novática

Revista de la Asociación de Técnicos de Informática

Novática, revista fundada en 1975 y decana de la prensa informática española, es el órgano oficial de expresión y formación continua de **ATI** (Asociación de Técnicos de Informática), organización que edita también la revista **REICIS** (Revista Española de Innovación, Calidad e Ingeniería del Software).

<http://www.ati.es/novatica/>
<http://www.ati.es/reicis/>

ATI es miembro fundador de **CEPIS** (*Council of European Professional Informatics Societies*), representa a España en **IFIP** (*International Federation for Information Processing*) y es miembro de **CLEI** (*Centro Latinoamericano de Estudios de Informática*) y de **CECUA** (*Confederation of European Computer User Associations*). Asimismo tiene un acuerdo de colaboración con **ACM** (*Association for Computing Machinery*) y colabora con diversas asociaciones informáticas españolas.

Coordinación Editorial, Redacción Central y Redacción ATI Madrid
Gutierre de Cetina 24, 28017 Madrid • Tlfn.914029391 <novatica@ati.es>
Administración y Redacción ATI Cataluña
Calle Avila 50, 3a planta, local 9, 08005 Barcelona
Tlfn.934125235 <secregen@ati.es>
Redacción ATI Andalucía <secreand@ati.es>
Redacción ATI Galicia <secregal@ati.es>
Suscripción y Ventas <novatica.subscripciones@atinet.es>
Publicidad Gutierre de Cetina 24, 28017 Madrid
Tlnf.914029391 <novatica@ati.es>
Imprenta: Impresión Offset Derra S.L., Llull 41, 08005 Barcelona.
Depósito legal: B 15.154-1975 -- ISSN: 0211-2124; CODEN NOVAEC
Portada: "La decisión" - Concha Arias Pérez / © ATI
Diseño: Fernando Agresta / © ATI 2003

sumario

Nº 237, julio-octubre 2016, año XLII

Monografía del próximo número: "Seguridad digital"

Presentación. *Big Data:* Conceptos y aplicaciones

José María Gómez Hidalgo[1], Ricardo Baeza-Yates[2]
[1]Analytics Manager de Pragsis Bidoop, coordinador de la sección técnica "Acceso y recuperación de información" de Novática; [2]Director de Tecnología (CTO) de NTENT, Catedrático part-time de la Universitat Pompeu Fabra y la Universidad de Chile

<jmgomez@pragsis.com>, <rbaeza@acm.org>

1. Introducción al *Big Data*

Big Data, como *Cloud*, *Machine Learning* y otras palabras clave de moda, aparece por todas partes hoy en día. Está en la prensa y en las noticias, en publicaciones de informática y de investigación, tiene sus propias conferencias, y no es raro que un amigo no informático nos pregunte, en una charla de café: "Oye, ¿qué es eso del *Big Data*?".

Existen dos visiones de lo que es el *Big Data*, una general y otra técnica. Por un lado, el público en general es consciente, a través de las noticias, de que las grandes compañías de Internet, y también los gobiernos, están acumulando gran cantidad de datos sobre nuestro comportamiento, con fines comerciales (la publicidad dirigida), o para reforzar la seguridad.

Para alguien de la calle, decir *Big Data* es decir "Gran Hermano". En parte al menos, los responsables de esta visión son los medios de comunicación, que dan una visión sesgada y alarmista de un conjunto de tecnologías que, a todos los efectos, se pueden considerar casi imprescindibles en la gestión de datos de hoy y mañana en cualquier tipo de organización o sector.

Por otro lado, *Big Data* hace referencia a volúmenes de datos tan grandes que han requerido el diseño de nuevos algoritmos y estrategias para manejarlos. Por asimilación, entendemos que también hace referencia al conjunto de tecnologías que, basadas en la idea de la distribución de los datos y los recursos para procesarlos, permiten una gestión del volumen, procesamiento y análisis de datos con una profundidad y escalabilidad sin precedentes, y virtualmente ilimitada.

Las tecnologías *Big Data*, especialmente a través del ecosistema Hadoop, cubren la adquisición, el almacenamiento, el acceso, el procesamiento (ya sea por lotes - *batch* o en tiempo real), la analítica y la visualización de inmensas cantidades de datos.

Una de las ideas más celebradas de *Big Data* es la de distribuir los datos y los programas de tal manera que los programas no tratan de acceder a los datos, provocando un cuello de botella, sino que los datos se ubican de manera natural a lo largo de un *clúster* o grupo de ordenadores, en los lugares donde residen los programas que los precisan.

La distribución de datos se realiza de forma replicada, aumentando asimismo la fiabilidad y disminuyendo el riesgo de pérdida de información. El uso de múltiples servidores de relativas bajas prestaciones individuales en un *clúster* hace además que el factor de disminución del coste sea un beneficio estratégico de esta tecnología.

Éstas son sólo algunas de las ventajas que proporciona el *Big Data*. En esta monografía pretendemos que se muestren éstas y muchas otras a través de los artículos que la componen.

2. Esquema de la monografía

Nuestra idea al preparar esta monografía, ha sido dar a sus lectores la oportunidad de obtener una visión global de cómo funcionan estas tecnologías, de cómo afectan a los enfoques tradicionales de análisis de datos, y cómo están contribuyendo a promover y a posibilitar la transformación digital de la empresa.

Para ello, contamos con una serie de artículos que tocan distintas áreas del *análisis de datos* (características de los datos, preprocesamiento, *streams* o flujos, arquitecturas), varias *aplicaciones* (sistemas de recomenda-

Editores invitados

José María Gómez Hidalgo ha sido profesor e investigador en la Universidad Complutense de Madrid y la Universidad Europea de Madrid, durante 16 años, y Director de I+D en la empresa multinacional de seguridad de Optenet (ahora Allot Communications). Actualmente es *Analytics Manager* en la empresa de *Big Data* Pragsis Technologies, donde realiza consultoría de *Big Data Analytics* para empresas de banca y finanzas, turismo, industria, medios de comunicación, etc. A nivel de investigación, José María se centra principalmente en el Procesamiento del Lenguaje Natural y Aprendizaje Automático sobre datos textuales, con aplicaciones en el acceso a la información de actualidad y biomédica, y la Recuperación de Información con Adversario (filtrado de correo basura, filtrado Web y protección del menor en Internet). Es autor de numerosos trabajos de investigación en estas áreas, coordinador de la sección técnica de "Acceso y recuperación de información" de ***Novática***, y fue coeditor de la monografía del número 185 de la revista, titulada "*Buscando en la Web del Futuro*". Asimismo, José María es miembro de SEPLN, ACM y ATI.

Ricardo Baeza-Yates es Director de Tecnología (*Chief Technology Officer*, CTO) de NTENT <http://www.ntent.com>, una compañía de tecnología de búsqueda semántica basada en New York y California, desde junio de 2016. Antes fue Vicepresidente de Investigación de Yahoo Labs, primero en Barcelona y luego en Sunnyvale, California, desde enero del 2006 hasta febrero del 2016. Entre 2008 y 2012 también supervisó Yahoo Labs Haifa y entre 2012 y 2015 estuvo a cargo de Yahoo Labs Londres. Hasta 2005 fue director del Centro de Investigación de la Web <http://www.cwr.cl/> en el Departamento de Ciencias de la Computación <http://www.dcc.uchile.cl> de la Escuela de Ingeniería <http://www.fcfm.uchile.cl> de la Universidad de Chile <http://www.uchile.cl>; y catedrático ICREA y fundador del Grupo de Investigación de la Web <http://wrg.upf.edu> en el Dept. de Tecnologías de la Información y las Comunicaciones <http://www.upf.edu/dtecn> de la Universitat Pompeu Fabra <http://www.upf.ed/> en Barcelona (España). Mantiene vínculos con ambas universidades como catedrático jornada parcial. Obtuvo su doctorado en Ciencia de la Computación en la Universidad de Waterloo (Canadá) en 1989. Sus intereses de investigación incluyen algoritmos y estructuras de datos, recuperación de información, búsqueda y minería de datos en la Web además de *data science* y visualización de datos. Tiene más de 500 publicaciones donde destaca el libro *Modern Information Retrieval*, cuya segunda edición fue publicada en 2011 por Addison-Wesley. Ha obtenido varios premios, incluyendo distinciones de la Organización de Estados Americanos, el Centro Latinoamericano de Estudios en Informática, el Instituto de Ingenieros de Chile y la Universidad de Waterloo. Es *Fellow* de la ACM y de la IEEE.

ción, análisis Web, redes sociales, analítica de texto), un caso de uso real y práctico que nos muestra una visión plenamente aplicada y sus beneficios de *negocio*, además de un artículo específicamente dirigido hacia la *privacidad*.

En lo que se refiere al *análisis de datos*, ***Ricardo Baeza-Yates***, uno de los editores de esta monografía, saca partido de su gran experiencia en la búsqueda y la minería de datos en la Web, para exponernos en su artículo todos los problemas de los datos de la Web, como por ejemplo su calidad, escalabilidad, los sesgos, la dispersión y la privacidad de los mismos.

A continuación, ***Salvador García***, ***Sergio Ramírez-Gallego***, ***Julián Luengo*** y ***Francisco Herrera*** nos describen los requisitos de los algoritmos de preprocesamiento de datos (limpieza, normalización, transformación, etc.) en entornos *Big Data*. También nos muestran un caso de uso en la tarea concreta de selección de atributos, donde se describe el algoritmo Fast-mRMR, desarrollado por su grupo para garantizar la escalabilidad y la eficiencia en esta tarea.

En el mismo ámbito, ***Albert Bifet*** y ***Jesse Read*** realizan una discusión de los requisitos necesarios para procesar grandes cantidades de datos en tiempo real (*streams* o flujos), con especial atención a una de sus fuentes principales: la Internet de la Cosas (*Internet of Things*). En particular, nos explican cómo los algoritmos usados tradicionalmente para la resolución de la evolución de conceptos, la clasificación y el agrupamiento, se pueden modificar para operar en ámbitos *Big Data*. También nos introducen algunas herramientas de código abierto que se pueden usar para procesar grandes cantidades de datos en tiempo real.

Adicionalmente, ***Alexey Cheptsov*** y ***Bastian Koller*** presentan su trabajo sobre la evolución de los sistemas de supercomputación (*High Performance Computing*) para poder hospedar tecnologías *Big Data*, con la idea de traer lo mejor de ambos mundos: la rapidez de la supercomputación y la escalabilidad del *Big Data*. Dado que las plataformas de supercomputación no pueden soportar por defecto arquitecturas *Big Data*, los autores proponen, por ejemplo, un cambio en los sistemas de paso de mensajes que afecta a la arquitectura de estas plataformas, desarrollado en el proyecto europeo Juniper.

En lo que se refiere a las *aplicaciones*, ***David C. Anastasiu***, ***Evangelia Christakopoulou***, ***Shaden Smith***, ***Mohit Sharma*** y ***George Karypis*** nos ofrecen un panorama de los avances más recientes en la adaptación de los sistemas de recomendación a entornos *Big Data*, con el fin de dotarles de la necesaria escalabilidad para realizar recomendaciones a millones de usuarios en cuestión de segundos. No debemos olvidar que la personalización, la recomendación y la publicidad son áreas críticas para el negocio hoy en día, ya sea digital o físico.

Por otra parte, ***Weizhong Zhao***, ***Gang Chen***, ***Venkata Swamy-Martha*** y ***Xiaowei Xu*** presentan un artículo muy interesante en el que se discuten los algoritmos de análisis de grafos, habitualmente empleados en el ámbito de la Minería de Redes Sociales. Debido a las deficiencias de los algoritmos tradicionales, los autores han tenido que extender un algoritmo para el análisis social a un entorno paralelo; este nuevo algoritmo, llamado PSCAN, es una contribución importante para esta monografía.

Asimismo, el otro editor de esta monografía, ***José María Gómez Hidalgo***, nos presenta las bondades que ofrece una de las tecnologías más extendidas en el ámbito *Big Data* (la plataforma Spark) para el análisis de información textual. En su artículo, se discuten dos de los elementos que permiten construir de manera rápida sistemas de clasificación de texto, que son la extensión Spark Streaming y la biblioteca Spark MLlib. Adicionalmente, nos demuestra la tecnología por medio de un ejemplo de aplicación al análisis de sentimiento de texto.

Con el fin de demostrar el potencial de las tecnologías *Big Data* para transformar el *negocio*, contamos con el artículo proporcionado por ***Xavier Ferrándiz Bofill*** y ***Alberto Alejo Marcos***, que nos detalla las motivaciones para realizar esta transformación en la Corporación Catalana de Medios Audiovisuales. Xavier y Alberto también nos explican qué casos de uso concretos han abordado hasta el momento, con qué objetivos y tecnologías, y cuáles son los resultados operativos que han obtenido.

Es importante subrayar que los beneficios obtenidos en esta iniciativa incluyen lecciones organizativas que les permiten avanzar en una transformación digital completa en el análisis y procesamiento de la información, con beneficios específicos para el negocio.

Para finalizar, y dada la preocupación en los medios y en los usuarios finales por los aspectos relacionados con la *privacidad*, hemos pedido a ***José María del Álamo Ramiro***, ***Esmeralda Saracibar Serradilla*** y ***Emilio Aced Félez***, que contribuyan a esta monografía con un análisis de la privacidad en entornos *Big Data*.

En su artículo, nos ofrecen una panorámica general sobre los principales riesgos que la analítica masiva de datos entraña para la privacidad y la protección de datos personales, y también presentan varios enfoques jurídicos, organizativos y técnicos para sacar el máximo provecho de los datos mientras se respetan los derechos individuales.

Referencias útiles

A continuación se proporcionan algunas referencias sobre *Big Data*, acompañadas por artículos científicos recientes y enlaces con herramientas útiles para profundizar en los distintos enfoques sobre el tema.

Libros y revistas

■ **Bart Baesens.** "*Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*". Wiley, 2014. ISBN: 978-1-118-89270-1. Un libro orientada a sacar partido de los datos y del *Big Data* en el marco empresarial.

■ **Viktor Mayer-Schönberger, Kenneth Cukier.** "*Big Data: A Revolution That Will Transform How We Live, Work, and Think*". Houghton Mifflin Harcourt, 2013. ISBN-10: 1848547927. Este libro presenta una visión general de cómo estas tecnologías afectarán a la sociedad.

■ **Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills.** "*Advanced Analytics with Spark: Patterns for Learning from Data at Scale*". O'Reilly Media, 2015. ISBN-10: 1491912766. En este libro se presenta cómo desarrollar aplicaciones de analítica de datos con Spark.

■ **Tom White.** "*Hadoop: The Definitive Guide, 4th Edition. Storage and Analysis at Internet Scale*". O'Reilly, 2015. ISBN-10: 1491901632. Un libro que detalla la arquitectura y funcionamiento de Hadoop.

Artículos y estudios

■ **Richard M. Burton, Dolly Mastrangelo, Fabrizio Salvador (editores).** "Big Data and Organization Design Special Issue". *Journal of Organization Design, Vol. 3, No. 1, 2014*. Monografía sobre como *Big Data* afecta a la organización de las empresas.

■ **Jeffrey Dean, Sanjay Ghemawat.** "MapReduce: simplified data processing on large clusters". *Communications of the ACM - 50th anniversary issue: 1958 - 2008, Volume 51 No 1*, enero 2008, pp. 107-113. Descripción del paradigma de programación funcional distribuída MapReduce.

■ **Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung.** "The Google File System". *Proceedings of the nineteenth ACM Symposium on Operating Systems Principles*, ACM, New York, NY, EEUU, pp. 29-43, 2003. El artículo clásico en el que se presenta el sistema de archivos de Google, que inspiró Hadoop.

Eventos

■ **Apache Big Data Europe 2016.** 14-16 de noviembre, 2016, Sevilla, España. La conferencia de la Fundación Apache sobre *Big Data* en Europa. <http://events.linuxfoundation.org/events/apache-big-data-europe>.

■ **Big Data Spain 2016.** 17-18 de noviembre, 2016, Madrid, España. La principal conferencia técnica sobre *Big Data* en España. <https://www.bigdataspain.org/>.

■ **IEEE BigData 2016.** 5-8 de diciembre, 2016. Washington D.C., EEUU. Una de las más importantes conferencias científicas sobre *Big Data*. <http://cci.drexel.edu/bigdata/bigdata2016/index.html>.

■ **Data Beers Madrid y Barcelona.** Los dos *meetups* más importantes sobre *Data Analytics* y *Big Data* en España. <http://databeers.tumblr.com/>, <http://databeersbcn.com/>.

Distribuciones y software

■ **Apache Hadoop.** La página oficial del proyecto Hadoop y de muchos otros que conforman su ecosistema. <http://hadoop.apache.org/>.

■ **Cloudera.** Probablemente la distribución de Hadoop más popular. <http://es.cloudera.com/>.

■ **Hortonworks.** Otra distribución muy popular de Hadoop. Hortonworks contribuye con frecuencia a proyectos del ecosistema Hadoop. <http://es.hortonworks.com>.

■ **MapR.** Otra de las distribuciones de Hadoop más extendidas. <https://www.mapr.com/>.

■ **SolR.** La herramienta de búsqueda del ecosistema Hadoop. <ttps://lucene.apache.org/solr/>.

Proyectos y organizaciones a seguir

■ **Indra.** Es miembro fundador de la principal iniciativa de investigación europea en *Big Data*: <http://www.indracompany.com/es/noticia/indra-miembro-fundador-principal-iniciativa-investigacion-europea-big-data>. INDRA ha colaborado en la puesta en marcha de una Asociación Público-Privada (APP) que marcará la estrategia de I+D+i sobre macrodatos.

■ **Pragsis Bidoop.** Participa en un proyecto europeo de seguridad: <http://pragsis.com/blog/pragsis_bidoop_participa_en_un_proyecto_europeo_de_seguridad>. Pragsis Bidoop participa en el proyecto DANTE que tiene como fin la detección de actividades terroristas en el ámbito digital usando tecnologías *Big Data*.

■ **Banco Santander.** *El Big Data aportará 2.500 millones en ingresos al Banco Santander*: <http://www.muycomputerpro.com/2016/03/03/el-big-data-aportara-2-500-millones-en-ingresos-al-banca-santander>. Una noticia que indica cuáles son los planes de futuro del Grupo Santander para su transformación digital.

■ **BBVA.** Proyecto *Big Data* del Centro de Innovación BBVA: <http://www.centrodeinnovacionbbva.com/proyectos/big-data>. Toda la información y conocimiento que se está generando en torno a este tema en BBVA.

■ **Unión Europea.** Proyecto Juniper. Un proyecto del Séptimo Programa Marco de la Unión Europea, para el desarrollo de una arquitectura que permita el procesamiento de grandes volúmenes de datos en tiempo real. <http://www.juniper-project.org/>.

■ **Telefónica, Universitat de Barcelona.** Telefónica y la *Universitat de Barcelona* aplican por primera vez *Big Data* al mundo de la gastronomía con Appetit. Una iniciativa para aplicar las tecnologías del *Big Data* al mundo de la gastronomía. <https://www.telefonica.com/es/web/sala-de-prensa/-/telefonica-y-la-universitat-de-barcelona-aplican-por-primera-vez-big-data-al-mundo-de-la-gastronomia-con-appetit>.