a los usuarios y ciudadanos en general, en el movimiento y las tendencias de datos abiertos.

«Lo primero que tienen que hacer los gobiernos es hacer más datos abiertos» ,[25] afirmaba Jeff Jaffe, presidente ejecutivo del W3C, en la *Bilbao Web Summit 2011*; y Tim Berners-Lee, en su conferencia en el mismo congreso, manifestó: "El futuro social de la educación está en los datos, en la calidad de los mismos, los datos abiertos (*open data*), la libertad de los datos, que éstos puedan fluir para el acceso de cualquier persona y que, a su vez, puedan ser aprovechados". Así se expresaban los dos principales directivos de W3C (Consorcio de la W3): Tim Berners-Lee es inventor de la Web, y actual director del W3C, y Jeff Jaffe, su presidente ejecutivo. Aquí cabe preguntarse qué es *Open Data* (datos abiertos) y cuáles son las iniciativas a nivel internacional que están difundiendo la iniciativa.

El W3C está impulsando en todo el mundo el movimiento a favor de la apertura de datos públicos. La enciclopedia Wikipedia define *Open Data* como: "Una filosofía y práctica que requiere que ciertos datos estén disponibles libremente para cualquier persona sin restricciones de *copyright*, patentes u otros mecanismos de control".

El movimiento de datos abiertos comenzó su explosión en 2010 y ha crecido a pasos agigantados en toda la década, sobre todo por el apoyo ofrecido por el gobierno de los Estados Unidos (`http://www.data.gov.gb`); en Europa, por el gobierno de Gran Bretaña (`http://www.data.gov.gb`); en la propia Unión Europea (`http://www.ec.europa.eu`); y en España (`datos.gob.es`) , con numerosos gobiernos autonómicos (regionales) como Euskadi, Asturias, Cataluña, Navarra, entre otros, y ciudades como la milenaria Córdoba. En América Latina y Caribe, como ya se ha comentado anteriormente y aunque de un modo más lento, también se han puesto en marcha iniciativas de *Open Data* en la mayoría de los países (Colombia, Perú, Chile Uruguay, México, entre otras naciones).

En la práctica Open Data[26, 27] es la puesta a disposición de la sociedad de gran cantidad de datos procedentes de diferentes organizaciones, fundamentalmente del ámbito de la administración pública o de aquellos proyectos que han sido financiados con dinero público, de manera libre. En general, los datos proporcionados se refieren a diferentes temáticas (médicos, geográficos, meteorológicos, biodiversidad, servicios públicos, etc.). Cuando hablamos de *Open Data* nos referimos a información general que es posible utilizar libremente, reutilizar y redistribuir por cualquier persona, y que puede incluir datos geográficos, estadísticos, meteorológicos, así como datos de proyectos de investigación financiados con fondos públicos o libros digitalizados de las bibliotecas.

El objetivo fundamental de abrir los datos a la sociedad es que ésta pueda obtener provecho de ellos; es decir, se trata de que cualquier persona, organización o empresa pueda sacarles utilidad como simple conocimiento, o bien con iniciativas altruistas o empresariales que le saquen el mayor rendimiento posible.

En la práctica, las administraciones gestionan bases de datos, listados, estudios, información general; es decir, materia prima con gran potencial y que, al haber nacido del dinero público y al ponerse al servicio de la ciudadanía, puede ofrecer oportunidades de negocios a emprendedores tanto en el aspecto personal como en el de la empresa.

"Las administraciones generan multitud de información en forma de datos propios que son de difícil acceso para la mayoría de los ciudadanos; datos de diversa índole que van desde tablas estadísticas, oportunidades laborales, recursos turísticos o incidencias de tráfico y que normalmente se encuentran perdidos en las páginas Web de los organismos" (Ruiz-Tapiador, 2010)[28]. Los datos abiertos son muy aprovechables y generan valor añadido a las empresas. En el sector público, tener acceso a los datos de la administración garantiza la transparencia, la eficiencia y la igualdad de oportunidades, a la vez que se crea valor (Generalitat de Cataluña, Gobierno de la Comunidad Autónoma de Cataluña en España)[29]. La *transparencia*, porque se puede consultar y tratar datos que vienen directamente de las fuentes oficiales; la *eficiencia*, porque ciudadanos y organizaciones pueden crear servicios en forma más ajustada en colaboración con la administración; y la *igualdad de oportunidades*, porque el acceso es el mismo para todo el mundo.

En cuanto a las licencias y términos de uso de los datos abiertos, éstos deben estar sometidos a las leyes de reutilización de la información del sector público del país donde se está poniendo en marcha la iniciativa de Open Data. En algunos casos, pueden tener derecho de propiedad intelectual, pero siempre se tratará de dejarlas abiertas con los términos de uso y licencias legales.

INICIATIVAS *OPEN DATA*

Las iniciativas de Open Data en el mundo son numerosas. Los proyectos más innovadores han nacido en los Estados Unidos con la primera administración del presidente Obama, y en Gran Bretaña; en España a nivel regional o autonómico y local, aunque también existieron iniciativas nacionales como el *Proyecto Aporta* dentro del *Plan Avanza*, que desde 2007 plantearon que todas las administraciones locales, autonómicas y centrales estaban llamadas a hacer pública la información que generan, hechos que se fueron confirmando de modo progresivo.

Estados Unidos (Data.gov)

Sin lugar a dudas el portal **DATA.GOV** es una referencia obligada en el estudio de Open Data . En mayo de 2013, el portal incluía páginas relativas a Data, Apps, Communities, Open Government, Metrics, Semantic Web y Blogs.

Figura 5.6. Pantalla inicial de www.data.gov (mayo de 2013)

En una visita realizada el 9 de mayo de 2013, a la opción (pestaña) *Open Data Sites* (sitios de datos abiertos) aparecen 39 estados, 35 ciudades, 41 países y 183 iniciativas regionales de Open Data. En Europa, son numerosos los países con iniciativas nacionales de Open Data (Gran Bretaña, Francia, Italia, Rusia, España...); en América del Norte, los Estados Unidos y Canadá; en América Latina, Argentina, Brasil, Chile, México, Perú y Uruguay figuraban en el portal oficial de data.gov con una iniciativa en marcha de todos los países de la zona. En Asia y África hay algunos países con iniciativas de datos abiertos.

En abril de 2017, el portal tal como se define en su página web, es: «The home of the U.S. Government´s open data» («La casa de los datos abiertos del Gobierno de Estados Unidos»). En la página se encuentran opciones de datos, herramientas y recursos para desarrollar investigaciones, desarrollar aplicaciones web y móviles, diseñar visualizaciones de datos y estrategias de cómo obtener datos o añadir sus datos al portal.

Los temas de datos abiertos del portal abarcan casi todos los sectores de impacto de la sociedad: agricultura, clima, consumidor, ecosistemas, educación, energía, finanzas, salud, gobierno local, industria y fabricación, marítimo, océanos, seguridad pública e investigación y ciencia. Así mismo proporciona datos de gobiernos abiertos para potenciar aplicaciones de software que ayuden a las personas a tomar decisiones de todo tipo.

Figura 5.7. Portal Open Data del Gobierno de Estados Unidos
Fuente: <www.data.gov> (abril 2017)

De igual forma, data.gov proporciona herramientas y recursos muy buenos para los hackers cívicos (*civic hackers*), emprendedores tecnológicos, científicos de datos y desarrolladores de "todas las franjas y colores (*stripes*)"

Figura 5.8. Portal Open Data del Gobierno de Estados Unidos
Fuente: <www.data.gov> (abril 2017)

Reino Unido (data.gov.uk)

El portal oficial *Opening Up Government*, del Reino Unido, es también otro modelo para el estudio de *Open Data*. Igual que sucede con el portal del gobierno de los Estados Unidos, ofrece una amplia oferta de opciones: Datos, Apps, Foros, Blogs, Biblioteca, Publicaciones.

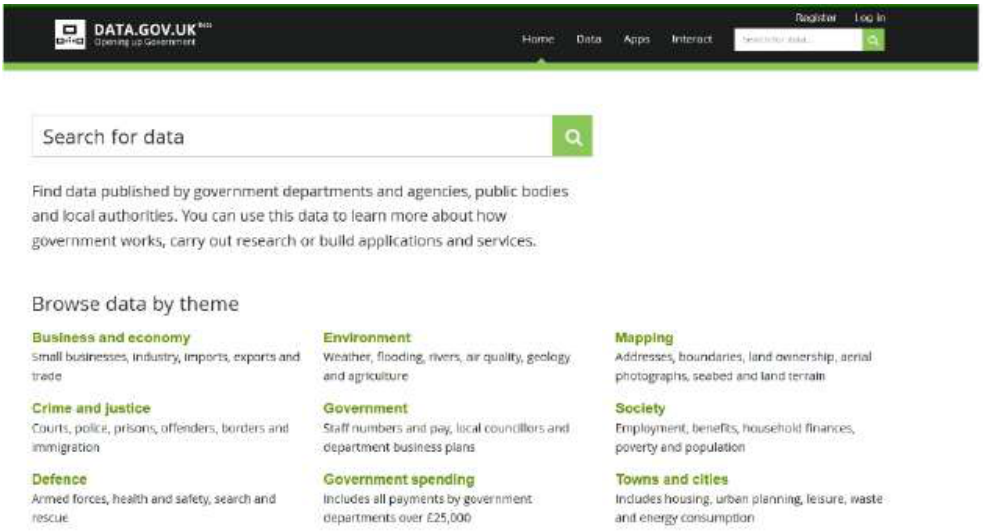

Figura 5.9. Pantalla de <www.data.gov.uk> [Consulta: 10 abril 2017].

España

Las iniciativas en España han sido numerosas tanto en la administración nacional como en las administraciones autonómicas y locales.

Las iniciativas pioneras comenzaron en los gobiernos autonómicos: el Principado de Asturias y el País Vasco, a las que pronto se sumaron la Generalitat de Cataluña y el Gobierno de Navarra, y nacieron, muy pronto, iniciativas locales como los ayuntamientos de Córdoba y Zaragoza. Prácticamente la mayoría de los gobiernos autonómicos y locales tienen portales de datos abiertos o están en proceso de su diseño y construcción.

Figura 5.10. Portal de datos abiertos del Gobierno de España (data.gob.es)

América Latina y el Caribe

Las iniciativas pioneras en América Latina y el Caribe fueron muy tempranas. Según el portal (**data.gov**)del gobierno federal de los Estados Unidos, a finales de octubre de 2011 en América Latina y el Caribe sólo existía un país con iniciativa de Open Data: Perú, con el proyecto *Open Data Perú*; a finales de febrero de 2012 se incorporó Uruguay, y en mayo de 2013 Argentina, Brasil, Chile, México, Perú y Uruguay. En los siguientes años la iniciativa se ha ido extendiendo a, prácticamente, todos los países de Latinoamérica y el Caribe,

Figura 5.11. Portal original de datos abiertos de Perú www.datosperu.org (año 2011)

Figura 5.12. Portal nacional de datos abiertos de Perú (abril 2017)

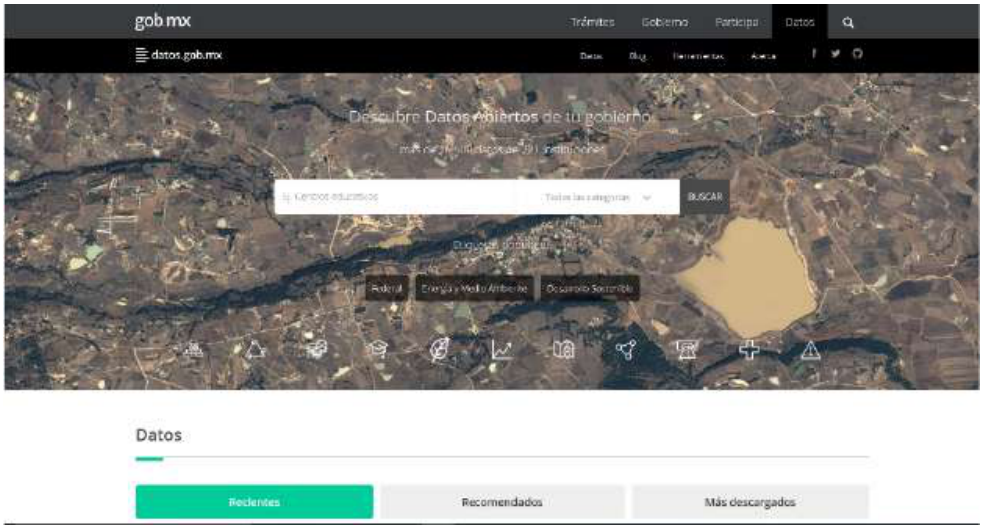

Figura 5.13. Portal de datos abiertos del Gobierno de México (abril 2017)

LA INICIATIVA DE LA UNIÓN EUROPEA

La Unión Europea lanzó a finales de diciembre de 2012 (el día de Nochebuena) la versión beta pública de su esperado portal Open Data, y su lanzamiento definitivo se desarrolló a lo largo del mes de enero de 2013. Sus objetivos iniciales se mantienen en 2017, pero le ha dado un énfasis especial a la sección de desarrolladores para facilitar las tareas de emprendimiento en la Unión Europea en este campo.

La Unión Europea anunció, en la inauguración de la página de su portal de datos abiertos sus objetivos -los cuales se mantienen y se han ampliado-:

> *Este portal trata de transparencia, gobierno abierto e innovación. El Portal de datos de la Comisión Europea proporciona acceso a datos públicos abiertos de esta institución. Pero además, permite que otras instituciones, organismos, oficinas y departamentos de la Unión accedan a los datos previa solicitud. Cualquier persona interesada puede descargar los datos publicados para reutilizarlos, vincularlos y crear servicios innovadores. Asimismo, este Portal de datos divulga y facilita el conocimiento sobre los datos de Europa. Los organismos editores, desarrolladores de aplicaciones y el público en general pueden aprovechar la tecnología semántica del Portal, que pone a su disposición esta nueva funcionalidad.*

La mayoría de los datos del portal, en una primera instancia, proceden de Eurostat, la oficina de estadística de la UE.

Figura 5.14. Portal datos abiertos de la Unión Europea https://data.europa.eu/euodp/es/data/ (abril 2017)

OPEN DATA CENTER ALLIANCE

La asociación internacional Open Data Center Alliance (http://www.opendatacenteralliance.org) es una organización sin ánimo de lucro, constituida en 2010 como un único consorcio de organizaciones líderes en IT, que ha nacido para trabajar en la configuración futura de ***cloud computing*** un futuro basado en estándares abiertos e interoperables, según la alianza. Las empresas miembros comparten una visión de los requerimientos, procesos y tecnologías que configuran la adopción de los servicios de la nube por las empresas de IT a nivel mundial. La importancia y fortaleza de la asociación es que incluye a más de 300 compañías mundiales y de las más variadas industrias. Su actual consejo de dirección está constituido por ejecutivos senior IT de las empresas BMW, Capgemini, CenturyLink, China Unicom, Deutsche Bank, Infosys. Intel, National Australia Bank y SAP.

La misión de la asociación es aumentar la velocidad de migración a *cloud computing*, facilitando el ecosistema de soluciones y servicios para dirigir los requerimientos de IT con el más alto nivel de interoperabilidad y estándares. Pretenden tener una voz unificada para buscar los requerimientos de *cloud computing* y los emergentes centros de datos. Una de las muchas virtudes que

tiene esta organización son sus publicaciones de carácter libre. Una que está directamente relacionada con los grandes volúmenes de datos es la *Big Data Consumer Guide*[30] y que fue publicada en el año 2012 coincidiendo con despliegue de la tendencia de *big data* en las empresas.

OPEN DATA INSTITUTE (ODI)

La organización Open Data Institute (ODI, www.theodi.org) fue creada en el año 2012, por Tim Berners-Lee, creador de la Web, y el catedrático (*Professor*) de Inteligencia Artificial Nigel Shadbolt. La ODI es una organización independiente sin ánimo de lucro, y ya en sus orígenes tuvo el apoyo económico del gobierno de Gran Bretaña (vía la agencia de innovación Technology Strategy Board) y de la organización Omidyar Network. Hoy día son innumerables sus miembros asociados tanto a nivel corporativo como personal Su sede es Londres y está dirigida a toda la comunidad de personas interesadas en desarrollar *Open Data*, a las que invitan a ponerse en contacto desde su página Web inicial.

El Open Data Institute pretende canalizar la evolución de una cultura de Open Data para crear valor económico, ambiental y social. Trata de desbloquear las fuentes, generar demanda y crear y diseminar el conocimiento, centrándose en temas locales y globales.

Entre sus objetivos fundacionales está convocar a expertos de nivel mundial para colaborar, incubar, nutrir y actuar de mentores de nuevas ideas, así como promover la innovación. Busca que cualquier persona pueda aprender a relacionarse con los datos abiertos, y la autonomía de los equipos para ayudar a los demás a través del *coaching* profesional y la tutoría.

El ODI define los *datos abiertos*[31] como: "Información que está disponible para cualquier persona que los utiliza, para cualquier propósito y sin ningún costo". Los datos abiertos tienen una licencia que deben aclarar que son datos abiertos. Sin una licencia, los datos no pueden ser reutilizados. La licencia también puede decir que:

- Las personas que utilizan los datos deben acreditar quién los está publicando. Esta característica se llama *atribución*.
- Las personas que mezclan los datos con otros datos tienen también que liberar los resultados. Esta característica se llama *compartir por igual*.

La ODI recomienda, en su definición, la palabra "abierta", dada por la organización Open Definition (http://www.opendefinition.org) para los términos: *Open Data*, *Open Content* y *Open Services*.

RESUMEN

Big Data, grandes datos, grandes volúmenes de datos o macrodatos, están constituidos por la avalancha de datos procedentes de las fuentes más diversas: movilidad, medios sociales, Internet de las cosas, M2M, sensores, computación en la nube.

- La cantidad de datos crece de manera espectacular. En 2011 fueron 1,8 zettabytes; en 2012, 2,8 zettabytes; y para 2020 se prevén 40 zettabytes (Informe Digital Universe de IDC/EMC 2012).
- Big Data no sólo se considera en términos de *grande* (*volumen*), sino en términos de **variedad** y **velocidad** (modelo de las 3 V). Este modelo se ha extendido para incluir las características de **veracidad** y **valor** (modelo de las 5 V).
- Los tipos de datos se clasifican en tres grandes grupos: estructurados (bases de datos tradicionales o relacionales), semiestructurados y no estructurados.
- Uno de los grandes riesgos que entrañan los Big Data son las implicaciones de privacidad que acompañan a muchas de las fuentes de datos, origen de los grandes datos.
- La integración de los datos tradicionales con los Big Data supone una gran oportunidad de negocio para organizaciones y empresas.
- La explosión de los Big Data se ha producido en los últimos años por las innumerables fuentes de datos que han ido proliferando desde los datos de texto y no textuales, de contenidos de audio, fotografía y video, datos de teléfonos inteligentes y tabletas, de los *social media*, sensores...
- Los Big Data no constituyen una amenaza como tal, sino más bien un reto y una oportunidad para organizaciones y empresas.
- La historia del término Big Data desde el punto de vista académico se remonta a 1984, y desde el punto de vista comercial o empresarial a 1987. En 2001, Laney publica un artículo profesional que titula "*Three V´s of Big Data (Volume, Variety and Velocity)*" donde conceptualiza el significado del término y el fenómeno. Estas características han sido aceptadas como las fundamentales en la definición. 2008, con la publicación del artículo de la "Era del exabyte", en *Wired*; y 2010, con la publicación de artículos e informes en diversos medios de comunicación como *The Economist* y *Forbes*, se consideran las fechas de partida de Big Data como fenómeno social, tecnológico, económico y empresarial.
- Los grandes volúmenes de datos existentes en la actualidad y utilizados por organizaciones, empresas y particulares, proceden de numerosas fuentes

que capturan y generan datos estructurados, no estructurados y semiestructurados, tales como sensores, medios sociales, dispositivos móviles (teléfonos, tabletas, videoconsolas…), dispositivos de detección y localización de posición geográfica de objetos y personas, datos climatológicos.

- Una taxonomía global de las fuentes de datos que alimentan a los Big Data y que se ha considerado en el capítulo (Soares, 2012) es:
- Web y Social Media (medios sociales: redes sociales, blogs, wikis, gestión de contenidos audio, video, fotografías, libros…).
- Máquina a Máquina (M2M, Internet de las cosas), sensores, chips NFC y RFID…
- Transacciones de todo tipo: banca, comercio, seguros…
- Biometría: datos biométricos de las personas e incluso animales.
- Las propias personas generan gran cantidad de datos: documentos, correos electrónicos, faxes, mensajes instantáneos, facturas, recetas médicas…
- Los datos abiertos (*Open Data*) se refieren a los datos públicos y privados que deberían estar a disposición de los ciudadanos y empresas para un uso eficaz y rentable. Naturalmente, los datos abiertos deberán respetar siempre la privacidad y la información que deba estar protegida, como datos de salud, personales, pero se requiere que se abran y que sean interoperables por las distintas plataformas utilizadas por los desarrolladores, y deben ser también legibles y entendibles por los ciudadanos.
- Los Estados Unidos, Canadá y Europa son pioneros en este movimiento mundial por la apertura de los datos, a los que poco a poco se van sumando otros países de los diferentes continentes. En el caso de América Latina, Perú y Uruguay han sido los primeros países oficialmente reconocidos por el portal Open Data (data.gov) del gobierno federal de los Estados Unidos.
- En Europa, diferentes países, y en España, diferentes comunidades autónomas, han puesto en marcha iniciativas de Open Data.

BIBLIOGRAFÍA

- **CABALLERO**, Rafael y **MARTÍN,** Enrique (2015). *Las bases de Big Data*. Madrd: Los Libros de la Catarata.
- **FRANKS**, Bill (2012). *Taming the Big Data Tidal Wave. Finding Opportunities in Huge Data Streams with Advanced Analytics*. New Jersey: Wiley.
- **JOYANES**, Luis (2014). *Big Data. Análisis de grandes volúmenes de datos*. Ciudad de México: Alfaomega; Barcelona: Marcombo.
- **JOYANES**, Luis (2015). *Sistemas de Información en la Empresa. El impacto en la nube, la movilidad y los medios sociales.* Ciudad de México: Alfaomega; Barcelona: Marcombo.
- **MARR,** Bernard (2016). *Big Data. La utilización del big data, el análisis y los parámetros SMART para tomar mejores decisiones y aumentar el rendimiento.* TEELL Editorial.
- **PÉREZ** Marqués, María (2015). *Big Data. Técnicas, herramientas y aplicaciones*. Madrid: Catarata,
- **SCHMARZO**, Bill (2013). *Big Data. El poder de los datos.* Madrid: Anaya Multimedia.
- **SOARES**, Sunil (2012). *Big Data Governance. An Emerging Imperative*. Boise, MC Press Online
- **SOLANA**, Albert y **ROCA,** Genis (2015). *Big Data para directivos*. Barcelona: Empresa activa.
- **TASCON**, Mario y **COULLAUT**, Arantza (2016). Big Data y el Internet de las Cosas. Madrid: Catarata.

NOTAS:

[1] *Adrian Merv: "Big Data", en Teradata Magazine, 2011 Q1. Disponible en:* <http://ww.nxtbook.com/nxtbooks/mspcomm/teradata_2011q1/index.php?startid=8#/40>

[2] *La consultora McKinsey a través de McKinsey Global Institute publicó el informe que se ha convertido en un clásico, consultado y referenciado por numerosas organizaciones y empresas, así como profesionales. Big data: The next frontier for innovation, competition, and productivity, mayo 2011. Disponible en:*

<http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation>

[3] *Consultora IDC. Disponible en:* <http://mx.idclatin.com/releases/news.aspx?id=1433>

[4] *Predicciones de Deloitte para el sector de tecnología, medios de comunicación y telecomunicaciones 2012. Disponible en:* <http://www.deloitte.com/assets/Dcom-Mexico/Local%20Assets/Documents/mx%28es-mx%29TMT2012_Esp.pdf>

[5] *CEO Advisory: "Big Data", en Equals Big Opportunity, 31 marzo, 2011*

[6] *Howard Elias: "El desafío de Big Data: Cómo desarrollar una estrategia ganadora", en CIO, julio 2012. Disponible en:* <http://cidperu.pe/articulo/10442/el-desafio-de-big-data-como-desarrollar-una-estrategia-ganadora>.

[7] *Sitio de IBM de big data: "Bringing big data to the enterprise". Disponible:* <http://www_01.ibm.com/software/data/bigdata>

[8] *Mark Beyer, Gartner "Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data",* http://www.gartner.com/newsroom/id/1731916 *, 27 junio, 2011.*

[9] *Mark Beyer, Gartner "Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data",* http://www.gartner.com/newsroom/id/1731916 *, 27 junio, 2011.*

[10] *Ibid, IBM, p. 8.*

[11] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[12] *Bernard Marr. Big Data: The 5 Vs Everyone Must Know.* https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know *Posteriormente en su libro: Big Data: Using Smart Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance, Wiley,2015, volvió a dar la misma definición de Big Data.*

[13] *IDC: "The Digital Universe Decade. Are You Ready?". Patrocinado por EMC, mayo 2011.*

[14] *Francis X. Diebold: "A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline", University of Pennsylvania, First Draft, August 2012. Este draft: 26 de noviembre de 2012. Disponible en:* <http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf>

[15] *El artículo lo publicó en 2001 como una nota de investigación del META Group, en la actualidad forma parte de Gartner. Tal vez aquí resida el hecho de que, en sus*

publicaciones, Gartner definía las características de los Big Data con las 3 V, y a Laney como el padre del modelo de las 3 V. Disponible en: < http://goo.gl/Bo3GS>

[16] *Steve Lohr: "How Big Data Became So Big", en The New York Times. Disponible en:* <http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html?_r=0>. *Publicado en la edición impresa del 12 de agosto de 2012.*

[17] *"The Petabyte Age: Because More Isn't Just More. More Is Different", en Wired. Disponible en:*
<http:www.wired.com/science/discoveries/magazine/16-07/pb_intro>

[18] *Chris Anderson: "Will the Data Deluge Makes the Scientific Method Obsolete?" [Consulta: 6.30.08].*

[19] *Sus autores han sido tres prominentes científicos de las Ciencias de la Computación: Randal E. Bryant (Carnegie Mellon University), Randy H. Katz (Universidad de California, Berkeley) y Edward D. Lazowska (Universidad de Washington). Disponible en:*
<http://www.cra.org/ccc/docs/init/Big_Data.pdf>

[20] *<http://www.significancemagazine.org/view/0/index.html>*

[21] *Op. cit. Big data: The next frontier for innovation, competition, and productivity, cuadro 7, p. 19.*

[22] *Sunil Soares (2003). Big Data Governance. An Emerging Imperative. Boise. MC Press Online. El autor de este libro mantiene un blog excelente sobre Big Data y Gobierno de Big Data.*

[23] *"An Overview of Biometric Recognition", disponible en:*
<http://biometrics.cse.nsu.edu/info.html>

[24]
datos.gob.es/sites/default/files/files/Herramientas_de_Visualización.docx

[25] *Entrevista en El Mundo a Jeff Jaffe, Madrid, 21 de mayo de 2011.*

[26] *"Comienza el movimiento Open Data", en Computer World, [Consulta: 21 de mayo de 2011].*

[27] *Definiciones académicas de Open Data se pueden ver en la organización Open Definition (*http://www.opendefinition.org *), y en el Open Data Institute (*http://www.theodi.org*). Ambas instituciones se describirán más adelante.*

[28] *Teresa Ruiz-Tapiador: "Suplemento PYMES RI+D+I", en Cinco Días, Madrid, 20 septiembre 2010, pp. 2-3. Analiza el fenómeno de Open Data (datos abiertos) desde una perspectiva de negocio y empresa.*

[29] *El portal de la Generalitat de Cataluña (*http://www.dadesobertes.gencal.cat *) ofrece una buena documentación de Open Data.*

[30] <http://www.opendatacenteralliance.org/docs/Big_Data_Consumer_Guide_Rev1.0.pdf>

[31] <http://www.theodi.org/guide/what-open-data>