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

CONTENIDO

OBJETIVOS

- Examinar y conocer la definición y características principales de Big Data.
- Identificar y analizar los diferentes tipos de datos manejados en los sistemas de Inteligencia de Negocios.
- Localizar e identificar las diferentes fuentes de datos.
- Introducción a la tendencia de *datificación*.
- Describir la arquitectura de Big Data.
- Identificar el ecosistema Hadoop (soporte de Big Data) y sus componentes principales.
- Conocer e identificar las herramientas más utilizadas en los sistemas de Big Data.
- Examinar y analizar el concepto de Open Data y su impacto en los sistemas de Inteligencia de Negocios.

5.1 INTRODUCCIÓN

El año 2016 fue el despliegue universal de *Big Data* y 2017 supuso 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 y 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. Según la consultora IDC, Durante 2011 se crearon 1.8 zettabytes de datos (1 billón de gigabytes), 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 ya señalaba Bill Franks, director de Analítica de Teradata en 2012[1]: "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 las profesiones relacionadas con estos temas serán las más demandadas en los años sucesivos.

Profundizaremos en el concepto ya introducido 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 en 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 que producirá el enorme volumen de datos que se irá creando y difundiendo.

Hoy en 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, etcétera. 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 en 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 estos grandes datos son:

- Datos de la Web.
- Datos de los medios sociales (redes sociales, *blog*s, 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 y cine).
- Datos procedentes de sensores en diferentes campos de la industria y la agricultura.

- Datos de videojuegos en locales recreativos, casinos y 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 (*smartgrids)*.
- 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.

5.2 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 (*whitepapers*), 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"[2].

Otra definición muy significativa es la del McKinsey Global Institute[3], 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", acerca de cuan "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 crece. 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 sean los tamaños comunes de los conjuntos de datos en ese sector o industria. Teniendo presente estas consideraciones, los *Big Data*, en muchos sectores, hoy en día variarán desde decenas de terabytes a petabytes, y ya casi exabytes, camino a zettabytes.

Otra fuente de referencia es la consultora tecnológica IDC[4] 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"[5].

Otra definición muy acreditada, por venir de la mano de la consultora Gartner, es: "*Big Data* son los grandes conjuntos de datos que tienen 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)"[6]. 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"[7].

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, la combinan 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.

5.3 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**.

5.3.1 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, etcétera. 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 (normalmente, en hojas de cálculo). 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, etcétera, 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 tienen para manejar datos no estructurados.

> Los datos estructurados son aquellos 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.

5.3.2 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 relacionales. 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.

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, Viber, Telegram, Snapchat, Skype, Messenger Facebook, Google Allo y 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.

5.3.3 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 *blog*s 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 logs* 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, etcétera; 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, etcétera, pero su contenido normalmente no está estructurado.

Los datos semiestructurados son aquellos 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.

5.4 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[8]. 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, logs, 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?

5.4.1 MODELO 3V DE *BIG DATA*

En 2001, DougLaney —analista de META Group, hoy Gartner[9]— definía el crecimiento constante de datos como una oportunidad y un reto para investigar el volumen, velocidad y variedad. Posteriormente, Mark Beyer[10], 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 V³): **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 a añadir hasta siete u ocho V, como veremos más adelante.

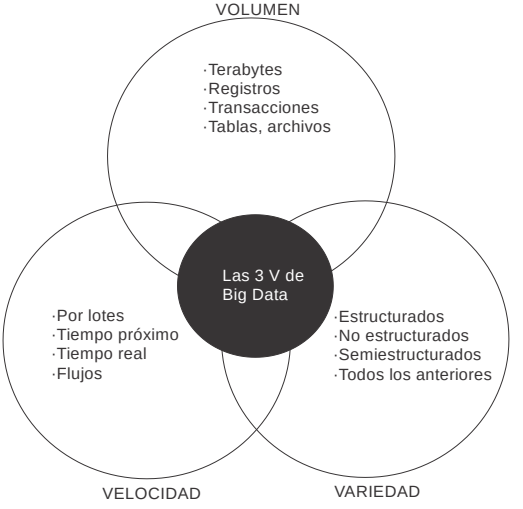

Figura 5.1. Las 3 V de *Big Data*
Fuente: Philip Russom: *Big Data Analytics*, en Teradata, Fourth Quarter 2011. Disponible en: <http://tdwi.org/blogs/philip-russom>

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). Sólo Twitter genera más de 9 terabytes (TB) de datos cada día; Facebook genera 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 en 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 de 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 para 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.

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, lo que incluirá tanto los datos tradicionales como los no tradicionales[11]. Por citar unos ejemplos, el video y las imágenes no se almacenan

de manera fácil ni eficaz 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.

5.4.2 EL MODELO DE LAS 5 V

IBM añadió una cuarta V y, posteriormente, una quinta V[12]. 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*[13], donde añade dos nuevas propiedades: **veracidad** y **valor**, que luego incluiría en su libro de *Big Data*, publicado en 2015.

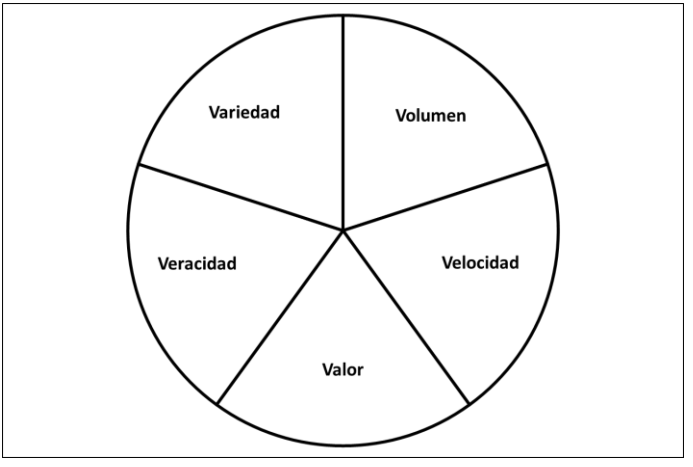

Figura 5.2. Las 5 V de *Big Data*

Veracidad

IBM define la característica de veracidad como "la incertidumbre de los datos"[14]. La veracidad hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos. 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, 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 lenguaje 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 cómo 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 *clúster* de centenares, o incluso millares, de computadoras 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"[15].

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.

5.4.3 EL MODELO DE LAS 7 V

A las cinco características anteriores se están uniendo, según algunos modelos de *Big Data*, dos nuevas e importantes V 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

5.4.4 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", y muchos almacenes de datos de empresas han roto la barrera del petabyte. En la feria CES 2017 de Las Vegas, se presentaron modelos de memoria externa (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, mientras que más del 20% por las empresas. IDC[16] 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, que es lo más importante. Es decir, el análisis de los grandes datos que realice su organización, combinado con las acciones que se tomen para mejorar su negocio, será 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.

5.5 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, con su uso restringido 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[17]—en un trabajo de investigación realizado sobre el origen e implantación del término *Big Data*, y publicado en 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 *smallcode, Big Data*. En 1989, y por último en 1993, se habla de *Big Data applications*.

Por último, Diebold menciona un trabajo de Laney (2001)[18] 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) destacan que comienza a ser utilizado en dos grandes disciplinas: Ciencias de la Computación (Informática) y Estadística/Econometría.

Diebold considera que el término Big Data nació a mediados de los años 90 y su autor John Maskey que trabajaba en Silicon Graphics Inc. (SGI) y, posteriormente fue utilizado por Weiss e Indurkhya en el área de computación y en el año 2000 por el propio Diebold en el área de estadística y econometría. En resumen, concluye Diebold, que el término se puede atribuir razonablemente a Marsey, Weiss e Indurkhya, Diebold y Laney.

5.5.1 EL ORIGEN MODERNO DE BIG DATA

En 2008, Steve Lohr[19], en *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 hemos seguido los avatares de *Big Data*.

Pero, sin duda, el detonante de la explosión de *Big Data* es el artículo que *Wired*[20] publicó en junio del mismo año; 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).*[21]"

En ese mismo número, Chris Anderson[22], su director y 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 *array*s 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 (*whitepaper*): *Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society*[23].

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 *dossier* especial: *Big Data Special Issue* en la revista *Significance,* publicación conjunta de la American Statistical Association y la Royal Statistical Society[24].

5.6 FUENTES DE DATOS

El gran volumen de datos procede de numerosas fuentes, especialmente de las nuevas fuentes, tales 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)[25], citando fuentes oficiales de estadística de los Estados Unidos, numerosas empresas de todos los sectores tenían almacenados, ya en 2009, al menos 100 terabytes, y muchas podían llegar a tener más de 1 petabyte. Algunos datos ilustrativos de almacenamiento utilizadas, en esas fechas, por sectores, eran: 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 ya hemos comentado, 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.

5.6.1 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, como se muestra en la figura 5.4 (Soares, 2012). 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.

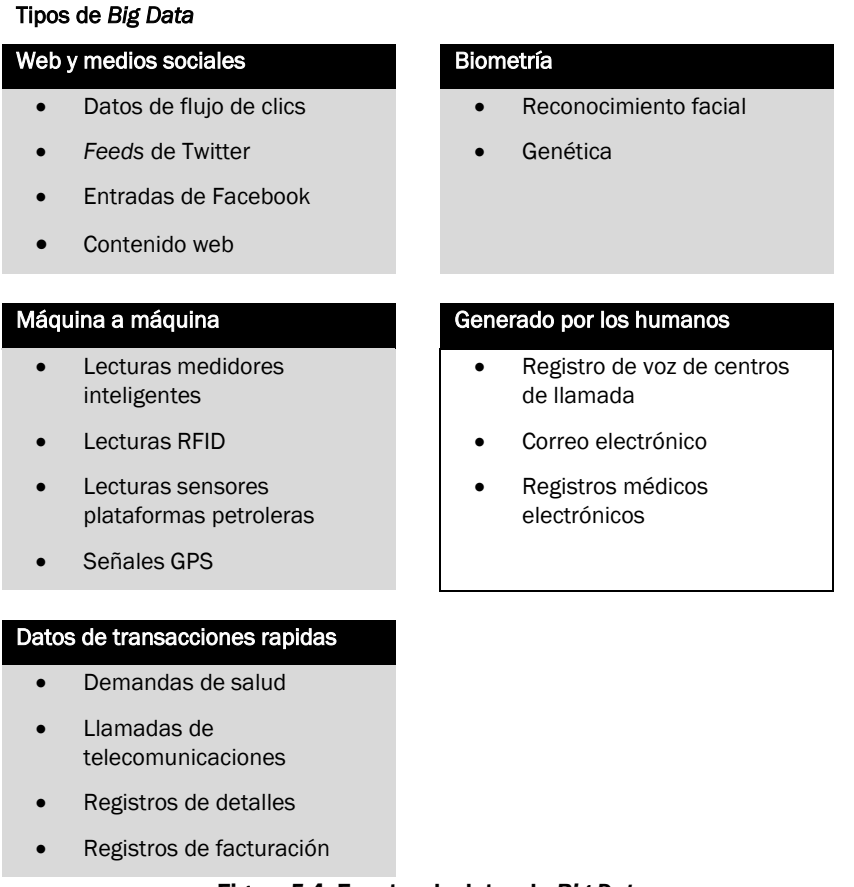

Figura 5.4. Fuentes de datos de *Big Data*
Fuente: Soares (2012)[26] (adaptada).

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 y 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 e industria). Estos datos se transmiten a través de redes cableadas, inalámbricas, móviles, y satelizadas a otros dispositivos y 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**.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[27] 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, lo que incluye huellas digitales, iris, escaneo de la retina, reconocimiento facial, genética, ADN, reconocimiento de voz e, 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 de 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 o utilizar Internet y los diferentes sitios que visitamos, tales como páginas de periódicos, revistas o *blog*s, el uso de las redes sociales, etcétera. 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, búsquedas que realizamos, etcétera).

5.7 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 **datificación**. 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.

5.8 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.

5.8.1 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 2016: 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 sea realizado por sus empleados. El uso, cada día más frecuente, de los robots conversacionales (*chatbots*) hace que una gran parte de la empresa pueda automatizarse, sobre todo con los relacionados con la gestión y administración de los datos de los clientes en sus comunicaciones con la empresa.

5.8.2 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 2016:55-66). Algunos ejemplos de datos externos citados por Marr (2016: 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 y otras aplicaciones de Google.

5.9 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 y procesamiento de datos.
- Análisis de datos.
- Visualización de datos (presentación resultados).

5.9.1 IDENTIFICACIÓN DE LAS FUENTES DE DATOS

Previamente a la recolección de datos, se requiere de una etapa de **identificación de las fuentes de datos,** 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, tal como vimos en apartados anteriores. Los

aspectos que se han de considerar en la identificación de las fuentes de datos son:

- Identificar las fuentes internas y externas.
- Calcular la cantidad de datos detectada (por 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*

5.9.2 RECOLECCIÓN (INGESTA) DE DATOS

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), en los cuales se buscan cambios realizados desde la última transacción de datos. Un ejemplo de recolección de datos por lotes es la migración de datos en un periodo determinado —una hora o un día— de una base de datos a otra.
- **Tiempo real** *(streaming).* Este tipo de recolección trata directamente con la fuente de información de manera continua, de modo 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. Están conectados de modo continuo a las fuentes de datos, descargando la información cada vez que se transmite. Existen herramientas que pueden manejar los dos tipos de recolección de datos de modo independiente o, incluso, de modo híbrido, por lotes o flujo continuo, según se requiera.

5.9.3 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. En los sistemas de almacenamiento de Big Data, el soporte son los archivos distribuidos. En esta fase se tratan sistemas de archivos tradicionales y distribuidos, bases de datos relacionales SQL y, sobre todo, bases de datos NoSQL y "en memoria".

Entre otras características, los nuevos sistemas de almacenamiento deben ser **escalables**, precisamente debido a los grandes volúmenes de datos que necesitan las compañías, y de cualquier tipo, los cuales tienen que procesarse de acuerdo con 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 y 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, integrará 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).

5.9.4 PROCESAMIENTO DE DATOS

Esta etapa suele considerarse como una o dos etapas, según la metodología utilizada (procesamiento y análisis de datos; en nuestro caso separaremos las dos etapas). 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 los ya citados: *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.

MapReduce es el algoritmo de programación para la manipulación de grandes volúmenes de datos en sistemas distribuidos y están diseñados para tratar archivos de gran tamaño (gigabytes, terabytes e, incluso, petabytes). Es soporte fundamental para la manipulación de datos no estructurados, ya que funciona a nivel de sistemas de archivos.

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.

5.9.5 ANÁLISIS DE DATOS

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.

5.9.6 VISUALIZACIÓN DE DATOS

Los resultados del análisis de datos es la etapa de consumo de datos 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 7 "Visualización de Datos" trataremos detalladamente las herramientas gráficas más recomendadas, cuya selección es un criterio importante a 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, QlickView, CartoDB, D3.js, etcétera. 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.*[28]

5.9.7 PLATAFORMAS Y HERRAMIENTAS DE BIG DATA

Las herramientas de *Big Data* que se utilizan en las diferentes capas de la arquitectura de *Big Data* son muy numerosas y se pueden encontrar como soluciones independientes o integradas en paquetes (*suites*) de proveedores de *software* propietario o de *software* abierto (*open source*). Sin embargo, y dada la complejidad de las herramientas, son numerosas las plataformas de *software* que han ido naciendo (en algunos casos, asociadas a plataformas de *hardware*), que se pueden usar en todas las etapas del ciclo de vida de *Big Data* o de modo individual en una o varias etapas.

La infraestructura o plataforma más empleada en el tratamiento de *Big Data* es **Hadoop**, un marco de trabajo (*framework*) de código abierto (*open source*), pertenenciente y gestionada por la comunidad de Apache Software Foundation. En los dos últimos años, ha adquirido una gran popularidad la plataforma **Spark**, también de la fundación Apache, que ha venido a resolver determinadas limitaciones que presenta la plataforma Hadoop, sobre todo, cuando se realiza tratamiento de datos en tiempo real, También han surgido otras plataformas no ligadas a la Fundación Apache y que comienzan a ser utilizadas en el proceso de *Big Data*. La herramienta más introducida en la actualidad es **Lambda.**

La rentabilidad de un proyecto de *Big Data* para una empresa, y su integración en sus sistemas de Inteligencia de Negocios, requiere del uso adecuado de herramientas en cada una de las capas del desarrollo de su arquitectura, con el objeto de desarrollar soluciones que puedan obtener los mejores resultados para tomas de decisiones eficaces y eficientes. Así, para conseguir unas presentaciones de resultados mediante las adecuadas herramientas de visualización, requieren que, previamente, se hayan utilizado herramientas idóneas en las etapas anteriores, desde la ingesta (recolección) de datos, pasando por el almacenamiento, procesamiento y análisis de datos.

Las plataformas de código abierto Hadoop y Spark tienen herramientas que abarcan todas las etapas del ciclo de vida de *Big Data*, aunque existen otras soluciones independientes, como es el caso del tratamiento de bases de datos, donde, además de los sistemas de archivos HDFS de Hadoop, se utilizan con grandes resultados las bases de datos NoSQL y "en memoria", así como la integración de bases de datos relacionales con no relacionales NoSQL. En la etapa final de visualización de datos (resultados), también existen numerosas soluciones de mercado —tanto propietarias como de código abierto— que no están contenidas en la plataforma Hadoop (o Spark), pero incluyen conectores (programas de enlace o interfaces) para su integración con dichas plataformas. Este es el caso de herramientas de visualización como Tableau, QlikView, Gephi, NodeXL, etcétera.

No obstante, la infraestructura más utilizada, bien de modo independiente o integrada en plataformas, es Hadoop y, por su importancia, le dedicaremos un apartado completo para tratar en profundidad sus diferentes herramientas.

El proyecto Apache Hadoop, de la Apache Software Foundation, es una gran biblioteca de *software*, que consituye un marco de trabajo (*framework*) de desarrollo de *software* de código o fuente abierta *(open source*), que está diseñado para permitir el procesamiento distribuído de grandes conjuntos de datos mediante *clústeres* de computadoras (conjunto voluminoso de servidores o nodos), utilizando modelos de programación sencillos de realizar.

Hadoop ha sido desarrollado por la Fundación Apache como un sistema de procesamiento paralelo y distribuido de grandes datos. La plataforma Hadoop ofrece una amplia variedad de herramientas para ayudar a ejecutar un gran número de funcionalidades requeridas para el análisis de *Big Data*. El desarrollo original de Hadoop se inspira en dos herramientas muy populares e innovadoras de Google, y cuyo código dejó abierto, MapReduce y GFS (GFS publicado en octubre de 2013 y MapReduce publicado en diciembre de 2004). Google File System (GFS) es un nuevo sistema de archivos distribuIdo y MapReduce es un algoritmo o modelo de programación para manejar y gestionar grandes volúmenes de datos, que permite el procesamiento de grandes volúmenes de datos (ambos componentes fueron la base de la creación de Google).

Hadoop, el sistema de código abierto de Apache, se apoya, fundamentalmente, en MapReduce y en el sistema de almacenamiento de archivos distribuIdos HDFS (evolución del sistema GFS de Google).

5.10 ECOSISTEMA HADOOP

En el sitio web oficial de la Fundación Apache[29], y tras el saludo inicial *Welcome to Apache™ Hadoop®!,* aparece la pregunta *What Is Apache Hadoop?*,cuya respuesta describimos a continuación:

"El proyecto Apache Hadoop desarrolla un software de fuente abierta de computación distribuida, escalable y fiable. La biblioteca de *software* Apache Hadoop es un marco de trabajo *(framework)* que permmite el procesamiento distribuido de grandes conjuntos de datos a través de *clusters* de computadorasutilizando modelos de programación sencillos. Está diseñado para escalar desde simples servidores a miles de máquinas, cada una ofreciendo computación local y almacenamiento. En lugar de depender del *hardware* para entregar alta disponiblidad, la biblioteca por sí misma, está diseñada para detectar y manipular fallos en la capa de aplicación de modo que entrega un servicio de alta disponibilidad sobre la parte superior de un *cluster* de computadoras cada una de las cuales puede ser propensa a fallos*"*.

De la definición anterior de Hadoop, y del análisis de las aplicaciones más comunes, se puede deducir sus características más notables:

- **Código abierto**. El desarrollo de Hadoop está a cargo de la comunidad gobernada por la licencia de la Apache Software Foundation. Se puede mejorar Hadoop mediante la adición de características, corrección de fallos de *software*, mejora del rendimiento o de la escalabilidad.
- **Almacenamiento y procesamiento distribuido**. Hadoop es un marco de trabajo que permite el proceso distribuido de grandes conjuntos de datos entre *clústeres* (agrupamiento de servidores) de computación. Los grandes volúmenes de datos se descomponen automáticamente en trozos más pequeños (bloques) a través de los nodos del *clúster*. Cada máquina procesa su bloque de datos localmente, lo que significa que el procesamiento se distribuye, a la vez, mediante cientos de CPU y gigabytes de memoria.
- **Escalabilidad**. Está diseñado para escalar desde un servidor a miles de máquinas, cada una ofreciendo capacidad de cálculo y almacenamiento.
- **Tiene tolerancia a fallos**. Los servicios de Hadoop son tolerantes a fallos por redundancia.
- ***Hardware* genérico**. Todos los servicos de Hadoop se pueden ejecutar en *hardware* genérico (*commodity hardware*). De esta forma, se reducen los costes de implementación, de soporte y de mantenimiento.
- **Incluye cuatro módulos y un gran número de aplicaciones en la biblioteca de *software***: Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN y MapReduce.

"En la página oficial de la fundación Apache (hadoop.apache.org) se describe el proyecto Hadoop con la última versión (*release*) disponible de Apache Hadoop. El proyecto Hadoop contiene cuatro módulos y un conjunto de proyectos relacionados. Los cuatro módulos que componen la infraestructura Hadoop según la página oficial son los siguientes:"

- **Hadoop Common**. Utilidades prácticas (comunes) que soportan los otros módulos de Hadoop.
- **Hadoop Distributed File System *(HDFS)***.Un sistema de archivos distribuido que proporciona acceso de alto rendimiento a los datos de la aplicación.

- **Hadoop YARN**. Un marco de trabajo para planificación de tareas y gestión de recursos del *clúster*. Gestiona acceso a recursos de las aplicaciones (memoria, CPU), monitoriza nodos y soporta distintos tipos de aplicaciones, no sólo MapReduce.
- **MapReduce.** Sistema basado en YARN para procesamiento en paralelo de grandes volúmenes de datos. Es un modelo de programación basado en el paradigma "divide y vencerás".

Hadoop v1 (1.0)

La versión original de Hadoop contiene dos módulos (en la figura 5.6 se muestran los ecosistemas Hadoop v1. y v2. con algunos proyectos relacionados)

HDFS. Sistema de almacenamiento propio de Hadoop. Es un sistema de almacenamiento distribuido de archivos, que oprtimiza el tratamiento de la información.

MapReduce. Algoritmo que reduce los grandes volúmenes de información a un volumen muy inferior, para poder realizar consultas mucho más rápidas y eficientes. Esta herramienta permite que se realicen los análisis a una mayor velocidad.

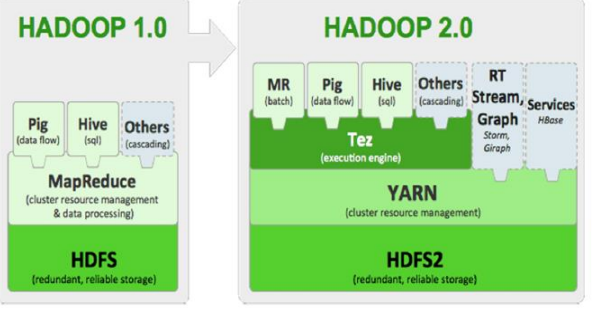

Figura 5.6. Ecosistema Hadoop (versiones 1.0 y 2.0)

5.11 HERRAMIENTAS MÁS UTILIZADAS DE HADOOP EN *BIG DATA*

Las herramientas del ecositema Hadoop más utilizadas en las etapas de un sistema de *Big Data* son las siguientes:

Adquisición de datos (Ingestión)

Se han de utilizar comandos de Hadoop para poder realizar las lecturas y posterior almacenamiento en los archivos HDFS. Las herramientas más empleadas son:

- Flume (es una de las mejores herramientas para recolección de datos. Una aplicación práctica: Mediante aplicaciones API REST que proporciona la red social Twitter, puede conectarse directamente para capturar datos de esta red social.
- Chukwa.
- Scribe.
- Sqoop.
- Kafka.

Almacenamiento de datos

Sistema HDFS. Bases de datos HBase, Cassandra y Hive.

Procesamiento y análisis de datos

Los componentes de Hadoop fundamentales para procesamiento de datos son: MapReduce y Spark. Herramientas de análisis: Hive, Pig y Avro.

Visualización de datos

Tableau. Es una herramienta de *software* propietario con versiones de uso gratuitas. Tiene una interfaz muy sencilla e intuitiva; ofrece gráficos estadísticos muy buenos sobre los datos analizados (ver capítulo 7). Conecta muy bien con numerosas plataformas de Hadoop, tal como Cloudera.

Administración

Zooquiper (sincronización del clúster), Avro (serialización de datos) y Hue (interfaz Web)

El sistema de programación MapReduce y YARN, con las herramientas

Hive es un software que facilita la búsqueda y gestión de grandes cantidades de datos alojados en unalmacén distribuido como estructura de datos. Dispone de un lenguaje similar a SAL, HiveQL que da soporte al almacenamiento.

Pig
Cascading
Spark
Spark SQL
Mahout
Oozie

5.11.1 CATÁLOGO DE HERRAMIENTAS DE HADOOP MÁS EMPLEADAS EN INTELIGENCIA DE NEGOCIOS

Las herramientas de Hadoop se utilizan, en su mayoría, en la parte superior de la infraestructura de Hadoop y permiten gestionar grandes volúmenes de datos que antes sólo podían realizar supercomputadoras y con un coste muy alto.

Breve descripción de herramientas Hadoop

Ambari. Aprovisionamiento, gestión (administración) y monitorización de clústeres de Hadoop.
Avro. Sistema de señalización de datos.
Chukwa. Monitorización y recolección de datos.
Flume. Recolector de fuentes de datos (recolección de *log*s, mensajes, etcétera).
Hama. Marco de trabajo de computación científica.
HBase. Base de datos no relacional (NoSQL) de Hadoop, modelo columnar. Está basada en BigTable de Google y puede gestionar tablas masivas de datos que combinan millones y miles de millones de filas y columnas.
Hive.*Software* de almacén de datos (*Data Warehouse*) de Hadoop. Lenguaje SQL Query, denominado HiveQL. Facilita la gestión de datos masivos que residen en el almacenamiento distribuido de archivos HDFS.
Hue. Interfaz web de Hadoop para análisis de datos.
Mahout. Plataforma escalable de aprendizaje automático.
MapReduce. Algoritmo utilizado para la programación y control de Hadoop.
Nutch.*Crawler* de Hadoop.
Oozie. Planificador de flujo de trabajo. Gestiona todas las tareas de Apache Hadoop. Es muy escalable y fiable y bastante extensible.
Pig. Lenguaje de alto nivel (*scripting*) para análisis de datos. Es una plataforma de alto nivel para crear programas MapReduce usando Hadoop. Se puede incorporar Pig en muchos lenguajes diferentes como JRuby, Python y Java, y, a la inversa, se pueden tener secuencias de comandos Pig en otros lenguajes.
Soir. Plataforma de búsqueda.
Sqoop. Herramienta de transferencia de datos en bruto. Lenguaje SQL para Hadoop. Importa bases de datos completas o tablas para el sistema de archivos HDFS y genera clases en Java que permiten interactuar e importar datos.
Storm. Sistema de computación distribuido en tiempo real.
Yarn. Soporte de MapReduce en Hadoop v2.
Zookeeper. Sistema de configuración centralizado de Hadoop.

5.11.2 FUNCIONALIDADES IMPORTANTES EN LA GESTIÓN MASIVA DE ARCHIVOS DE HADOOP

Sistemas de archivos distribuidos. HDFS, sistemas de archivos FTP, Windows Azure Storage, Amazon S".

Gestión de recursos distribuidos. Marco de trabajo YARN de proceso distribuido

SQL Query. Hive (componente de *software* de *Data Warehouse*).

Aprendizaje automático. Mahout (componente de aprendizaje automático).

Procesamiento de flujo. Storm (motor computacional de tiempo real).

Procesamiento de gráficos. Giraph. Un *framework* para procesamiento de gráficos a gran escala.

Gestión de interfaces. Zookeeper. Herramientas de gestión para clúster de Hadoop.

Conectividad con bases de datos relacionales. Sqoop (realiza transferencia de datos entre sistemas de bases de datos relacionales y Hadoop).

Procesamiento de flujo de datos. Pig (lenguaje de alto nivel de flujo de datos de *scripting*).

Herramientas de flujo de datos. Flume (servicio para transferencia eficiente en *streaming* de datos en el sistema de archivos HDFS).

Bases de datos NoSQL. HBase basada en BigTable y orientado a columnas (almacenamiento columnar).

Aplicaciones de estadística. R Connectors.

Recolección de datos basados en la Web. Nutch, Solr, Gora y HBase.

Programación. MapReduce (con lenguajes Java, Pig, Perl, Hive).

Planificación. Oozie.

Movimiento y flujo de datos. Comandos Hadoop, Sqoop. Flume y Storm.

***Monitorización*.** Hue, Nagios, Gangha.

Administración de clústeres. Ambari, CDH.

Análisis de SQL. Impala, Hive y Spark.

ETL. Talend y Pentaho.

Reporting. Splunk, Talend.

En la citada página oficial de Hadoop, se describen los siguientes proyectos (si se navega por el sitio web, se puede encontrar gran documentación sobre los innumerables proyectos o aplicaciones existentes):

Ambari. Herramienta basada en la Web para provisionar, gestionar y monitorizar clústeres de Apache Hadoop, que incluyen soporte para el sistema de archivos HDSFS y MapReduce, además Hive, HCatalog, HBase, Zookeeper, Oozie, Pig y Sqoop. Ambari proporciona también un cuadro de mando para visualizar la salud del clúster, tal como mapa de calor (*heatmaps*) y capacidad para visualizar MapReduce, Pig y aplicaciones Hive viusalmente junto con características para diagnosticar sus características de rendimiento de un modo amistoso al usuario.

Avro. Sistema de serialización de datos.

Cassandra. Base de datos multimaestra escalable para gestión de sistemas distribuidos.

Chukwa. Sistema de recolección de datos para la gestión de grandes sistemas distribuidos.

HBase. Base de datos distribuida, escalable, que soporta almacenamiento de datos estructurados para grandes tablas.

Hive. Una infraestructura de *Data Warehouse* que proporciona sumarización de datos y consultas *adhoc* (*querying*).

Mahout. Aprendizaje automático escalable y biblioteca de Minería de Datos.

Pig. Un lenguaje natural de alto nivel de flujo de datos y ejecución de marco de trabajo para computación paralela.

Spark. Un motor de cómputo general rápido para datos de Hadoop. Spark proporciona un modelo de programación rápido, sencillo y expresivo que soporta una amplia gama de aplicaciones, incluyendo ETL, aprendizaje automático, procesamiento de flujo (*stream*) y computación gráfica.

Tez. Un *framework* de programación de flujo de datos generalizado, construido en Hadoop YARN, que proporciona un motor potente y flexible para ejecutar un DAG arbitrario de tareas para procesar datos de casos de uso por lotes e interactivos. Tez está siendo adoptado por Hive, Pig y otros marcos de trabjo en el sistema Hadoop.

Zookeeper. Servicio de coordinación de alto rendimiento para aplicaciones distribuidas.

5.11.3 Distribuciones comerciales de Hadoop

Hadoop es el marco de trabajo *open source* para el desarrollo de *Big Data* y *Analytics*, tal como se ha mencionado en apartados anteriores. Proporciona soluciones para almacenamiento de datos de la empresa y soluciones de analítica con alta escalabilidad (prácticamente ilimitada). Desde su lanzamiento en 2011, ha crecido rápidamente en popularidad y ha emergido un gran ecosistema de distribuidores, vendedores y consultores con el objetivo de dar soporte a la industria.

Hadoop es un sistema *open source* (gratuito), disponible para cualquiera que lo desee utilizar. Sin embargo, las empresas necesitan alinear las soluciones de Hadoop con sus necesidades para el desarrollo de las soluciones específicas para ellas. Por estas razones, las distribuciones comerciales vienen empaquetadas para resolver las necesidades de gestión de datos y soluciones de analítica. ¿Cómo elegir, entonces, una distribución comercial? Bernard Marr[30], ya

citado como uno de los grandes expertos mundiales en *Big Data*, en un excelente artículo sobre la cuestión, describe las plataformas comerciales que considera de mayor impacto en las empresas y que contienen soluciones tecnológicas y más económicas. Un factor muy importante que analiza previamente Marr, antes de señalar las plataformas más recomendables, es la necesidad por parte de la empresa de seleccionar si se desea una solución *on premise* (en la empresa) o una solución en la nube. A continuación, se detalla una selección de distribuidores comerciales extraída del informe de Marr y de las consultoras Gartner y Forrester:

Cloudera. Fue la primera distribución comercial del mercado. Tiene como principal característica, la figura de su arquitecto jefe, Dough Cuttings, considerado uno de los creadores de Hadoop. Su producto principal Cloudera Enterprise es el más comercializado. La plataforma se ofrece en código abierto y una ventaja importante es el programa de formación y de certificaciones profesionales, uno de los más reconocidos a nivel internacional.

Amazon Elastic MapReduce. La solución de Amazon ofrece una plataforma de Hadoop-as-a-Sevice (HaaS), a través del prestigioso servicio AWS (Amazon Web Service). Sus grandes ventajas son el pago por uso y su alta capacidad de almacenamiento y escalabilidad.

Hortonworks. Es un modelo muy innovador con una arquitectura propia, fiable y muy eficiente.

MapR Technologies. Se enfoca en ofrecer el máximo rendimiento y tolerancia a fallos, aprovechando el potencial de Hadoop para trabajar a gran escala con el menor esfuerzo. Es el distribuidor de Hadoop que mayor empeño ha hecho en hacer fiables y eficientes las mayores implementaciones de clústeres Hadoop. MapR tiene una arquitectura distinta, con un enfoque más distribuido que se traduce en mejores rendimientos.

Pivotal. Es una solución muy innovadora, cuyos productos han sido seleccionados por consultoras como Gartner y Forrester.

Altiscale. Esta empresa fue adquirida por SAP en el año 2016. Ofrece una solución basada en la nube, como Amazon y HaaS. Al contar con el soporte de SAP, ofrece servicios operacionales adicionales que ofrecen soluciones de seguridad, escalado, rendimiento, etcétera, integradas con el marco de trabajo Hadoop.

Además de estas soluciones anteriores basadas en código abierto, los grandes distribuidores de software comercial ofrecen soluciones también para la integración de Hadoop con sus propias soluciones. Este es el caso de IBM, Oracle y Microsoft, además de SAP y otros proveedores como SAS y Microstrategy.

5.12 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**), y se referíaa 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 es 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 en 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 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.

El W3C está impulsando en todo el mundo el movimiento a favor de la apertura de datos públicos. La 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 durante toda la década, sobre todo por el apoyo ofrecido por el gobierno de los Estados Unidos(`http://www.data.gov`);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*[31,32] 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, etcétera). 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.

Enla 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 ques on 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. Los datos abiertos son muy aprovechables y generan valor añadido a las empresas. Enel 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: **transparencia**, porque se puede consultar y tratar datos que vienen directamente de las fuentes oficiales; **eficiencia**, porque ciudadanos y organizaciones pueden crear servicios en forma más ajustada en colaboración con la administración; **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.

5.13 INICIATIVAS E ÍNDICES INTERNACIONALES DE *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, en Gran Bretaña y 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.

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, 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.

Open Data Center Alliance

La asociación internacional Open Data Center Alliance (ODCA, `http://www.opendatacenteralliance.org`) es una organización sin fines 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***, según la Alianza, un futuro basado en estándares abiertos e interoperables. 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[33], que fue publicada en el año 2012, coincidiendo con el 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 fines 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 en 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, se encuentra el 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 (Overseas Development Institute) define los **datos abiertos**[34] 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 ella, 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*.

5.13.1 ÍNDICES INTERNACIONALES DE DATOS ABIERTOS

Los índices y barómetros de *Open Data* pretenden reflejar el *ranking* internacional de los países en su aplicación de políticas de datos abiertos. El día 4 de marzo se celebra en todo el mundo el Día Mundial de *Open Data*. Con ocasión de esa

festividad, y alrededor de esa fecha o meses siguientes, se suelen publicar los índices de mayor reputación internacional por las organizaciones más prestigiosas del mundo en *Open Data*, en los que se valoran diferentes aspectos relacionados con los datos abiertos.

El portal *Iniciativa de datos abiertos del Gobierno de España* publicó[35], en junio de 2017, un informe sobre los cuatro índices globales más relevantes del mundo de *Open Data*:

1. ***Open Data Barometer (ODB)*[36]: World Wide Web Foundation**
   El ODB es una medida global de cómo los gobiernos están publicando y utilizando datos abiertos de responsabilidad, innovación e impacto social. Se publica con carácter anual por la World Wide Foundation desde el año 2013 (el índice de 2017 fue publicado el 1 de junio de 2017).

   El Barómetro *Open Data* evalúa las políticas y prácticas de datos abiertos desarrolladas en 115 países de todo el mundo. Mide cómo los gobiernos ponen sus datos a disposición de sus ciudadanos a partir de tres fuentes principales de información: cuestionarios de evaluación completados por expertos, auto-evaluaciones por parte de los gobiernos y otros datos provenientes de bases de datos internacionales.

2. **Informe anual[37]: Portal Europeo de Datos Públicos (European Data Portal)**
   Se publicó el 4 de marzo de 2017 (segunda edición anual). El portal evalúa **a 31 países** (los 28 miembros de la Unión Europea junto a Noruega, Suiza y Liechtenstein) y el informe analiza el nivel de **madurez del ecosistema de datos abiertos en Europa**. Según este informe, el estado de los datos abiertos en toda Europa ha mejorado en términos globales, pero destaca también discrepancias significativas entre los países, así como barreras políticas, legales y técnicas.

3. **Global Open Data Index (GDDI)[38]: Open Knowledge Foundation**
   El índice GODI (Global Open Data Index) es elaborado por la Open Knowledge Foundation. El 15 de junio de 2017, OKF publicó su índice global de datos abiertos de 2017, en el que analiza 94 estados y territorios en el mundo. El índice mide el nivel de apertura de diez categorías de datos y se basa en su calificación en las respuestas de un cuestionario que incluye asuntos sobre el formato, el nivel de apertura, la actualización o periodicidad de publicación de los datos. La puntuación permite crear un *ranking* final y visualizar el nivel de apertura de cada conjunto nacional de datos. Es una iniciativa de carácter ***colaborativo.***

4. **OUR DataIndex[39]: OCDE**
   El Índice OUR Data (Open, Useful, Re-usable Government Data) analiza la situación de datos abiertos de 28 países de la OCDE (Organización para la Cooperación y el Desarrollo Económico), y evalúa los esfuerzos de los gobiernos por implementar datos abiertos en tres áreas críticas: apertura,

utilidad y reutilización de los datos gubernamentales. El índice se centra en ver los esfuerzos del gobierno para asegurar la accesibilidad y disponibilidad de los datos del sector público, para estimular una mayor reutilización. El índice se basa en una metodología propia de la OCDE y el análisis incluye: información de negocios, registros, información sobre patentes y marcas, bases de datos de licitaciones públicas, información geográfica, información legal, información meteorológica, datos sociales e información de transporte.

5.14 RESUMEN

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

- 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.
- La integración de los datos tradicionales con los *Big Data* supone una gran oportunidad de negocio para organizaciones y empresas.
- La explosión de *Big Data* se ha producido en los últimos años por las innumerables fuentes de datos que han ido proliferando desde los datos textuales y no textuales, de contenidos de audio, fotografía y video, datos de teléfonos inteligentes y tabletas, de los *socialmedia*, sensores, etcétera.
- Los Big Data no constituyen una amenaza como tal, sino más bien un reto y una oportunidad para organizaciones y empresas.
- 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* (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, etcétera.
- Los datos abiertos (*Open Data*) se refieren a los datos públicos y privados que deberían estar a disposición de los ciudadanos y de las empresas para un uso eficaz y rentable. Naturalmente, los datos abiertos deberán respetar siempre la privacidad y la información que deba estar protegida, comodatos de salud y personales, pero se requiere que se abran y que sean interoperables por las distintas plataforma sutilizadas por los desarrolladores; deben ser, también, legibles y entendibles por los ciudadanos.
- 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, aunque casi todos los países latinoamericanos y caribeños tienen en la actualidad iniciativas oficiales de *Open Data*.
- En Europa, diferentes países, y en España, diferentes comunidades autónomas, han puesto en marcha iniciativas de *Open Data.*

NOTAS

[1] Bill Franks. (2012). Taming the Big Data Tidal Wave. Finding Opportunities in Huge Data Streams with advanced Analytics. New York. Wiley

[2] Adrian Merv: "Big Data", en Teradata Magazine, 2011 Q1. Disponible en:

<http://ww.nxtbook.com/nxtbooks/mspcomm/teradata_2011q1/index.php?startid=8#/40>.

[3] *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>.

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

[5] *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>.

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

[7] *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>.

[8] *Sitio de IBM de big data: "Bringing big data to the enterprise". Disponible:*

<http://www_01.ibm.com/software/data/bigdata>.

[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] *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.*

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

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

[13] *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 MetricsTo Make Better Decisions and Improve Performance, Wiley,2015, volvió a dar la misma definición de Big Data.*

[14] *IBM. Analytics: el uso de big data en el mundo real. IBM Institute for Business Value. 2012.* http://www-

05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf

[15] Bernard Marr. Big Data. Madrid: Editorial TEELL, 2016

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

[17] 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>.

[18] 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>.

[19] 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.*

[20] "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>.

[21] Chris Anderson. Editor jefe de Wired. "The Petabyte Age: "Because More Isn't Just More-More is Different ". Wired, junio 2008. www.wired.com/2008/06/pb-intro.

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

[23] 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>.

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

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

[26] *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.*

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

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

[29]http://hadoop.apache.org/. *Versión disponible el 8 de agosto de 2018: release 3.1.1.*

[30] *Bernard Marr, How to Choose a Commercial Hadoop Distribution in 2017, 30 noviembre, 2017.* http://data-informed.com/how-to-choose-a-commercial-hadoop-distribution-in-2017/

[31] *"Comienza el movimiento Open Data", en ComputerWorld, [Consulta: 21de mayo de 2011].*

[32] *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.*

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

[34] http://www.theodi.org/guide/what-open-data

[35] *29 de junio, 2017.* http://datos.gob.es/es/noticia/indices-internacionales-de-datos-abiertos.

[36] *1 de junio, 2017.*
http://opendatabarometer.org/?_year=2016&indicator=ODB&lang=es

[37] *4 de marzo, 2017.* https://www.europeandataportal.eu/

[38] https://index.okfn.org/;
https://blog.okfn.org/files/2017/06/FinalreportTheStateofOpenGovernmentDatain2017.pdf

[39] http://www.oecd.org/gov/digital-government/open-government-data.htm