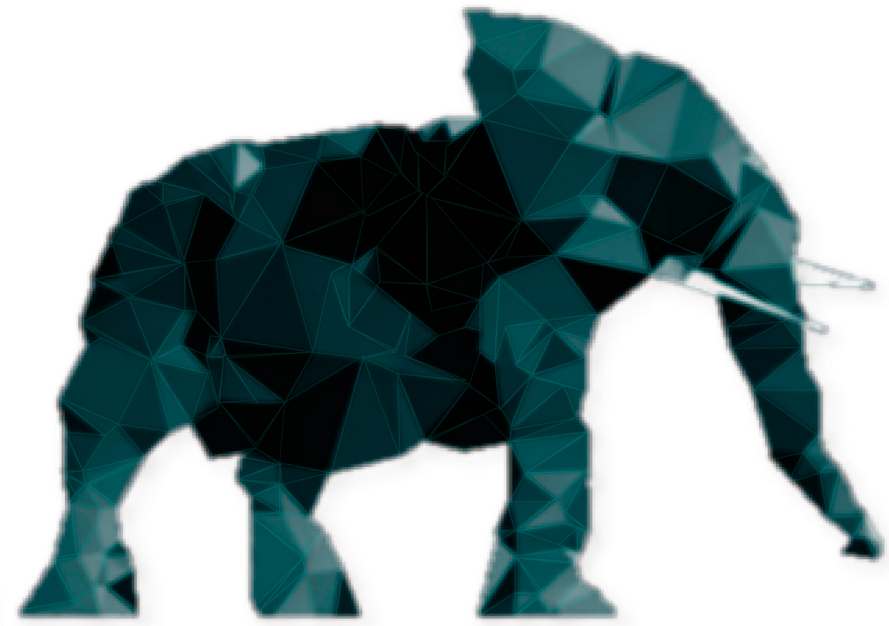
NTICS

tecnología

Big Data

ANÁLISIS DE GRANDES VOLÚMENES DE DATOS EN ORGANIZACIONES

Luis Joyanes Aguilar



Alfaomega

marcombo
ediciones técnicas

Big Data

Análisis de grandes volúmenes de datos en organizaciones

Luis Joyanes Aguilar

Big Data

Análisis de grandes volúmenes de datos en organizaciones

Luis Joyanes Aguilar

Alfaomega

marcombo
ediciones técnicas

Big Data. Análisis de grandes volúmenes de datos en organizaciones

Luis Joyanes Aguilar

Primera edición: Alfaomega Grupo Editor, México, julio 2013

Primera edición: MARCOMBO, S.A. 2014

www.marcombo.com

Diseño de cubierta: Iris Biaggini

ISBN: 978-84-267-2081-8

D.L.: B-20395-2013

Printed in Spain

A Inés y Olivia, que ya viven en la era de la nube y de Big Data disfrutando de sus enormes beneficios y como muestra del inmenso cariño que os tengo.
Luis Joyanes Aguilar

Mensaje del editor

Los conocimientos son esenciales en el desempeño profesional, sin ellos es imposible lograr las habilidades para competir laboralmente. La universidad o las instituciones de formación para el trabajo ofrecen la oportunidad de adquirir conocimientos que serán aprovechados más adelante en beneficio propio y de la sociedad; el avance de la ciencia y de la técnica hace necesario actualizar continuamente esos conocimientos. Cuando se toma la decisión de embarcarse en una vida profesional, se adquiere un compromiso de por vida: mantenerse al día en los conocimientos del área u oficio que se ha decidido desempeñar.

Alfaomega tiene por misión ofrecerles a estudiantes y profesionales conocimientos actualizados dentro de lineamientos pedagógicos que faciliten su utilización y permitan desarrollar las competencias requeridas por una profesión determinada. Alfaomega espera ser su compañera profesional en este viaje de por vida por el mundo del conocimiento.

Alfaomega hace uso de los medios impresos tradicionales en combinación con las tecnologías de la información y las comunicaciones (IT) para facilitar el aprendizaje.

Libros como éste tienen su complemento en una página Web, en donde el alumno y su profesor encontrarán materiales adicionales.

Esta obra contiene numerosos gráficos, cuadros y otros recursos para despertar el interés del estudiante, y facilitarle la comprensión y apropiación del conocimiento. Cada capítulo se desarrolla con argumentos presentados en forma sencilla y estructurada claramente hacia los objetivos y metas propuestas.

Los libros de Alfaomega están diseñados para ser utilizados dentro de los procesos de enseñanza-aprendizaje, y pueden ser usados como textos para diversos cursos o como apoyo para reforzar el desarrollo profesional.

Alfaomega espera contribuir así a la formación y el desarrollo de profesionales exitosos para beneficio de la sociedad.

Acerca del autor

Luis Joyanes Aguilar

Doctor Ingeniero en Informática y Doctor en Sociología, Catedrático de Lenguajes y Sistemas Informáticos de la Universidad Pontificia de Salamanca en el campus de Madrid y profesor invitado en diferentes universidades del mundo. Conferenciante habitual en congresos, seminarios, jornadas y talleres a nivel mundial. Ha escrito numerosos libros y artículos relativos a Tecnologías de la Información.

Patrono de la Fundación de I+D Software Libre de Granada, miembro del Instituto Universitario "Agustín Millares" de la Universidad Carlos III de Madrid y presidente de SISOFT.

Contenido

Prólogo

Big Data (*grandes datos* o ***macrodatos*** según la Fundación Fundéu BBVA) supone la confluencia de una multitud de tendencias tecnológicas que venían madurando desde la primera década del siglo XXI y se han consolidado durante los años 2011 y 2012 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 volúmenes de datos* han ido creciendo de modo espectacular. Durante 2012, se crearon 2,8 zettabytes (ZB) de datos (1 ZB = 1 billón de gigabytes) según datos de la consultora IDC en el estudio "El Universo Digital de Datos 2012" publicado en diciembre de 2012 y esta cifra se dobla cada dos años. Un dato significativo, Walmart, la gran cadena de almacenes de 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. La ignorancia de los *Big Data* producirá grandes riesgos en las organizaciones y no las hará competitivas. Para ser competitivas en el siglo actual, como señala Franks (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".

Big Data ya es una realidad consolidada. La consultora Gartner ha cuantificado el gasto en *Big Data* en 2012 en 28.000 millones de dólares y prevé para el año 2013, la cantidad de 34.000 millones de dólares. A su vez, la auditora Deloitte estima que a finales de 2012 más del 90% de las empresas del índice *Fortune 500* podrían poner en marcha iniciativas de *Big Data*. Por estas razones, los profesionales de *Big Data*, los *analistas de datos* y *científicos de datos*, tienen mucho trabajo por delante y será una de las profesiones más demandadas en 2013 y años sucesivos.

En el libro, además de introducir al lector en los fundamentos de los volúmenes masivos de datos, las tecnologías y herramientas de *Big Data*, se estudiarán las diferentes formas en las que una organización puede hacer uso de esos grandes datos para sacar mayor rendimiento en su toma de decisiones y trataremos de las

oportunidades que traerán 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.

El diluvio de datos

La avalancha o aluvión de datos que cada día genera, captura, almacena y analiza las organizaciones y empresas y, por ende, los particulares, ha dado lugar a la nueva tendencia *Big Data*. Situémonos en un día cualquiera del año; imaginemos los millones de usuarios que visitan Facebook, los millones de *tuits* (*tweets*) que se publican a diario, los millones de mensajes y conversaciones que se realizan a través de WhatsApp, Joyn o Line, los millones de correos electrónicos que envían y reciben millones de personas de todo el mundo, los miles de llamadas telefónicas y videoconferencias a través de Skype. Sumemos a toda ese ingente volumen de información, las páginas que visitan dichos usuarios, las noticias que leen, las ofertas de anuncios, ventas, alquileres, etcétera; las visitas a sitios de turismo, de ocio, de cultura... Multiplique esa información personal de más de 2.800 millones de internautas en el mundo. En esencia, hablamos de datos de la Web y de los medios sociales (Social Media).

Por si ese volumen de información no fuera significativo, añadamos ahora los datos que se transfieren entre sí los miles de millones de objetos o cosas, que se comunican entre sí, a través de sensores, chips NFC chips/etiquetas de RFID, etcétera, es decir, la interconexión de datos entre máquinas (M2M) origen del conocido Internet de las cosas o también de los objetos.

Sigamos sumando, datos médicos de los millones de hospitales, hoy prácticamente digitalizados en su gran mayoría; datos de las administraciones públicas, prácticamente todos en línea −al menos en la mayoría de los países del mundo−; datos de posición y geolocalización, sistemas de información geográfica (SIG/GIS) a través de sistemas GPS y teléfonos inteligentes (*smartphones*), además de miles de satélites de comunicaciones, etcétera.

En resumen, la explosión de los grandes volúmenes de datos (*Big Data*) no para de crecer y parece que de modo exponencial. Eric Schmidt, el presidente ejecutivo de Google, ya advertía hace unos años que: “entre el origen de la Tierra y el año 2003 se crearon 5 exabytes de información. Hoy día creamos la misma cantidad cada dos días”. El estudio citado de IDC/ EMC ya confirmaba también que las cifras del Presidente de Google: alrededor de 2,5 exabytes de datos se creaban cada día en el año 2012 y −es más− el número se doblaba cada 40 meses aproximadamente. El estudio de IDC pronostica que el año 2020 se alcanzará en la Tierra, los 40 zettabytes (40 ZB), se han creado 2,8 ZB de datos durante el año 2012, lo que significa que se generarán 5,247 gigabytes (GB) por cada persona existente en el mundo en ese año. Pero lo sorprendente no sólo es este inmenso caudal de datos, sino que en la presentación del informe se revela que menos del

1% de los datos del mundo se analizan para aprovecharse de esa gran ventaja y valor añadido que suponen los *Big Data* y también que menos del 20% de los datos no están protegidos. El informe advierte de las grandes oportunidades que se ofrecen a las empresas para la protección y extracción del valor que suponen este inmenso volumen de datos.

La revolución de la gestión

Andrew McAfee[2] y Erik Brynjolfsson[3] profesores del MIT publicaron un artículo significativo, en el número de octubre de 2012 de la prestigiosa revista *Harvard Business Review*, "Big Data: The Management Revolution"[4], Las conclusiones fundamentales de su estudio son claras: "La explotación de los nuevos y espectaculares flujos de información pueden mejorar radicalmente el desempeño (rendimiento) de su empresa, Sin embargo, será necesario cambiar su cultura de toma de decisiones". La propuesta final de su artículo es concluyente: "La evidencia es clara: las decisiones controladas por los datos tienden a ser mejores decisiones: los líderes empresariales o bien adoptan esta situación o serán remplazados por otros que lo hagan".

Es decir, las organizaciones se debaten entre adoptar o no *Big Data*, al igual que estos primeros años de la actual década, el debate empresarial era la adopción o no de la computación en la nube[5] con las consiguientes conclusiones para su adopción y migración a la misma de modo radical o gradual, dependiendo de las estrategias de cada organización. La adopción de *Big Data* parece que es un hecho que tarde o temprano deben realizar las organizaciones; los retos y oportunidades que ofrecen compensarán los gastos económicos y de talento que se requerirán y serán compensados con la ventaja competitiva que supondrá dicha adopción y el análisis de esos grandes volúmenes de datos implicarán una gran mejora en la toma de decisiones.

La investigación realizada por los profesores McAfee y Brynjolfsson, no deja lugar a dudas, "las empresas que inyectan *Big Data* y analítica de *Big Data* en sus operaciones muestran tasas de productividad y rentabilidad que son del orden del 5 al 6% más altas que aquellas de la competencia o compañías homólogas"[6].

El científico de datos (*data* scientist): la nueva profesión sexy del siglo XXI

Así titulan su artículo en el citado número de HBR, los prestigiosos analistas Thomas Davenport[7] y D. J. Patil[8]. Evidentemente, los *Big Data*, como ya sucedió con la Web 2.0 y el advenimiento de los medios sociales (*social media*), ha traído nuevos roles el mundo del trabajo así como nuevas profesiones.

La Web 2.0 y los medios sociales han traído: Analista Web, especialista SEO y SEM, Community Manager (Gestor de comunidades), Social Media Manager (Director de medios sociales), analistas sociales... Estas nuevas profesiones están dando paso a los analistas de *Big Data* y, de modo muy especial, al científico de datos (*data scientist*) que convive con el analista de datos y el analista de negocios tradicionales.

Las profesiones de la Web y Social Media ya han llegado a las organizaciones y empresas y su formación académica ya ha entrado en la Universidad, en las Escuelas de Negocios y en los departamentos de formación de las grandes empresas así como en las universidades corporativas. Ahora comienzan a llegar los analistas de *Big Data* y de manera muy significativa, por ser el más demandado y más escaso, el **científico de datos**.

¿Qué es un científico de datos?

Aunque el término, como casi siempre sucede con las ideas de impacto, no está totalmente definido, en cuanto a las personas, sí parece claro que nación en las grandes compañías clásicas de Internet controladas por datos, desde el punto de vista de uso en las empresas y en la industria, tales como Google, Amazon, Facebook, Twitter o LinkedIn, y algunas otras también del mundo de negocios de Internet como eBay, PayPal o de gran éxito en *retailing* o en ventas al por menor (grandes almacenes) como es el caso de Walmart.

Sin embargo, hay cierta unanimidad en fijar a D. J. Partil y Jeft Hammerbadier entonces líderes respectivos de análisis de datos en LinkedIn y Facebook, en el año 2008, como acuñadores del término[9], aunque en 2009, Troy Sadkovsky creó un grupo de investigación en LinkedIn y usó el término para definir una nueva profesión (la suya, por otra parte).

Entonces, ¿qué es un científico de datos? Precisamente, Davenport y Partil, centran todo su artículo de HBR para tratar de definir el nuevo rol del científico de datos ¿Qué tipo de persona debe ser? ¿Qué formación académica le debe respaldar? ¿Qué capacidades y competencia ha de poseer? En el citado artículo, se decanta por un híbrido de hacker, analista, comunicador de datos y asesor de confianza. La combinación es extremadamente potente y rara, confiesan los autores.

Las grandes empresas informática y de Internet parece se decantan por definiciones y roles diversos:

1. "Un buen *data scientist* ha de tener diferentes capacidades: saber matemáticas, tener capacidad analítica y formación en estadística, pero ha de saber contar una historia y tener curiosidad porque se trata de crear significado y valor sobre los datos" (Sonderegger, director senior de *Analytics* en Oracle).
2. "Evolución del analista de datos o de negocios en el contexto de *Big Data*: se considera mitad analista, mitad artista" (Gonzalo Smith, responsables de "*smart analytics*" de IBM GBS España).
3. El científico de datos será aquel que tenga el trabajo más sexy del siglo XXI (Davenport y Partil).
4. Es una persona con habilidades serias en diversas disciplinas técnicas, como ciencias de la computación (informática), analítica, matemáticas, generación de modelos y estadística. Además, debe ser un buen comunicador que sea capaz de entender un problema de negocios, transformar ese problema en un plano analítico, ejecutar el plan y luego dar [10]una solución d negocios (Ani Kaul, CEO de AbsoluData, empresa de analítica e investigación de Alameda, California):

En síntesis, el científico de datos es una profesión emergente. Existen muchos científicos de datos en Google, Amazon, Facebook, LinkedIn, Twitter, y... Todavía existen pocas ofertas de formación en el mundo académico, no sólo iberoamericanas sino de Europa y de Estados Unidos, pero, sin lugar a dudas, estas ofertas irán creciendo poco a poco. En el capítulo 5, se analiza el rol del científico de datos; se citan algunas iniciativas importantes como el caso del prestigioso ITAM de México DF que puso en marcha en 2012, una maestría en Ciencia de Datos.

El análisis de los grandes volúmenes de datos

El análisis de datos y de negocios, son disciplinas antiguas que han experimentado notable crecimiento en todos los campos del saber y, en particular, en organizaciones y empresas, por la necesidad de disponer de herramientas que analicen datos y que éstos sirvan para toma de decisiones eficaces y eficientes.

El análisis de datos ha ido evolucionando a medida que los grandes volúmenes de datos crecían. Las herramientas de inteligencia de negocios han ido recogiendo las tecnologías de OLAP (procesamiento analítico en línea), de informes y consultas (*reporting and query*), de visualización y, especialmente de minería de datos con sus

ya asentadas categorías de minería Web y minería de texto, y las innovadoras minería social en el análisis de datos en medios sociales, que se ha apoyado en técnicas de análisis de sentimiento y de opinión, o minería de opinión y minería de sentimiento como también se la conoce.

La analítica de *Big Data* está emergiendo a la vez que la avalancha de los grandes volúmenes de datos sigue creciendo. "La era de los grandes datos (Big Data) está evolucionando rápidamente y toda nuestra experiencia sugiere que la mayoría de las compañías deben actuar ahora [...]. A medida que las compañías aprendan las destrezas fundamentales (*core skills*) para utilizar los *Big Data*, la construcción de capacidades superiores se pueden convertir pronto en un activo competitivo decisivo" (Barton, Court 2012)[11].

En el libro, trataremos las categorías de Analítica de Datos que consideramos fundamentales para el estudio de *Big Data*: analítica Web, analítica Social, analítica de sentimientos, analítica M2M y, en general, analítica de *Big Data*.

Arquitectura de Big Data

Una arquitectura de Big Data debe considerar la integración de las nuevas tecnologías y herramientas de los grandes volúmenes de datos y su integración con los datos tradicionales (bases de datos relacionales y heredadas "*legacy*") así como la integración con la infraestructura existente de las organizaciones y empresas.

Así pues, en el libro, se ha tratado de considerar los conceptos y componentes básicos que las compañías deberán considerar en la gestión y explotación de sus *Big Data* y que enumeramos a continuación:

- *Fuentes de Big Data*. Los datos proceden de la Web, de los Social Media, de interconexión de objetos M2M mediante sensores conocida como "Internet de las cosas", de la movilidad, biometría, datos procedentes de las propias personas, etcétera.
- Los *tipos de datos* se clasifican en tres grandes categorías: estructurados (datos transaccionales de las bases de datos relacionales), no estructurados (audio, vídeo, fotografía, textos...) y semiestructurados (procedentes, fundamentalmente de archivos HTML, XML).
- Almacenes de datos empresariales (**EDW**, *Enterprise data warehouse*).
- Bases de datos no relacionales (**NoSQL**) que no siguen, normalmente el estándar SQL.
- Bases de datos analíticas "**en memoria**" y **MPP** (procesamiento masivo paralelo).

- Hadoop: el marco de trabajo por excelencia (*framework)* para procesar y analizar los grandes volúmenes de datos, especialmente los datos no estructurados y semiestructurados.
- Analítica de Big Data. Herramientas de analítica, de informes (*reporting*), de consultas (*query*) y de visualización (*dashboard*) así como los cuadros de mando integral (*balanced scorecard*) que conducirán a analítica de datos, en su sentido general, y analítica Web, analítica social, analítica de sentimientos, analítica M2M, etcétera.

En la actualidad, las tecnologías y herramientas de *Big Data* se deben centrar en la integración de datos estructurados y datos no estructurados o semiestructurados, así como la integración de los datos tradicionales en las bases de datos relacionales con los datos no estructurados en las bases de datos analíticas y NoSQL. Otro aspecto fundamental que se considerará en esta obra será el tema de la seguridad y la privacidad de los grandes volúmenes de datos.

Toda esta arquitectura de *Big Data* requerirá de plataformas que gestionen estos grandes datos para que las organizaciones y empresas puedan obtener el máximo rendimiento. Para ello, se requieren de proveedores de soluciones que hoy día son muy numerosos y que pueden ser agrupados en proveedores de código propietario o de código abierto (*open source*). Entre los cuales podemos destacar: *Soluciones de Big Data propietarias* (SAP, Oracle, IBM, EMC, HP, SAS...), *Bases de datos NoSQL* (Cassandra, MongoDB, CouchDB...), *bases de datos "en memoria"* donde sobresalen SAP HANA, aunque Oracle, IBM, Teradata, EMC ofrecen soluciones similares, integración de todas estas herramientas en el *marco de trabajo Hadoop* con plataformas eficientes e innovadoras como Cloudera, Hortonworks o MapR entre otras.

Innovaciones tecnológicas que han acelerado los Big Data

Las innovaciones que han acelerado la explosión y avalancha de los grandes volúmenes de datos son muchas, pero cuatro son los grandes pilares sobre los que se sustentan las tecnologías de *Big Data*: Los medios sociales ***"Social media***", la ***Movilidad*** (teléfonos inteligentes, tabletas... y aplicaciones (*apps*), ***Cloud Computing*** (Computación en la Nube) e ***Internet de las cosas*** (**M2M**, sensores de todo tipo, chips **NFC, RFID**...).

Sin embargo, las tendencias se segmentan y las tecnologías, dispositivos y aplicaciones Web más innovadoras crecen casi exponencialmente y son fuentes y origen de datos de todo tipo. Así se pueden considerar las tendencias tecnológicas que vendrán controladas por las multitudes inteligentes (*crowds*) y que influirán en la explosión de los grandes volúmenes de datos: *crowdsourcing, crodwfunding,*

BYOD (*Bring Your Own Device*), *consumerización* y *gamificación*, fundamentalmente. El gran volumen de datos que se irán generando se verán notablemente influenciadas por las tecnologías anteriores y las tendencias que se presuponen tendrán gran impacto en organizaciones y empresas.

Organización del libro

El libro se ha estructurado en cuatro partes que pretenden abarcar las partes fundamentales de las tecnologías y las estrategias de *Big Data*, así como la arquitectura nuclear de los *Big Data*, unidas a la verdadera razón de ser de los *Big Data*, la teoría y herramienta de análisis de los grandes volúmenes de datos con el objetivo esencial de ayudar en la toma de decisiones de organizaciones y empresas. El libro se complementa con varios anexos prácticos donde se describen las plataformas, proveedores y soluciones comerciales más implantadas en la actualidad a nivel mundial, precedidas de un informe sobre el panorama o paisaje actual de las *Big Data* donde se mostrarán de un modo integrado las diferentes plataformas, proveedores y herramientas que componen la oferta comercial a disposición de organizaciones y empresas. Asimismo, se incluye un glosario de términos de *Big Data* que faciliten la comprensión por parte del lector de los numerosos y variados conceptos relacionados con estas tecnologías y tendencias.

PARTE I. LA ERA DE LOS BIG DATA

Capítulo 1. ¿Qué es Big Data?

Capítulo 2. Fuentes de grandes volúmenes de datos

Capítulo 3. El Universo Digital de Datos: El almacén de Big Data

Capítulo 4. Sectores estratégicos de Big Data y Open Data

Capítulo 5. Big Data en la empresa: La revolución de la gestión, la analítica y los científicos de datos

PARTE II. INFRAESTRUCTURA DE LOS BIG DATA

Capítulo 6. Cloud Computing, Internet de las cosas y SoLoMo

Capítulo 7. Arquitectura y gobierno de Big Data

Capítulo 8. Bases de datos analíticas: NoSQL y “en memoria (*in-memory*)”

Capítulo 9. El ecosistema Hadoop

PARTE III. ANALÍTICA DE BIG DATA

Parte IV. EL FUTURO DE BIG DATA

APÉNDICES

Agradecimientos

En primer lugar, deseo expresar mi agradecimiento a mis alumnos de Ingeniería Informática e Ingeniería de Organización Industrial de la Universidad Pontificia de Salamanca en las asignaturas de Sistemas Informáticos, Sistemas de Información, Gestión del Conocimiento e Inteligencia de Negocios, que me han permitido experimentar las tecnologías, herramientas y aplicaciones de *Big Data* tanto en las clases teóricas como en los numerosos talleres y trabajos que ellos han realizado como actividades académicas. A mis alumnos de máster y doctorado de la Universidad Pontificia de Salamanca (campus Salamanca), Universidad Carlos III de Madrid (Documentación) y Universidad Nebrija de Madrid (Empresa) donde he podido implantar todos los conocimientos que he ido adquiriendo en la ciencia y arte de los grandes volúmenes de datos. También quiero agradecer a mis estudiantes de doctorado de la Pontificia de Salamanca del campus de Madrid que investigan conmigo en *cloud computing*, *Big Data*, movilidad y *social media* así como en áreas aplicadas como negocios digitales, educación, gestión del conocimiento,

inteligencia de negocios y sistemas de información geográfica, entre otras, a los que ya han leído su tesis doctoral y a los que espero que lean muy pronto.

Asimismo, deseo agradecer a los alumnos y profesores de las universidades latinoamericanas donde he impartido conferencias, cursos y talleres en los últimos dos años y donde he tratado los temas de **Cloud Computing** y *Big Data*: **México** (TEC de Monterrey, campus Cuernavaca, Universidad del Valle, Instituto Politécnico de México, Instituto Tecnológico de Tijuana, Instituto Tecnológico Superior de Coatzacoalcos (ITESCO), Universidad Autónoma de Baja California –sedes de Tijuana y de Ensenada), **Perú** (Universidad San Martín de Porres, Universidad Tecnológica de Perú y Universidad Garcilaso de la Vega), **Panamá** (Universidad Tecnológica de Panamá), **República Dominicana** (ITLA "Instituto Tecnológico de las Américas", Universidad Unibe, Universidad APEC y la Fundación Funglode).

Por último, quiero agradecer al director editorial Alberto Umaña y al gerente editorial Marcelo Grillo por su apoyo a la colección de libros **NTiCS** que tengo el honor de dirigir y que espero que continúen bien pronto con los siguientes números y en esta ocasión particular a mi editor y, no obstante, amigo, Damián Fernández que desde Buenos Aires me ha acompañado día a día en esta ardua tarea que ha sido la producción de esta obra de *Big Data* y que ha colaborado estrechamente conmigo no sólo como editor sino como un gran amigo que me ha ayudado cuando ha sido menester en esta larga y prolífica tarea de lanzar esta obra sobre un tema tan innovador y de futuro como es *Big Data*. Gracias al equipo editorial.

En **Carchelejo (Sierra Mágina)**, Andalucía (España) y en **México DF** (México), a diecisiete de Mayo de 2013, ***Día Mundial de Internet.***

Luis Joyanes Aguilar

Catedrático de Lenguajes y Sistemas Informáticos de la Universidad Pontificia de Salamanca

NOTAS

[1] Bill Franks (2011). *Taming the Big Data Tidal Wave*, NewJersey: Wiley, p.3.

[2] Investigador del MIT´s Center for Digital Business y autor de Enterprise 2.0 (Harvard Business School Press, 2009).

[3] Director del Center for Digital Business, del MIT´s Solan School of Management. Con Andrew McAfee, son autores de Race Against the Machime Digital Frontier, 2012.

[4] Andrew McAfee[4] y Erik Brynjolfsson, "Big Data: The Management Resolutor", Harvard Business Review, 2012.

[5] Luis Joyanes. *Computación en la nube. Estrategias de cloud computing en las empresas*. México: Alfaomega, 2012.

[6] Op. cit.p. 78.

[7] Profesor visitante de Harvard Business School y autor de varios libros de Gestión del Conocimiento y de Analítica. El autor de esta obra leyó varias de sus obras mientras se formaba en el área de Gestión del Conocimiento.

[8] Científico de datos y antiguo director de productos de datos de la red social LinkedIn.

[9] Desde el punto de vista científico, el origen se remonta a los años sesenta donde comenzó a utilizarse en artículos científicos.

[10] Entrevista realizada en la revista Information Week, edición de México, publicada el 22 de noviembre de 2012, en su edición digital: www.informationweek,com.mx.

[11] Dominic Barton y David Court. Making Advanced Analytics Work For You. HBR, Octubre 2012, p. 88.

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.