Big data. Un nuevo paradigma de análisis de datos

There was five exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days, and the pace is increasing.

Eric Schmidt, former CEO of Google, 2010

Carlos Maté Jiménez

Profesor Propio de la ETSI (ICAI) de la Universidad Pontificia Comillas de Madrid, adscrito al Departamento de Organización Industrial y al Instituto de Investigación Tecnológica (IIT). Doctor en Ciencias Matemáticas y diplomado en Ciencias Económicas y Empresariales por la Universidad Complutense. Actualmente imparte las asignaturas de Análisis de Datos, Economía y Gestión de Empresas y Estadística. Reconocido experto en predicción y en análisis de datos simbólicos, ha escrito varios libros sobre Estadística y publicado diversos artículos sobre aplicaciones de los métodos estadísticos en prestigiosas revistas internacionales y nacionales, tanto del ámbito industrial, informático y de organización como del económico.

Palabras clave: cálculo distribuido, conjuntos de datos masivos, estadística, minería de datos.

Resumen:

En nuestros días, es un hecho incuestionable la ingente cantidad de información que se genera cada segundo en nuestro planeta. Dicha información puede ser estructurada, semiestructurada o no estructurada. También puede aportar enorme valor a cualquier entidad o puede suponer un consumo excesivo de recursos humanos, informáticos, etc.

El análisis inteligente (y la mayoría de las veces en tiempo real) de este tipo de información está empezando a ser un requisito innegable para la supervivencia de muchas empresas y organizaciones. Como consecuencia de ello han surgido en los últimos años términos de nuevo cuño como *big data,* Mapreduce, Hadoop o computación en la nube. Así, la demanda de los llamados "científicos de datos" está creciendo exponencialmente.

Este artículo plantea una introducción divulgativa a todos estos términos y analiza las estructuras más conocidas para el tratamiento de los *big data,* así como las cuestiones legales y éticas.

Key words: *data mining, distributed computing, massive datasets, statistics.*

Abstract:

An unquestionable fact is the vast amount of information that each second is now generated on our planet. This information can be structured, semi-structured or unstructured. It can also bring tremendous value to any entity or may lead to undue consumption of human or computing resources. Intelligent analysis (generally in real time) of this information is becoming absolutely undeniable for the survival of many companies and organizations. As a result demand for the so-called "data scientist" is growing exponentially and new concepts like big data, *Mapreduce, Hadoop or cloud computing have emerged.*

This article presents an informative introduction to all these terms and analyzes the best known structures for the treatment of big data, as well as legal and ethical issues.

Introducción

Tradicionalmente la estructura de un conjunto de datos se presenta como una matriz de *n* filas y *p* columnas, representando cada fila información sobre *p* variables medidas en cada unidad (individuo, empresa, inmueble, calle de una gran ciudad, procedimiento judicial, etc.). Por ejemplo, la hoja de cálculo Excel 2013 puede utilizarse para mostrar 1.048.576 filas por 16.384 columnas en cada hoja, siendo los límites máximo de almacenamiento en memoria de 2 gigabytes (GB) en un entorno de 32 bits, y los límites del sistema y su memoria en un entorno de 64 bits.

Recordamos que un bit es la mínima cantidad de información procesada, sólo puede ser 1 o 0; mientras que un byte es un conjunto de 8 bits. La Tabla 1 muestra los distintos múltiplos del byte con algunos ejemplos de los ámbitos estático y dinámico de la información, tomando como base el año 2014.

Una solución a las limitaciones de Excel procedió de los sistemas de gestión de bases de datos relacionales (RDBMS), que utilizan lenguaje de consultas estructurado (SQL) para definir consultas y actualizar la base de datos. Las empresas líderes en el mercado de sistemas de bases de datos son Oracle, IBM y Microsoft.

Estos sistemas se diseñaron para la retención de datos estructurados, en lugar de para asimilar un crecimiento vertiginoso de los mismos y la mayoría de las veces presentándose en forma no estructurada o semiestructurada, lo que hace de ellos una herramienta extraordinariamente costosa si la quisiéramos utilizar para manejar y almacenar datos masivos. Por ejemplo, se pueden consultar las especificaciones de capacidad máxima para un servidor SQL en 2014 en la web:

http://msdn.microsoft.com/en-us/library/ms143432.aspx

La conclusión a la que se llega es la incapacidad de las bases de datos tradicionales para dar respuesta a muchos de los datos que aparecen ahora en las empresas. Por ejemplo, la información que se genera cada día a través de la opinión de los clientes de una marca en las redes sociales como Facebook, Twitter, etc.

El término "big data": definiciones y tipos. Internet de las cosas

Desde hace unos años (especialmente los dos últimos años), se ha venido observando que las cantidades masivas de datos recogidas a lo largo del tiempo responden al concepto de *big data*. Se han propuesto varias definiciones para este término, aunque todavía no hay una definición universal al respecto (http://datascience.berkeley.edu/what-is-big-data/ recoge más de 40 definiciones). La Organización Mundial de Normalización (ISO) ha creado un grupo de trabajo que va a redactar la norma de vocabulario ISO 3534-5, dedicada al mundo del *big data* y la analítica predictiva. Mientras llega esa definición universal comentamos algunas de las más utilizadas.

La definición que proporciona el diccionario de inglés de Oxford es "datos de tamaño muy grande, típicamente hasta el extremo de que su gestión presenta retos logísticos significativos".

El estudio publicado por McKinsey Global Institute (MGI) en junio de 2011:

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

iluminó el sentido de la definición anterior al definir *big data* como "*conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos*".

Tabla 1. Unidades básicas de información y tratamiento de datos

Nombre	Símbolo	Sistema internacional	Ejemplo 2014 estático	Ejemplo 2014 dinámico
Byte	B	10^0 bytes	1 B es un número de 0 a 255	
Kilobyte	KB	10^3 bytes	2 KB es aproximadamente un sector de CD-ROM	
Megabyte	MB	10^6 bytes	3 MB es aproximadamente una canción de 3 minutos	4 MB/min en llamadas de vídeo por Skype
Gigabyte	GB	10^9 bytes	8/16 GB es el tamaño estándar de mercado de un pen-drive	4 GB/hora de vídeo de alta calidad
Terabyte	TB	10^{12} bytes	4 TB es el tamaño de un disco de 120 € que almacena 800.000 fotos o canciones mp3	20 TB/hora es la información generada por un motor de avión en el aire
Petabyte	PB	10^{15} bytes	2 PB es la información almacenada en todas las bibliotecas de investigación académicas de USA	24 PB/día es la información recogida por Google
Exabyte	EB	10^{18} bytes	5 EB es aproximadamente todas las palabras pronunciadas por todos los seres humanos	966 EB es aproximadamente la predicción del volumen total de Internet en 2015
Zettabyte	ZB	10^{21} bytes	Se estimó que en 2012 la capacidad instalada de almacenamiento de información en el mundo sería de 2,5 ZB.	5 ZB/año es la cantidad de datos digitales promedio que se van a generar en la Tierra en los próximos 8 años
Yottabyte	YB	10^{24} bytes	1 YB equivale a la capacidad del Data Center inaugurado por la NASA en 2013	
Xerabyte	XB	10^{27} bytes	1 XB equivale a 1.257.000 iPad 3 de máxima capacidad por cada habitante de la tierra	

En 2012 Gartner definió *big data* como "*activos de información caracterizados por su volumen elevado, velocidad elevada y alta variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y la toma de decisiones en las organizaciones*". Esta definición hace mención a las 3 famosas "V" de los *big data:* Volumen, Velocidad y Veracidad (Figura 1); cuyos detalles se pueden consultar en el libro blanco de Fujitsu –Mitchell *et al.* (2012)– y en Zicari (2014). Adicionalmente se han propuesto nuevas "V" como Valor, Veracidad y Visualización; o incluso Volatilidad, Validez y Viabilidad.

Los tipos de datos en las aplicaciones de *big data* se muestran en la Tabla 2.

Las redes sociales como Facebook, Twitter, Linkedin, etc., son uno de los más reconocidos caladeros para obtener datos masivos, habiendo dado lugar a una línea de investigación importante, que es el análisis del sentimiento. Una de sus ramificaciones es la incidencia que tiene en las finanzas (ver, por ejemplo, Cerchiello and Giudici [2014]).

Otra fuente de generación ingente de *big data* en los próximos años va a ser el Internet de las cosas, cuyos detalles se pueden consultar en:

http://www.cisco.com/web/LA/soluciones/executive/assets/pdf/internet-of-things-iot-ibsg.pdf

Se trata de todos los datos que se generan entre persona y máquina o entre máquina y máquina (Tabla 2), que como ocurre con los datos de las redes sociales también suelen ser no estructurados.

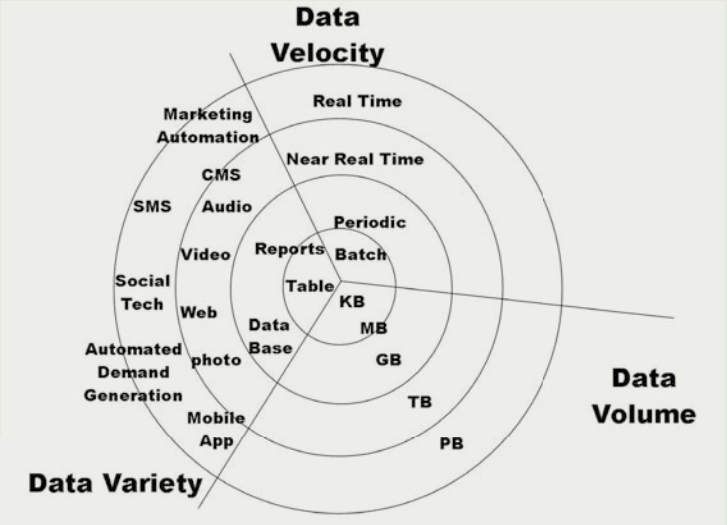
Distintos ejemplos de contextos sobre *big data* se muestran en Akerkar (2014).

Soluciones informáticas para el tratamiento de *big data*

El procesar la información asociada a conjuntos de datos cuyo tamaño es del orden de 10 TB plantea utilizar sistemas distribuidos en nodos en lugar de sistemas con un único nodo. La razón principal es la rapidez. Si un nodo procesa 50 MB/s requerirá 2,3 días para procesar la información anterior. Sin embargo, con un clúster de 1.000 nodos sólo necesitaremos 3,3 minutos.

Una parte importante de los inicios de desarrollo de plataformas informáticas para el tratamiento de *big data* se encuentra en dos artículos que escribieron los investigadores de Google. Ghemawat *et al.* (2003) diseñaron e implantaron el sistema de ficheros de Google (GFS) como un sistema de ficheros distribuido y escalable para aplicaciones intensivas en datos. Dean y Ghemawat (2008) crearon la herramienta MapReduce y en 2004 (primera versión de su artículo) solicitaron la patente del sistema y método para el procesado eficiente de datos a gran escala, que fue concedida seis años después (Dean y Ghemawat [2010]).

Figura 1. Las 3 primeras "V" de *big data*



Fuente: http://velvetchainsaw.com/2012/07/20/three-vs-of-big-data-as-applied-conferences/

MapReduce

Es un modelo de programación y una implantación para procesar y generar grandes conjuntos de datos que tiene sus orígenes en el lenguaje LISP. Los usuarios tienen que especificar varias funciones *Map* (M en la Figura 2). Una función mapa *(Map)* procesa un par clave/valor generando un conjunto intermedio de pares clave/valor. Es decir:

Map (clave, valor) → (clave', valor')

A continuación actúan varias funciones *Reduce* (R en la Figura 2). Una función de reducción *(Reduce)* mezcla todos los valores intermedios (clave', valor') asociados con la misma clave intermedia (clave'). Cada *Reduce* genera una salida de fichero única (o cero).

Tabla 2. Tipos de datos en el paradigma *big data*

Datos estructurados	Datos semiestructurados	Datos no estructurados
Fichas de clientes Fecha de nacimiento Nombre Dirección Transacciones en un mes Puntos de compra	**Correos electrónicos** Parte estructurada: destinatario, receptores, tema Parte no estructurada: cuerpo del mensaje	**Persona a persona** Comunicaciones en las redes sociales **Persona a máquina** Dispositivos médicos Comercio electrónico Ordenadores, móviles **Máquina a máquina** Sensores, dispositivos GPS Cámaras de seguridad

La Figura 2 muestra el marco MapReduce, cuya empresa pionera fue Google.

EJEMPLO:

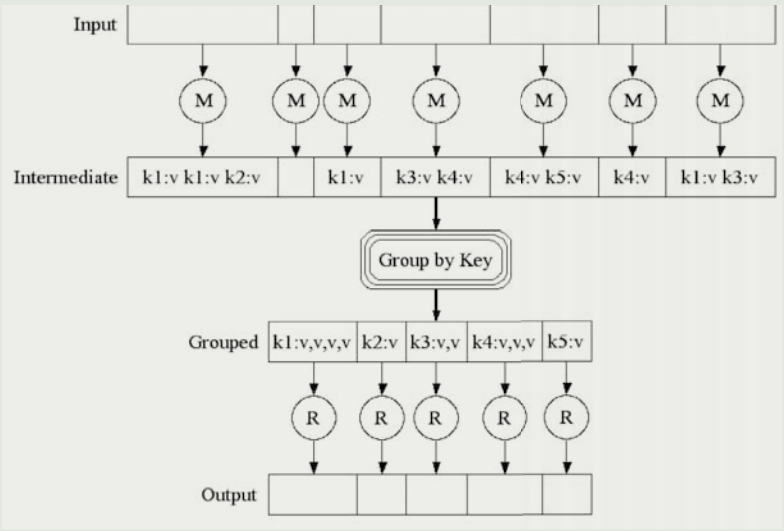
Un caso de uso frecuente es aplicar *Map* y *Reduce* de forma sucesiva, primero se prepara un conjunto de datos vía *Map* y luego se extrae información vía *Reduce*. Por ejemplo, siguiendo la información de la siguiente web:

http://www.infosun.fim.uni-passau.de/cl/MapReduceFoundation/ la Figura 3 muestra una tarea de MapReduce en la que contabiliza las ocurrencias de cada palabra (datos de salida a la derecha) en los datos de entrada (izquierda).

Es decir, el ejemplo anterior nos muestra el cálculo de la frecuencia absoluta en términos de Estadística Descriptiva de cada una de las modalidades presentes en los datos de entrada. Lógicamente con los datos de salida se pueden obtener frecuencias relativas y aplicar procedimientos gráficos como pictogramas, diagramas de barras, etc. En el caso de que la información de entrada sea numérica, una de las tareas claves en la generación de gráficos de cajas y búsqueda de los cuantiles consiste en ordenar los datos de entrada.

Los programas escritos en este estilo funcional automáticamente se configuran en paralelo y se ejecutan sobre un gran clúster de máquinas, siendo altamente escalable. Por ejemplo, un cálculo típico de MapReduce procesa decenas de TB en miles de máquinas.

Figura 2. Marco MapReduce de Google



Fuente: Dean y Ghemawat (2008).

Hadoop

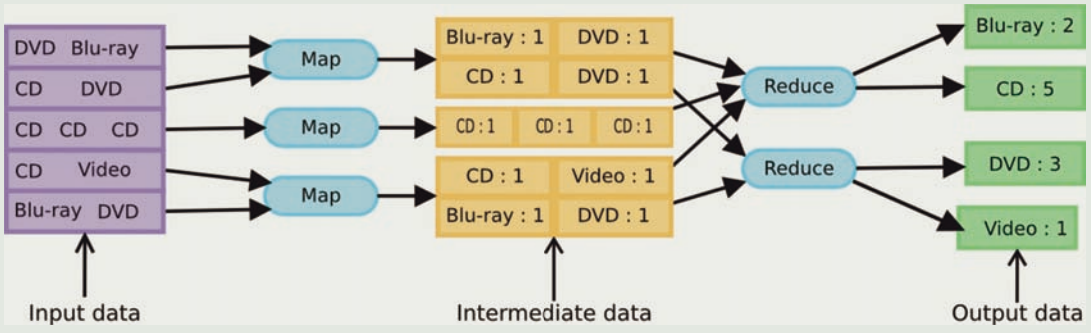
El proyecto Apache™ Hadoop® (http://hadoop.apache.org/) desarrolla *software* libre para el cálculo distribuido, fiable y escalable. Conocido popularmente por Hadoop y representado por un elefante amarillo (Figura 4), se trata de una plataforma de *software* que permite escribir con facilidad y ejecutar aplicaciones que procesan ingentes cantidades de datos. Incluye:

- MapReduce (motor de cálculo *offline*).
- HDFS (sistema de ficheros distribuidos de Hadoop).
- HBase (acceso de datos *online*).

El mayor contribuyente a los desarrollos de Hadoop es por el momento Yahoo. Las características de Hadoop que lo hacen especialmente útil son:

- **Escalable:** diseñado para escalar de servidores individuales a miles de máquinas, cada una ofreciendo cálculo local y almacenamiento; puede llegar a procesar y almacenar petabytes de manera fiable.
- **Económico:** distribuye los datos y los procesa a través de clústers de ordenadores comúnmente disponibles (en miles).
- **Eficiente:** al distribuir los datos puede procesarlos en paralelo sobre los nodos donde los datos están localizados.
- **Fiable:** automáticamente mantiene copias de datos y también de manera automática realiza de nuevo tareas de computación basadas en fallos.

Figura 3. Ejemplo de MapReduce

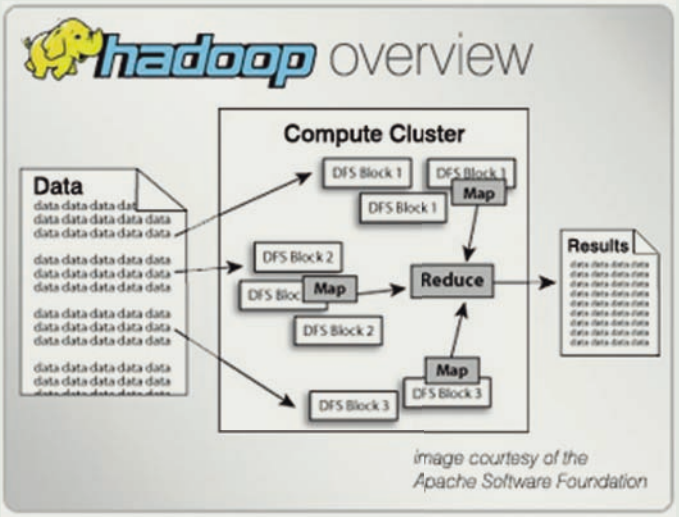


Fuente: http://www.infosun.fim.unipassau.de/cl/MapReduceFoundation/

EJEMPLOS:

- **Amazon.** Para construir los índices de búsqueda de producto de Amazon dentro de su analítica se procesan diariamente millones de sesiones. Se emplean JAVA y API de *streaming*, variando los clústers de 1 a 100 nodos.
- **Yahoo.** Hadoop se ejecuta en más de 100.000 CPU que se encuentran en aproximadamente 20.000 ordenadores. El clúster más grande es de 2.000 nodos (cada disco tiene 4 TB y está montado en cajas de 2 x 4 CPU). Su uso está vinculado a búsquedas en la web.
- **Facebook.** Emplea Hadoop para almacenar copias de log internos y fuentes de dimensión de datos. Lo utiliza como fuente para generar informes de analítica y aprendizaje de máquina. El sistema tiene un clúster de 320 máquinas con 2.560 núcleos y alrededor de 1,3 PB de almacenamiento bruto. Más detalles en Zicari (2014).

Figura 4. Esquema de Hadoop y MapReduce

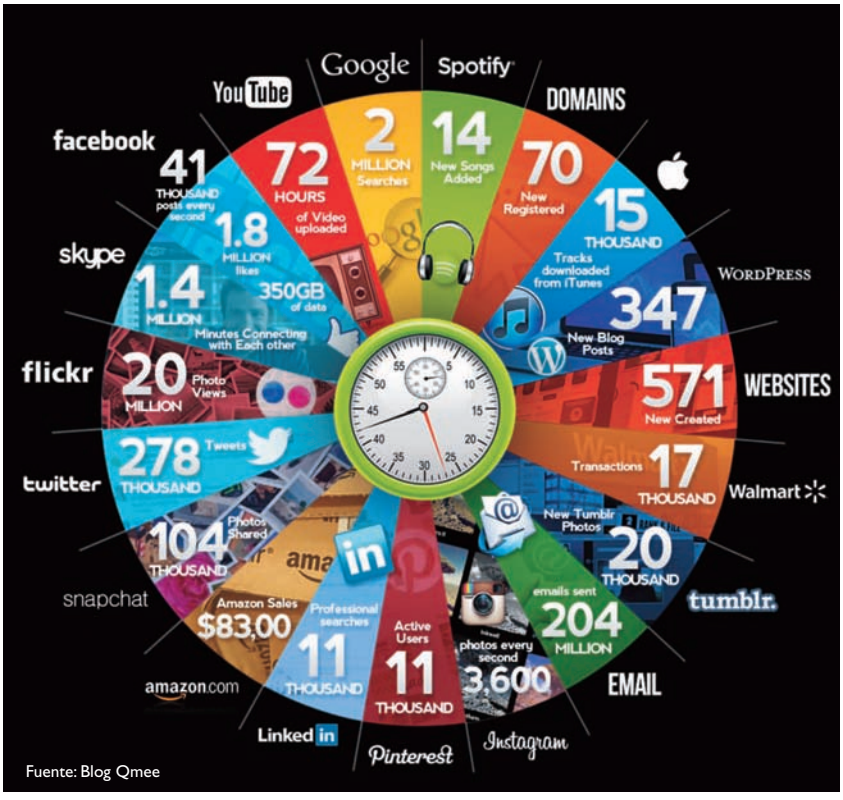


Fuente: Apache Software Foundation

NoSQL y Hadoop

El término NoSQL (Not Only SQL) hace referencia a amplias clases de bases de datos que se diseñan para manejar datos semiestructurados. No utilizan el lenguaje de consultas o SQL. Más detalles en Pokorny (2013).

Hadoop y NoSQL son sistemas abiertos o libres, poseen alta velocidad y muestran un elevado grado de tolerancia al fallo. Son eficientes en costes porque almacenan los datos en pequeños trozos a través de varios servidores. Pueden procesar consultas con rapidez al enviar varias consultas a múltiples máquinas al mismo tiempo. Debido a estas ventajas, Microsoft, Oracle, IBM, EMC, Teradata y otras empresas los han incorporado en sus propias plataformas.



Fuente: Blog Qmee

Computación en la nube

El término computación en la nube (*cloud computing*) es una solución de las tecnologías de la información (IT) para ofrecer recursos y servicios sobre Internet. Según la definición del NIST (National Institute of Standard and Technology), el *cloud computing* es un modelo tecnológico que permite el acceso ubicuo, adaptado y bajo demanda en red a un conjunto de recursos de computación configurables compartidos (por ejemplo, redes, servidores, equipos de almacenamiento, aplicaciones y servicios) que pueden ser rápidamente aprovisionados y liberados con un esfuerzo de gestión reducido o interacción mínima con el proveedor del servicio.

La idea básica es que toda la información se almacena de forma distribuida en servidores, siendo accesible en cualquier momento por el usuario sin que éste se preocupe de nada, el propio sistema de "cloud" es el que se encarga de mantener siempre la

información disponible. En el caso de que se esté almacenando una aplicación en la nube, el propio sistema es el que se encarga de subir la capacidad de computo, memoria, etc., en función del uso que se le está dando a la aplicación, con lo cual en la nube no sólo se delega la capacidad de almacenamiento sino que también se distribuye en los servidores el procesamiento de datos. Esto hace que en un sistema en la nube las capacidades de cálculo y almacenamiento sean muy elevadas.

La computación en la nube ha supuesto una reducción de costes, una mayor flexibilidad y una utilización óptima de los recursos, por lo que se considera que es una herramienta de ventaja competitiva de las empresas. Entre sus usos destaca la analítica de los *big data*. Kambatla *et al.* (2014) han indicado que una de las principales aplicaciones de la generación futura de sistemas distribuidos y de cálculo paralelo se encuentra en la analítica de los datos enormes. Los repositorios de datos para tales aplicaciones exceden actualmente la magnitud de exabytes y están creciendo rápidamente en tamaño. Los datos residen en plataformas con capacidades computacionales y de red que varían ampliamente. Ello hace que las consideraciones de tolerancia a fallos, seguridad y control de acceso sean críticas.

El territorio emergente de entornos basados en la nube con centros de datos que acogen grandes repositorios de datos plantea la necesidad de algoritmos distribuidos/paralelo efectivos. Se trata de un tema de investigación en la frontera del conocimiento de las técnicas de inteligencia artificial de aprendizaje de máquina como las redes neuronales, las técnicas de clasificación o los diagramas en árbol.

Para más detalles acerca de las cuestiones relativas al tratamiento de los *big data* a través de la computación en la nube veáse la revisión de Hashem *et al.* (2015).

Cuestiones legales y éticas

La obtención, tratamiento y explotación de los *big data* plantea importantes cuestiones de índole legal. El antecedente legislativo más conocido es la Ley Orgánica de Protección de Datos (LOPD), que se puede consultar en la Agencia Española de Protección de Datos (AEPD), cuya web es:

https://www.agpd.es/portalweb AGPD/index-ides-idphp.php

Su modificación por la influencia de los datos masivos, computación en la nube, internet de las cosas, etc., todavía no ha sido propuesta en España pero es posible que sea acometida en la próxima legislatura. La imperiosa necesidad de esa modificación de la LOPD vendrá de la toma de conciencia por parte de la sociedad de las implicaciones éticas correspondientes que analizamos más adelante. Un documento actualizado de la AEPD sobre todo ello es:

http://www.agpd.es/portalweb AGPD/canaldocumentacion/publicaciones/common/Guias/Guia_EIPD.pdf.

La reflexión sobre las implicaciones éticas de los *big data* suele estar presente en los distintos eventos que se organizan sobre este tema, como la clausura del Año Internacional de la Estadística en diciembre de 2013, entre otros; concluyendo que se va a poner a prueba el nivel ético de los distintos usuarios de estos datos ya sean gobiernos, organizaciones o empresas.

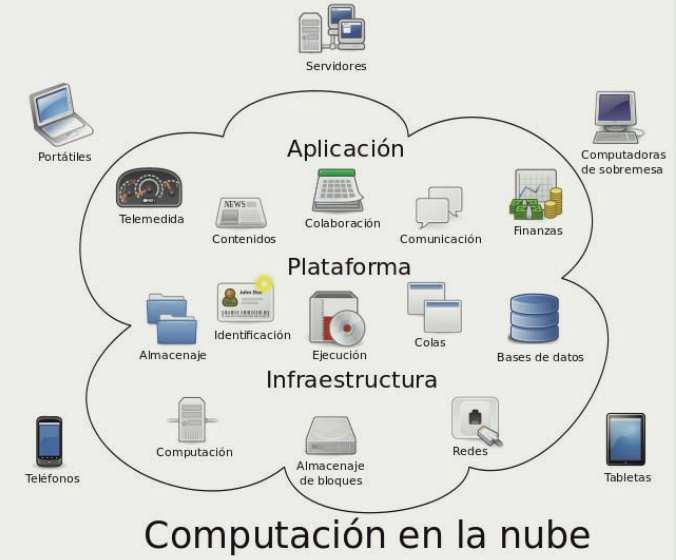
Recientemente, Pulido (2014) en la lección inaugural del curso 2014-2015 en la UAM ha identificado las siguientes cuestiones éticas sobre los *big data*: privacidad, transparencia, pérdida de identidad, discriminación y castigo anticipado y peligro de exclusión. Remitimos a los lectores a dicho documento para profundizar sobre estas cuestiones. En el caso del marketing se puede consultar Nunan y Di Domenico (2013).

Digamos que igual que es necesario un carné de conducir para dirigir los movimientos de una moto, coche o camión con las consiguientes responsabilidades penales; será necesario también un carné de conducción de datos para tratar y analizar los datos, también con las consiguientes responsabilidades penales. Las modalidades de este futuro carné de datos probablemente dependerán de la complejidad y tamaño de los datos a analizar.

Conclusiones

En el año 2010 el término *big data* era prácticamente desconocido. A mediados de 2011 se convertía en una palabra que aparecía con frecuencia entre las últimas tendencias. Lo que parecía iba a ser una palabra de moda (*buzzword*) y, por ende pasajera, se ha convertido en todo un área de interés enorme para las empresas, orga-

Figura 5. Estructura de *cloud computing*



Fuente: http://es.wikipedia.org/wiki/Computación_en_la_nube

nizaciones y administraciones públicas; generando un mercado profesional emergente, que es el de los científicos de datos, y abriendo nuevas líneas de investigación y nuevas revistas específicas en el contexto académico como *Big Data* o *Big Data Research*. Desde el punto de vista de la normalización, ISO está desarrollando la nueva norma de vocabulario ISO 3435-5, donde se recogerán los términos asociados a este concepto.

Este artículo ha presentado la evolución del análisis de datos clásico, básicamente pensado para datos estructurados (cualitativos y/o cuantitativos), desde las típicas matrices de datos de *n* filas por *p* columnas (con *n* no superando el millón de unidades o casos y *p* no superando un par de decenas de miles de variables), hasta el nuevo marco del análisis de los datos enormes donde los datos pueden ser estructurados, semiestructurados o no estructurados, y se presentan en cientos de billones de filas y millones de variables.

Los sistemas de gestión de bases de datos tradicionales RDBMS basados en SQL son incapaces de procesar este nuevo contexto de los *big data*, lo que ha hecho necesaria la aparición de nuevos paradigmas en el tratamiento de datos masivos como MapReduce, Hadoop, NoSQL o *cloud computing*; términos que han sido presentados de forma divulgativa en este artículo.

La analítica de los *big data* no se ha podido desarrollar por cuestión de espacio. Sólo queda comentar que la estadística de los *big data* tiene, como es lógico, sus cimientos en la Estadística Clásica, pero requiere un enfoque de la llamada Ingeniería Estadística, donde las técnicas de Inteligencia Artificial proporcionan una base imprescindible y crítica para el futuro analítico de los datos masivos.

Todo lo anterior nos lleva a afirmar con rotundidad que *big data* representa un nuevo paradigma dentro del Análisis de Datos.

Mención aparte merecen las cuestiones legales y éticas de los *big data* que han sido someramente expuestas. Con ellas se abre un campo de investigación fascinante y multidisciplinar donde la colaboración entre distintos centros de una misma Universidad podría dar unos resultados absolutamente espectaculares en el medio plazo. ▮

Bibliografía

Akerkar R. (Ed.). (2014). *Big data computing*. CRC Press.

Cerchiello P, Giudici P. (2014). *How to measure the quality of financial tweets* (No. 069). University of Pavia, Department of Economics and Management.

Dean J, Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Dean J, Ghemawat S. (2010). *U.S. Patent No. 7,650,331*. Washington, DC: U.S. Patent and Trademark Office.

Ghemawat S, Gobioff H, Leung ST. (2003). The Google file system. In *ACM SIGOPS Operating Systems Review* (Vol. 37, No. 5, pp. 29-43). ACM.

Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, 98-115.

Kambatla K, Kollias G, Kumar V, Grama A. (2014). Trends in big data analytics. *Journal of Parallel Distributed Computation*, 74, 2561-2573.

McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.

Mitchell I, Locke M, Wilson M, Fuller A. (2012). The White Book of Big Data. The definitive guide to the revolution in business analytics. Fujitsu. http://www.fujitsu.com/global/Images/WhiteBookofBigData.pdf

Nunan D, Di Domenico M. (2013). Market research and the ethics of big data. *International Journal of Market Research, 55*(4), 505-520.

Pokorny J. (2013). NoSQL databases: a step to database scalability in web environment. *International Journal of Web Information Systems, 9*(1), 69-82.

Pulido E. (2014). Big data: solución o problema. *Lección inaugural curso 2014-2015 en la Universidad Autónoma de Madrid.* http://arantxa.ii.uam.es/~epulido/bigdata.pdf

Zicari RV. (2014). Big Data: Challenges and Opportunities. In Akerkar R. (Ed.). (2014). *Big data computing.* CRC Press.