CAPÍTULO 1
¿QUÉ ES BIG DATA?

Big Data (grandes datos, grandes volúmenes de datos o *macrodatos* como recomienda utilizar la Fundación Fundéu BBVA "Fundación del español urgente") supone la confluencia de una multitud de tendencias tecnológicas que venían madurando desde la primera década del siglo XXI, y que se han consolidado durante los años 2011 a 2013, cuando han explosionado e irrumpido 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 de su coste de conexión a Internet, medios sociales (en particular las redes sociales), Internet de las cosas, geolocalización, y de modo muy significativo la computación en la nube (*cloud computing*).

Los grandes datos o grandes volúmenes de datos han ido creciendo de modo espectacular. Durante 2011, se crearon 1,8 zettabytes de datos (1 billón de gigabytes) según la consultora IDC, y esta cifra se dobla cada dos años. Un dato significativo, Walmart, la gran cadena de almacenes de los Estados Unidos, posee bases de datos con una capacidad de 2,5 petabytes, y procesa más de un millón de transacciones cada hora. Los Big Data están brotando por todas partes y utilizándolos adecuadamente proporcionarán una gran ventaja competitiva a las organizaciones y empresas. En cambio, su ignorancia producirá grandes riesgos en las organizaciones y no las hará competitivas. Para ser competitivas en el siglo actual, como señala Franks (2012): "Es imperativo que las organizaciones persigan agresivamente la captura y análisis de estas nuevas fuentes de datos para alcanzar los conocimientos y oportunidades que ellas ofrecen".

Los profesionales del análisis de datos, los analistas de datos y científicos de datos, tienen mucho trabajo por delante y serán una de las profesiones más demandadas en el 2013 y años sucesivos.

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

DEFINICIÓN DE BIG DATA

No existe unanimidad en la definición de Big Data, aunque sí un cierto consenso en la fuerza disruptiva que suponen los grandes volúmenes de datos y la necesidad de su captura, almacenamiento y análisis. Han sido numerosos los artículos (*white papers*), informes y estudios relativos al tema aparecidos en los últimos dos años, y serán también numerosos los que aparecerán en los siguientes meses y años; por esta razón, hemos seleccionado aquellas definiciones realizadas por instituciones relevantes y con mayor impacto mediático y profesional. En general, existen diferentes aspectos donde 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 de su no adopción.

La primera definición que daremos es la de Adrian Merv, vicepresidente de la consultora Gartner, que en la revista *Teradata Magazine*, del primer trimestre de 2011, define este término como: "Big Data excede el alcance de los entornos de *hardware* de uso común y herramientas de *software* para capturar, gestionar y procesar los datos dentro de un tiempo transcurrido tolerable para su población de usuarios"[1].

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

Otra fuente de referencia es la consultora tecnológica IDC[3], que apoyándose en estudios suyos propios, 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 actualmente oscilan entre algunas decenas de terabytes hasta muchos petabytes para un conjunto de datos individual"[4].

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

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

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

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

TIPOS DE DATOS

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

nuevas categorías dentro de los tipos de datos no estructurados: *datos semiestructurados* y *datos no estructurados* propiamente dichos.

DATOS ESTRUCTURADOS

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

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

DATOS SEMIESTRUCTURADOS

Los datos semiestructurados tienen un flujo lógico y un formato que puede ser definido, pero no es fácil su comprensión por el usuario. Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos dato. La lectura de datos semiestructurados requiere el uso de reglas complejas que determinan cómo proceder después de la lectura de cada pieza de información. Un ejemplo típico de datos semiestructurados son los registros *Web log*s de las conexiones a Internet. Un *Web log* se compone de diferentes piezas de información, cada una de las cuales sirve para un propósito específico. Ejemplos típicos son el texto de etiquetas de lenguajes XML y HTML.

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

DATOS NO ESTRUCTURADOS

Los datos no estructurados son datos sin tipos predefinidos. Se almacenan como "documentos" u "objetos" sin estructura uniforme, y se tiene poco o ningún control sobre ellos. Datos de texto, video, audio, fotografía son datos no estructurados. Por ejemplo, las imágenes se clasifican por su resolución en píxeles. Datos que no tienen campos fijos; ejemplos típicos 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, Joyn, Viber, Line, WeChat, Spotbros. Al menos, el 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 (capítulos 8 y 9).

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

INTEGRACIÓN DE LOS DATOS: OPORTUNIDADES DE NEGOCIO DE LOS BIG DATA

¿Qué puede hacer una organización con Big Data? ¿Cómo puede tomar ventaja de sus grandes oportunidades? ¿Cómo puede evitar sus riesgos? Un número creciente de organizaciones les hacen frente desplegando herramientas especializadas como bases de datos de procesamiento masivamente paralelo (MPP, *Massively Parallel Processing*), sistemas de archivos distribuidos Hadoop, algoritmos MapReduce, computación en la nube. La pieza clave es la integración de datos. Es crucial para las organizaciones facilitar que los negocios accedan a todos los datos de modo que se pueden aplicar sobre ellos infraestructuras de Big Data.

La integración de datos facilita a su organización la combinación de los Big Data con los datos transaccionales tradicionales para generar valor y conseguir la mayor eficacia posible. Por esta razón uno de los aspectos más interesantes no es tanto lo que harán ellos mismos por el negocio, sino lo que se podrá conseguir para el negocio cuando se combinan con otros datos de la organización. Un buen ejemplo puede ser enriquecedor: utilizar las preferencias y

rechazos de los perfiles de los clientes en los medios sociales con el objetivo de mejorar la comercialización de destino.

El mayor valor de los Big Data puede producirse cuando se los combinan con otros datos corporativos. Colocándolos en un contexto más grande se puede conseguir que la calidad del conocimiento del negocio se incremente exponencialmente. Incluso la estrategia de Big Data dentro de la estrategia global de la compañía es mucho más rentable que tener una estrategia independiente.

Frank (2012: 22) considera que es muy importante que la organización no desarrolle una estrategia de Big Data distinta de su estrategia tradicional de datos, ya que en ese caso fallará toda la estrategia del negocio. Big Data y datos tradicionales son ambas partes de la estrategia global. Para que las organizaciones tengan éxito se necesita desarrollar una estrategia cohesiva donde los Big Data no sean un concepto distinto y autónomo. Frank (2012: 22) insiste en: "La necesidad desde el comienzo de pensar en un plan que no solo capture y analice los grandes datos por sí mismo, sino que también considera como utilizarlos en combinación con otros datos corporativos y como un componente de un enfoque holístico a los datos corporativos".

Es importante insistir en la importancia para las organizaciones de desarrollar una estrategia de Big Data que no sea distinta de su estrategia de datos tradicionales y conseguir una idónea integración de datos. Esta circunstancia es vital ya que ambos forman parte de una estrategia global, aunque los Big Data irán creciendo de modo exponencial deberán coexistir de modo híbrido con los datos tradicionales durante muchos años. Dicen las grandes consultoras de datos que los Big Data deben ser otra faceta de una buena estrategia de datos de la empresa.

Son numerosos los ejemplos que se pueden dar sobre la integración de datos de todo tipo en estrategias corporativas.

En el caso de la industria eléctrica, los datos de las redes inteligentes (*smart grids*) son una herramienta muy poderosa para compañías eléctricas, que conociendo los patrones históricos de facturación de los clientes, sus tipos de vivienda y otros indicadores, unidos con los datos proporcionados por los medidores inteligentes (*smart meters*) instalados en las viviendas pueden conseguir ahorros de coste considerables para la compañía proveedora del servicio eléctrico, y grandes reducciones del consumo eléctrico de los clientes.

Otro caso típico se da en el caso del comercio electrónico donde el análisis de los textos de los correos electrónicos, mensajes de texto SMS o de aplicaciones como WhatsApp, *chat,* se integran junto con el conocimiento de las especificaciones detalladas del producto que se está examinando; los datos de ventas relativas a esos productos, y una información histórica del producto proporcionan un gran poder al contenido de los textos citados cuando se ponen en un contexto global.

La integración de datos, mezcla de Big Data y datos tradicionales, supone una gran oportunidad de negocio para organizaciones y empresas.

CARACTERÍSTICAS DE BIG DATA

Cada día creamos 2,5 *quintillones* de bytes de datos, de forma que el 90% de los datos del mundo actual se han creado en los últimos dos años[7]. Estos datos proceden de todos los sitios: sensores utilizados para recoger información del clima, entradas (*posts*) en sitios de medios sociales, imágenes digitales, fotografías y videos, registros de transacciones comerciales y señales GPS de teléfonos celulares, por citar unas pocas referencias. Estos datos, son, según IBM, Big Data.

Big Data al igual que la nube (*cloud)* abarca diversas tecnologías. Los datos de entrada a los sistemas de Big Data pueden proceder de redes sociales, *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?

IBM plantea como también hizo Gartner que Big Data abarca tres grandes dimensiones, conocidas como el "Modelo de las tres V" (3 V o V^3): *volumen*, *velocidad* y *variedad* (*variety*). Existe un gran número de puntos de vista para visualizar y comprender la naturaleza de los datos y las plataformas de software disponibles para su explotación; la mayoría incluirá una de estas tres propiedades V en mayor o menor grado. Sin embargo, algunas fuentes contrastadas, 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 a la definición y características de los Big Data. Otras fuentes notables añaden una quinta característica, *valor*.

VOLUMEN

Las empresas amasan grandes volúmenes de datos, desde terabytes hasta petabytes. Como se verá más adelante (capítulo 3), 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 2015 a 2020, se espera entremos en la era del zettabyte. IBM da el dato de 12 terabytes para referirse a lo que crea Twitter cada día solo 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 35 zettabytes (ZB). Solo Twitter genera más de 9 terabytes (TB) de datos cada día. Facebook, 10 TB; y algunas empresas generan terabytes de datos cada hora de cada día del año. Las organizaciones se enfrentan a volúmenes masivos de datos. Las organizaciones que no conocen cómo gestionar estos datos están abrumadas por ello. Sin embargo, la tecnología existe, con la plataforma tecnológica adecuada para analizar casi todos los datos (o al menos la mayoría de ellos, mediante la identificación idónea) con el objetivo de conseguir una mejor comprensión de sus negocios, sus clientes y el *marketplace*. IBM plantea que el volumen de datos disponible en las organizaciones hoy día está en ascenso mientras que el porcentaje de datos que se analiza está en disminución.

VELOCIDAD

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

La importancia de la velocidad de los datos se une a las características de volumen y variedad, de modo que la idea de velocidad no se asocia a la tarea de crecimiento de los depósitos o almacenes de datos, sino que se aplica la definición al concepto de los datos en movimiento, es decir, la velocidad a la cual fluyen los datos. Dado que 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 de vigilancia de calles y edificios.

En los sistemas de Big Data las fuentes de datos son diversas y no suelen ser estructuras relacionales típicas. Los datos de redes sociales, de imágenes 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. Diferentes navegadores envían datos diferentes; los usuarios pueden ocultar información, pueden utilizar diferentes versiones de software, bien para comunicarse entre ellos, o para realizar compras, o leer un periódico digital. Sin embargo, los riesgos por 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 tratar 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 solo 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.

Bastante simple, *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 la variedad. Sin embargo, el éxito de una organización dependerá de su capacidad para resaltar el conocimiento de los diferentes tipos de datos disponibles en ella, que incluirá tanto los datos tradicionales como los no tradicionales[8]. Por citar un ejemplo, el video y las imágenes no se almacenan fácil ni eficazmente en una base de datos relacional, mucha información de sucesos de la vida diaria como el caso de los datos climáticos cambian dinámicamente. Por todas estas razones, las empresas deben capitalizar las oportunidades de los grandes datos, y deben ser capaces de analizar todos los tipos de datos, tanto relacionales como no relacionales: texto, datos de sensores, audio, video, transaccionales.

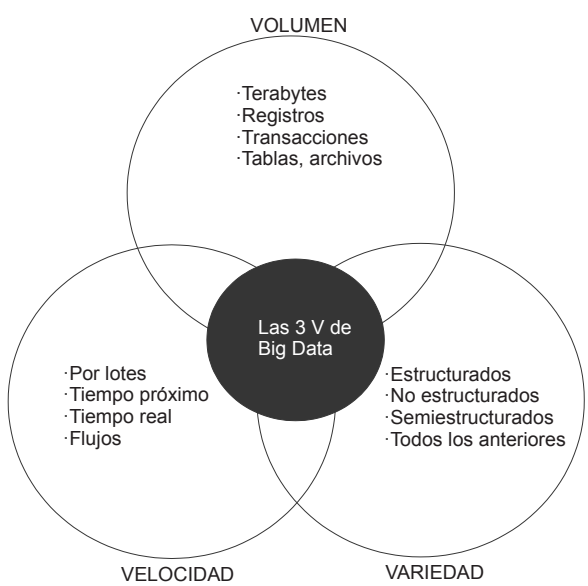

Figura 1.1. Las 3 V de Big Data. Fuente: Philip Russom: "Big Data Analytics", en *Teradata*, Fourth Quarter 2011. Disponible en: <http://tdwi.org/blogs/philip-russom>.

VERACIDAD

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

VALOR

Además de las 3 V clásicas con las que todas las fuentes coinciden, y la cuarta que suele señalar IBM, existe una quinta característica que también se suele considerar: el valor. Las organizaciones estudian obtener información de los grandes datos de una manera rentable y eficiente. Aquí es donde las tecnologías de código abierto tales como Apache Hadoop se han vuelto muy populares. Hadoop, que se estudiará más adelante en el libro, es un software que procesa grandes volúmenes de datos a través de un *cluster* de centenares, o incluso millares de computadores de un modo muy económico.

EL TAMAÑO DE LOS BIG DATA

La megatendencia de los Big Data como ya hemos considerado anteriormente, no está directamente relacionada con la cantidad específica de datos. Recordemos que hace una década los almacenes de datos (*data warehouse*) de las grandes empresas cuando tenían de 1 a 10 terabytes se consideraban enormes. Hoy se puede comprar en cualquier gran almacén, unidades de disco de 1 a 5 terabytes por precios inferiores a 100 euros (Soares, 2012), y muchos almacenes de datos de empresas han roto la barrera del petabyte.

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

Recordemos que según IDC, el universo digital de datos se dobla cada dos años, y que más del 70% de los datos creados se generarán por los consumidores; y por encima del 20% por las empresas. IDC[9] predice que el universo digital se multiplicará por un factor de 44 para llegar a 35 zettabytes en 2020.

Un informe más reciente al citado de IDC, realizado por los científicos de computación de la Universidad de California en San Diego[10], y publicado en abril de 2011, aumentaba las cifras del Universo Digital, y consideraba que los servidores de las empresas del mundo procesaban 9,57 ZB de datos en 2008 (no se contaban los 3,6 ZB que calculaba el estudio generaban los hogares de América).

Aunque el estudio de la UCSD daba cifras 10 veces superiores al estudio de IDC, en ninguno de los casos se ponía límites inferiores. Tal vez una respuesta más ajustada a la situación actual es que ni la parte *big* ni la parte *data* son la parte más importante de Big Data. Como señala Frank (2012: 6): "Ni por asomo es más importante una parte que otra. Lo importante es lo que hacen las organizaciones con los grandes datos; es lo más importante. El análisis de los grandes datos que realice su organización combinado con las acciones que se tomen para mejorar su negocio es lo realmente importante".

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

¿CÓMO SE HA LLEGADO A LA EXPLOSIÓN DE BIG DATA?

Big Data supone la confluencia de tendencias que venían madurando desde la última década: redes sociales, movilidad, aplicaciones, caída del coste de la banda ancha, interconexión de objetos a través de Internet (M2M, *machine to machine*, o Internet de las cosas) y *cloud computing*. Todas estas tendencias tienen una cosa en común: producen una ingente cantidad de datos que necesitan ser captados, almacenados, procesados y analizados.

Empresas, organizaciones y gobiernos trabajan con miles de sensores digitales que arrojan información de todo tipo a la Red. En equipos industriales, automóviles, aviones, trenes, barcos, electrodomésticos, en las calles, estos sensores pueden medir y comunicar la posición o localización, el movimiento, la vibración, la temperatura, la humedad y hasta cambios químicos en el aire, emisión de CO_2. Estas situaciones han existido siempre y eso ha ayudado en la toma de decisiones para prevenir desastres naturales, detectar movimientos sísmicos, ¿dónde está la diferencia actual? Pues que antes los entornos o ambientes estaban controlados por datos estructurados, y ahora los datos provienen de todos lados y son datos no estructurados.

Hoy día en Madrid, México DC, Bogotá, Buenos Aires o Santiago de Chile, por citar algún ejemplo, cualquier usuario puede entrar en Google Maps, introducir una dirección, elegir la visión del satélite y ver en tiempo real la congestión del tráfico de la zona que desea visitar con información que los usuarios envían en tiempo real con sus teléfonos Android.

El gran volumen de datos procede de correos electrónicos, videos, documentos, mensajes de texto SMS, etiquetas RFID, fotografías, imágenes digitales, redes de sensores y dispositivos, índices de búsqueda, condiciones ambientales, redes sociales, exploraciones médicas, información gubernamental, historial de pulsaciones (clics), archivos de música, textos, transacciones en línea (*online*), incidencias telefónicas, junto a todo aquello que se pueda digitalizar y transformar en datos.

Algunas cifras relevantes

Un zettabyte corresponde a 75.000 millones de tabletas iPad de 16 gigabytes o a mil millones de discos duros (rígidos) de una típica computadora de escritorio.

En un minuto en Internet se generan 98.000 tuits se bajan 23.148 aplicaciones, se juegan 208.333 minutos de Angry Birds, 27.000 personas se conectan *(logean)* a Facebook y se ven 1,3 millones de videos en Youtube.

Muchos de estos datos se necesitan analizar en tiempo real, otros estarán almacenados durante años y solo para consultas puntuales. Esta memoria gigante no para de crecer y será necesario dotarla de inteligencia.

La Red va coleccionando datos de nuestro perfil (sexo, edad, gustos, hábitos, preferencias, aficiones, profesión); estos datos sirven para proporcionar mejores resultados en las búsquedas y pueden servir para tomar decisiones o generar políticas públicas que impacten de manera positiva en la sociedad. Los Big Data permiten responder a preguntas tales como: ¿cómo sabe Facebook qué personas me gustaría conocer? ¿Cómo sabe la Web que páginas me interesa visitar?

EL BIG DATA ECLOSIONA EN ESPAÑA (IDC)

Según un estudio de IDC España, patrocinado por EMC, JasperSoft, Microsoft y Sybase, el mercado de Big Data está en auge en el país. Los datos recabados a partir de 502 entrevistas a expertos españoles confirman que un 4,8% de las empresas ya han incorporado estos procesos a su negocio, y las previsiones indican que en 2014 la adopción será del 19,4%, lo que supone un incremento del 304% con respecto a 2012. El Big Data se empieza a mostrar como un factor imprescindible en las empresas españolas. Los beneficios en el 2010 que generaba esta tecnología eran en torno a los 3.200 millones de dólares en todo el mundo. Según las estimaciones de IDC, esta cifra podría llegar a alcanzar los 16.900 millones de euros en 2015.

Las cifras demuestran, según IDC España, que a pesar de la crisis, las empresas están interesadas por tecnologías que generan una mayor eficiencia organizacional y que proporcionan nuevas oportunidades de negocio. El Big Data tiene sentido cuando hablamos de empresas con un alto volumen de información, generada muy rápidamente, procedente de diversas fuentes, con distintos formatos y con datos de calidad. El volumen de información aparece como la principal razón para la explosión del Big Data. IDC preveía que la información digital ascendería a los 2,7 zettabytes durante 2012. Los procesos de Big Data son, por tanto, una tecnología que pronto será indispensable para todas las empresas. Otras razones extraídas del informe son el ahorro de costes y la mejora de la toma de decisiones

Barreras de entrada

Según el informe, cuatro son los principales obstáculos a la hora de la adopción de esta tecnología: la ausencia de expertos, la falta de presupuesto, la dificultad de la integración con los procesos de negocio y la calidad de los datos.

La escasez de profesionales especializados se debe a que se trata de un sector demasiado joven para las empresas españolas y que aún debe volcarse en la formación de talentos que abarquen esta área. El profesional *científico de datos* será esencial para poder realizar la organización de información de una empresa. En cuanto a la falta de presupuestos, existe la posibilidad de recurrir al *software* distribuido y desarrollado libremente, *open source*, ya que "permite reducir costes a la hora de atender a las nuevas tecnologías". La integración de los procesos en el negocio tampoco está exenta de dificultades. Para incluir Big Data en la estructura organizativa de una empresa es preciso que ésta atienda al menos a las cuatro dimensiones: volumen, variedad, velocidad y valor (IDC se apunta al modelo de 4 V). IDC considera que es preciso determinar cómo recoger los datos generados, clasificarlos, almacenarlos, construir arquitecturas con escalabilidades muy altas y crear nuevos modelos de bases de datos que pongan énfasis en el tratamiento de la información *in-memory* (que se analizará más adelante). Por último, el estudio establece que la velocidad de adopción entre unas empresas y otras varía en función de sus características, influyendo sobre todo el tamaño y el campo de actividad en el que operen. El sector financiero, el dedicado a infraestructuras, el público, y el sanitario son los que, según el informe, más recurren a Big Data para gestionar la información. En cuanto a tamaño, las compañías de más de 500 empleados son las más adelantadas en conocimientos, aunque son las empresas de 100 a 500 trabajadores las que dominan la implantación del modelo.

CÓMO CREAR VENTAJAS COMPETITIVAS A PARTIR DE LA INFORMACIÓN: IDC BIG DATA 2012

En septiembre de 2012, IDC hizo público un informe sobre la situación actual y perspectivas de futuro del Big Data en España y a nivel mundial, informe que fue un adelanto de su prestigioso estudio "*The Digital Universe*", publicado en diciembre del mismo año. Según IDC, el volumen de los contenidos digitales crecerá hasta 2.7 ZB en 2012, situándose en 8 ZB en 2015. Este crecimiento se verá impulsado por la exponencial proliferación de usuarios de Internet, redes de sensores y objetos interconectados, dispositivos inteligentes que permiten nuevas formas de trabajo y de comunicación así como las redes sociales que redefinen los modelos de negocio y la forma de interaccionar con los consumidores. Más de un 90% de los datos digitales estarán desestructurados, encapsulando una gran cantidad de información de valor, pero difícil de analizar y comprender. Ante el rápido desarrollo del universo digital, las empresas no pueden seguir confiando en los sistemas tradicionales para la toma de decisiones, que ya no son capaces de ofrecer respuestas ágiles y precisas. Cada vez más, las empresas buscan el valor que proporciona el acceso unificado a la información, y el análisis como soporte para la toma de decisiones de usuarios, grupos y sistemas.

> Aquellas empresas que desplieguen iniciativas de Big Data, recomienda IDC, no solo conseguirán ser capaces de analizar grandes volúmenes de datos, sino que también aumentarán su capacidad para rediseñar los procesos de negocio, e incluso crear nuevos servicios basados en la información. El Big Data es una aproximación crítica para generar ventajas competitivas basadas en la información (IDC, 2012).

RETOS EMPRESARIALES DE BIG DATA

La adopción de la filosofía de Big Data en organizaciones y empresas implica mucho más que la instalación y puesta en marcha del *software* adecuado. Es necesario un cambio organizacional en la empresa y en su personal. Los datos corporativos ya no son responsabilidad exclusiva de un departamento, dado que la asimilación de Big Data implica que todos los grupos de trabajo y departamentos se verán afectados. Por todo ello es imprescindible una formación especializada al personal en la utilización de las herramientas de Big Data con el objetivo principal de capturar, almacenar y manipular los grandes volúmenes de datos en beneficio de la productividad de la empresa.

Así como 2012 fue el año del asentamiento de las tecnologías de *cloud computing,* y una mayoría de organizaciones y empresas además de administraciones utilizan la nube, el año 2013 será el año del lanzamiento de Big Data, ya que se están convirtiendo sus técnicas y tecnologías en viables desde enfoques muy eficaces en coste y calidad que van a permitir controlar y dominar el volumen, la velocidad y variabilidad de los grandes volúmenes de datos.

Hasta ahora las grandes corporaciones como Walmart y Google han tenido a su alcance los grandes datos, pero a un coste elevadísimo. Hoy el *hardware*, las arquitecturas *cloud* y el *software* de fuente abierta están llevando al procesamiento de los grandes volúmenes de datos al alcance de aquellas corporaciones con menos recursos. El procesamiento de los Big Data es factible para incluso pequeñas empresas *startup*s que pueden alquilar tiempo de servidores en la nube.

La emergencia de los grandes datos en la empresa trae consigo una contrapartida: la agilidad. Explotar con éxito el valor de los grandes volúmenes de datos requiere de altas dosis de experimentación y exploración.

EL GRAN NEGOCIO DE BIG DATA

La consultora Deloitte prevé que en 2012 el negocio del Big Data acelerará su crecimiento y aumentará su penetración en el mercado (Deloitte, 2012):

> Big Data representa una oportunidad y un reto. Oportunidad para que las organizaciones sean más eficientes y competitivas aportando servicios de valor añadido a sus clientes, y por otro

lado, les plantea el reto de tener que gestionar grandes volúmenes de datos de muy diversos formatos y fuentes, que crecen año tras año; En este escenario, la tecnología es la clave.

BIG DATA, *THE NEXT THINK* (LA SIGUIENTE GRAN TENDENCIA)

Las empresas pueden sacar gran beneficio en el uso de los grandes volúmenes de datos. El gran caudal de información de las organizaciones y empresas permitirá deducir las necesidades de sus potenciales consumidores.

El volumen de datos por gestionar por las empresas va en aumento cada día merced a la infinidad de datos procedentes de los medios sociales (redes sociales, blogs, *wikis*, dispositivos móviles, objetos del Internet de las cosas, datos de geolocalización). Esta inmensidad de datos no solo será una gran oportunidad, sino también un riesgo al que han de enfrentarse las organizaciones y empresas para intentar no ser sepultados por esa inmensa avalancha de datos.

Las tecnologías y técnicas de Big Data no se deben plantear como un problema, sino como una oportunidad cargada de retos. Aquellas organizaciones que consigan analizar de una forma más inteligente y eficaz la información conseguirán controlar el sector o destacarse en sus mercados, anticipándose con sus decisiones y adquiriendo gran ventaja competitiva.

LA EMPRESA INTELIGENTE

En el año 2006, Andrew McAfee, profesor de la Universidad de Harvard, publicó el artículo "Enterprise 2.0" en el que planteaba una nueva visión de empresa apoyada en la naciente Web 2.0. Esta nueva empresa se apoyaba, esencialmente, en las tecnologías de medios sociales (espina dorsal de la Web 2.0): blogs, *wikis*, RSS, redes sociales. Pasados seis años, la empresa 2.0, cuyo concepto sigue teniendo fuerza, está evolucionando a una nueva empresa social que es cada día más inteligente y que constituye un nuevo concepto de empresa donde el acceso a los recursos está garantizado desde cualquier lugar, con cualquier dispositivo y en cualquier momento, y donde el análisis de la información procedente de los medios sociales se pone al servicio del negocio.

La nueva empresa que se comienza a denominar *empresa inteligente* se sustenta en la interacción entre la nube (*cloud computing*), la movilidad, los negocios sociales (*social business*) y los Big Data. Estas cuatro tendencias unidas al análisis de datos (*analytics*) se están transformando en grandes cambios disruptivos de los negocios, las organizaciones, las empresas, y, en un sentido amplio, la sociedad.

Las tecnologías actuales sustentadas en *cloud computing* y Big Data se han convertido en transversales y esta característica actúa como aglutinador de los departamentos de ventas, de marketing, de recursos humanos, y naturalmente, del propio departamento de tecnologías de la información (TI).

Ricardo Miguez[11] considera que las organizaciones se enfrentan a un nuevo ecosistema y tendrán que reinventarse y adaptarse al cambio de una forma proactiva para optimizar las nuevas oportunidades de negocio; Miguez plantea que: "Estamos en un momento de inflexión tecnológica en el que toda información obtenida con el *social business* debe ser explotada con soluciones de Big Data para reinterpretar los procesos y, a partir de ahí, reinventar la organización y la cultura corporativa".

En este panorama, es preciso tener conciencia de que estas tecnologías disruptivas aparecen en la lista de requisitos de los clientes de las grandes empresas tecnológicas, como señala Enrique Bertrand[12], director de Tecnología de Software AG de España, y se les ha obligado a tener que certificar todos los productos en plataformas *cloud* de múltiples fabricantes, porque ya es una experiencia del usuario. Otra característica importante a tener presente, como también señala Bertrand[13], es la necesidad de estándares en estas tecnologías que se irán consolidando, al existir ya una masa crítica que facilita el desarrollo de estos estándares, ya que hay en la industria un entorno más colaborativo además de organismos dedicados a esta tarea.

En el Foro Económico Mundial, celebrado en junio en Suiza, el concepto de Big Data fue protagonista destacado. El informe desarrollado durante el encuentro declara a Big Data, los grandes volúmenes de datos, como un nuevo activo económico al nivel del oro, del dinero, o del petróleo.

CASOS DE ESTUDIO

ROLLS ROYCE

La compañía británica ha comenzado a incluir sensores en sus motores, que proporcionan información en tiempo real sobre las piezas. Esta acción ha supuesto un cambio esencial, ya que ha pasado de vender un producto a vender, además, un servicio. Es decir, ha obtenido una ventaja competitiva a través de los datos.

GOOGLE

Google ha desarrollado la aplicación Flu Trends[14] que permite descubrir cómo ciertos términos de búsqueda sirven como buenos indicadores de la actividad de la gripe. Cualquier usuario puede entrar y ver la evolución de la gripe a través de datos globales de las búsquedas de los internautas en Google. Con estos datos se pueden hacer cálculos aproximados de la actividad de la enfermedad de la gripe en determinadas regiones, lo que es de gran utilidad en acciones preventivas para evitar la propagación.

La aplicación de evolución de la gripe en Google utiliza los datos globales de las búsquedas de Google para realizar, casi en tiempo real, cálculos aproximados de la actividad actual de la

gripe en todo el mundo. Al ejecutar la aplicación se pueden ver estimaciones históricas de los países donde está implantado (Estados Unidos, Alemania, Francia y España entre otros países).

SMART METERS

IBM lanzó en marzo de 2011 la estrategia *smart meters* dentro de su entidad global para realizar mediciones del consumo energético en los hogares, organizaciones y empresas.

Se trata de analizar el consumo de electricidad de un barrio o en cualquier zona urbana a través de sensores que envían datos de consumo. Sobre la base de esa información, la compañía fue capaz de determinar los hábitos de los vecinos en cada momento del día, ver cómo variaba la demanda y hasta cambiar algunos de esos hábitos con estrategias de premios y bonificaciones a sus clientes.

Estas iniciativas de IBM y de numerosos operadores de electricidad forman parte del despliegue de las redes inteligentes (*smart grid*). La estructura *smart grid* exige la lectura de datos en tiempo real para aspectos a nivel del sistema como la gestión de recursos y su supervisión, además de aspectos a nivel de usuario como la facturación automática o el control del consumo energético, que son, entre otros, algunos de los nuevos servicios y aplicaciones que requiere la generación distribuida y el consumo sostenible.

Esta forma de lectura de datos con los nuevos equipos *smart meter* se basan en la capacidad de gestionar tanto los contadores como los grandes volúmenes de datos medidos mediante lo que se denomina *smart metering*[15].

OPEN DATA

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

Aunque la iniciativa de *Open Data* nació en los Estados Unidos, hoy día forma parte de la Agenda Digital Europea, donde numerosos países (entre ellos, España) han promovido iniciativas de datos abiertos, así como en América Latina, donde países como la Argentina, Colombia y Perú están promoviendo también iniciativas nacionales. Más adelante (capítulo 4) se analizará en profundidad cómo uno de los sectores estratégicos como Big Data puede proporcionar una gran ventaja competitiva a las empresas y grandes beneficios a los usuarios y ciudadanos en general.

UNA BREVE RESEÑA HISTÓRICA DE BIG DATA

La historia del término Big Data se puede dividir en dos etapas. Primero, con el nacimiento y expansión del concepto en el campo científico y de negocios restringido su uso a su conceptualización como tal en la jerga técnica y académica; este período 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 período se puede considerar que se inicia en el año 2008.

El profesor Francis X. Diebold[16], en un trabajo de investigación que está realizando sobre el origen e implantación del término Big Data, y que está publicando con diferentes borradores (el más reciente de noviembre de 2012), hasta conseguir cerrar su investigación, analiza el término desde su aparición en escritos académicos y de negocios, y desde su perspectiva de economista/estadístico. Según Diebold, el uso académico del término Big Data se remonta a Tilly, en 1984, y en el lado no académico cita una primera reseña, publicada en 1987, relativa a una técnica de programación denominada *small code, big data*. En 1989, y por último en 1993, se habla de *Big Data application*s.

Por último Diebold menciona un trabajo de Laney (2001)[17] que se titula *Three V´s of Big Data (Volume, Variety and Velocity*), donde se conceptualiza el significado del término y el fenómeno de Big Data. Las conclusiones de la investigación de Diebold (él también interviene como uno de los primeros científicos, en este caso en el área de la estadística y la econometría, que utiliza el término en el año 2000) es que el término comienza a ser utilizado en dos grandes disciplinas: Ciencias de la Computación (Informática) y Estadística/Econometría, y que nació a mitad de los años noventa, en Silicon Graphics Inc (SGI), en la persona de John Mashey; y posteriormente en 1998, Weiss y Indurkbya, en computación; y Diebold (2000), en estadística/econometía, y Douglas Laney (META Group, hoy Gartner). En resumen, concluye Diebold que el término se puede atribuir razonablemente a Marsey, Weiss e Indurkhya, Diebold y Laney.

EL ORIGEN MODERNO DE BIG DATA

En 2008, Steve Lohr[18], del *The New York Times*, publicó que, de acuerdo con diferentes científicos de computación y directivos de la industria, el término Big Data fue calando en ambientes tecnológicos y comenzó a generar ingresos económicos. Estamos totalmente de acuerdo con Lohr, ya que también de modo ininterrumpido he seguido los avatares de Big Data.

Pero, sin duda, es el artículo que *Wired* [19]publicó en junio del mismo año, el detonante de la explosión de los Big Data; así también lo considera Lohr.

Wired publica un artículo en el 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 (*Ware house*) 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 solo más: más es diferente (Because in the era of big data, more isn't just more, more is different").

En ese mismo número, Chris Anderson[20], su director editor, publicaba otro artículo en el que cuestionaba el hecho de que el diluvio de datos podía dejar obsoleto el método científico. En el artículo plantea que hacía diez años, los *crawlers* de los motores de búsqueda hacían una única base de datos. Ahora Google y compañías similares están tratando el *corpus* masivo de datos como un laboratorio de la condición humana. Ellos son los hijos de la era del petabyte. La era del petabyte es diferente porque más es diferente. Los kilobytes se almacenaban en discos flexibles; los megabytes se almacenaban en discos duros. Los terabytes se almacenaron en *arrays* de discos. Los petabytes se almacenan en la nube. A medida que nos movemos en paralelo a la progresión anterior, nos desplazamos de la analogía de las carpetas (*folders*) a la analogía de los gabinetes de archivos, y de ahí a la analogía de la biblioteca (*library),* y en la era de los petabytes a la analogía de las organizaciones en la nube.

Lohr (2012), en el artículo antes citado, considera que a finales de 2008 se produjo el espaldarazo del mundo científico, ya que los Big Data fueron adoptados por un grupo de investigadores muy reconocidos del mundo de la computación y agrupados en torno a la prestigiosa Computing Community Consortium, un grupo que colabora con el National Science Foundation (NSF) de los Estados Unidos, y la Computing Research Association, también de los Estados Unidos, que a su vez representa a investigadores académicos y corporativos. Este consorcio publicó un influyente artículo (*white paper*) "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society"[21].

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*) del que hablaremos con bastante profusión a lo largo del libro.

Desde un punto de vista popular que demuestra la penetración del término, ya no solo 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, han publicado en agosto de 2012, un dosier especial "Big Data Special Issue", en la revista *Significance,* publicación conjunta de la American Statistical Association y la Royal Statistical Society[22].

RESUMEN

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

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