Rev. X del dd/mmm/aaaa



Economía del Cambio
Tecnológico – 5306.02
Innovación Tecnológica

# ENTENDIENDO EL BIG DATA: ANTECEDENTES, ORIGEN Y DESARROLLO POSTERIOR

Mikel Niño, Arantza Illarramendi

Departamento de Lenguajes y Sistemas Informáticos. Universidad del País Vasco (UPV/EHU). Donostia-San Sebastián.
mikel.nino@ehu.es

Recibido: 1/oct/2015 -- Aceptado: 16/dic/2015 - DOI: http://dx.doi.org/10.6036/NT7835

## 1. ANTECEDENTES: LA ANALÍTICA DE DATOS APLICADA A LOS NEGOCIOS

Uno de los primeros términos clave que nos encontramos en el campo del análisis de datos de negocio es el de ***Business Intelligence***, cuya primera referencia se remonta a 1958 a cargo de Hans Peter Luhn [1], investigador de IBM, aunque en dicha referencia el término aún estaba alejado de la evolución que sufrió posteriormente con la progresiva informatización de los procesos de negocio. Tras el desarrollo en dicho campo en los años siguientes, es en los años 80 cuando se consolida la idea de *Business Intelligence* (principalmente con la propuesta de Howard Dresner [2]) para referirse a un conjunto de sistemas *software* para el apoyo a la toma de decisiones de negocio, basados en la recogida de análisis de *hechos* o datos. Estos sistemas se enfocan en un análisis de tipo descriptivo, consultando datos históricos de manera agregada y cruzando indicadores para obtener una mejor visión de lo que ha pasado y está pasando en la organización.

El enfoque de *Business Intelligence* deja al margen, por tanto, un análisis de tipo predictivo que busca la extracción de conocimiento de los datos en forma de patrones, tendencias o modelos que permitan una cierta certeza sobre el resultado de potenciales acciones futuras [3]. Para denominar este tipo de análisis, a finales de los 80 surge la expresión ***Data Mining*** *(minería de datos).* El origen del término proviene de la analogía con las técnicas de minería en las que se extrae un material valioso (en este caso, conocimiento) a partir de yacimientos (bancos de datos). Junto al término *Data Mining*, como probablemente el más conocido y utilizado para referirse a este tipo de análisis de entre un conjunto de expresiones similares [4], en la misma época empieza también a utilizarse la expresión ***Knowledge Discovery in Databases*** **(KDD)**. De hecho, el primer seminario académico sobre esta materia se organiza en 1989 [5], que en 1995 deriva en la *First International Conference on Knowledge Discovery and Data Mining* [6].

El desarrollo de proyectos de *Data Mining* para la búsqueda y explotación de patrones en bancos de datos, empleando técnicas de ***Machine Learning*** *(aprendizaje automático)* [7] para la construcción de modelos predictivos, comienza a extenderse en contextos de negocio durante la década de 1990, destacando su aplicación en el sector bancario y de seguros, donde se busca emplear los resultados de este tipo de análisis para facilitar procesos de toma de decisiones ligadas a productos de dichos sectores (por ejemplo, la detección de fraudes a compañías aseguradoras, o para la concesión o denegación de créditos).

Este tipo de aplicaciones de *Data Mining* nos permite explicar el origen del concepto de ***Data Science*** *(ciencia de datos).* Dicho término surgió a principios de milenio para denominar una propuesta de revisión de las áreas técnicas en torno a la Estadística [8], y así adecuarse mejor a las prácticas de análisis de datos que venían desarrollándose en la época, principalmente con el desarrollo del *Data Mining* y su aplicación en diferentes contextos de negocio, y con la progresiva informatización de la recolección y análisis de datos. A partir de dicha propuesta, el concepto de *Data Science* ha ido desarrollándose como la integración de principios de las diferentes disciplinas (estadística y matemáticas, informática y computación, fundamentos del área específica de aplicación) que sustentan la práctica moderna del análisis de datos y guían la extracción de conocimiento de los mismos. De esta manera, podemos entender el *Data Mining* como la extracción en sí de ese conocimiento a través de herramientas y técnicas que incorporan los principios de la ciencia de datos [9].

***Business Intelligence***: Análisis descriptivo de datos, en el que se consultan y visualizan de manera agregada datos provenientes de diferentes indicadores de negocio para obtener una visión de lo que ha pasado y lo que está pasando.

***Data Mining*** (minería de datos): Extracción de conocimiento (patrones, tendencias, modelos) en bancos de datos, enfocado a un análisis de tipo predictivo. El concepto de *Knowledge Discovery in Databases* (KDD) comprende un área similar. En muchas ocasiones se usan indistintamente, aunque también se usa el término *Data Mining* para referirse específicamente a la etapa analítica dentro del KDD.

***Machine Learning*** (aprendizaje automático): Área de la Informática enfocada al estudio y creación de algoritmos capaces de tomar decisiones (hacer predicciones) basadas en la experiencia acumulada en una batería de casos (bancos de datos) resueltos con éxito. Sus fundamentos se derivan de la inteligencia artificial, la estadística y la optimización matemática. El uso de algoritmos de *Machine Learning* representa una parte destacada dentro de las técnicas de *Data Mining*.

***Data Science***: Conjunto de principios y fundamentos, tanto científicos como aplicados, que guían la extracción de conocimiento de los datos, y en los que se basan las herramientas, técnicas y procedimientos de *Data Mining*.

***Big Data***: Conjunto de tecnologías específicas, de entre las utilizadas en *Data Mining* o *Business Intelligence*, que facilitan el procesamiento y análisis de datos cuando su volumen o su complejidad de tratamiento es excesivamente grande para las capacidades de cómputo de una máquina de uso convencional.

*Glosario de términos habituales relacionados con el Big Data*

## 2. ORIGEN DE LAS TECNOLOGÍAS *BIG DATA*: EL MODELO MAP-REDUCE

La necesidad de las tecnologías que posteriormente se han definido como *Big Data* tiene su origen en el uso de *Data Mining* por parte de las grandes empresas tecnológicas que surgen (principalmente en Silicon Valley) con la eclosión y popularización del *World Wide Web* entre finales de los 90 y principios del 2000. El problema al que se enfrentaron dichas empresas no difiere del que hemos descrito en el caso de compañías financieras o aseguradoras: potenciar su negocio explotando sus bancos de datos. El elemento clave y diferencial surge al comparar la dimensión de los datos en ambos escenarios: mientras que en los ejemplos anteriores la cantidad de datos era procesable usando las herramientas y las capacidades de máquinas convencionales, en el caso de las grandes empresas en torno a la Web, por el contrario, el gran volumen de datos a analizar hacía inviable en la práctica su procesamiento usando las técnicas tradicionales. De hecho, es también en esta época cuando se comienza a apuntar al volumen, velocidad y variedad de los datos [10] (modelo posteriormente conocido como "3 V") como aspectos clave dentro de una estrategia para la óptima gestión de los datos en contextos de negocio.

El principal ejemplo de este nuevo escenario lo tenemos en la empresa Google, que tuvo que enfrentarse a este problema inicialmente [11] para procesar de manera eficiente su algoritmo *PageRank* [12] aplicado a grandes cantidades de datos provenientes del análisis de multitud de sitios web. Frente a otras soluciones y estrategias ya existentes para el procesamiento paralelo de grandes volúmenes de datos, basadas en el uso de máquinas de grandes prestaciones (*High-Performance Computing*, HPC) y con gran cantidad de procesadores o núcleos, Google optó por desarrollar una solución propia con una estrategia diferente, buscando la automatización eficiente de gran parte del trabajo que suponía repartir la tarea de procesar grandes volúmenes de datos de manera distribuida entre un conjunto de máquinas de prestaciones más modestas que las utilizadas para el HPC. Dicha solución fue construida sobre dos elementos fundamentales: un **sistema de ficheros distribuidos** para gestionar el almacenamiento de grandes cantidades datos de manera segmentada y replicada entre dicho conjunto de máquinas (nodos en un clúster) [13] y un *software* que proporcionaba ya implementadas de manera eficiente las tareas más complejas de las aplicaciones distribuidas que debían trabajar con datos almacenados en dicho sistema. Este software simplificaba enormemente el trabajo de creación



DYNA New Technologies
c) Mazarredo nº69 -4º 48009-BILBAO (SPAIN)
Tel +34 944 237566 – www.dyna-newtech.com - email: info@dyna-newtech.com

de programas para trabajar con datos en este tipo de contextos, al dar ya resuelto al programador gran parte del problema. Dicho software y, por extensión, el modelo de programación que facilitaba recibieron el nombre de **MapReduce** [14], marcando el hito principal en el origen de las tecnologías *Big Data*.

Gracias a la utilización este modelo, la tarea del programador de aplicaciones queda reducida a definir los detalles de dos funciones ("*map*" y "*reduce*") que corresponden a las dos etapas principales de procesamiento de los datos. La función que se define para el paso *map* es la que transforma un conjunto de datos de partida en otra serie diferente de datos intermedios en forma de (clave, valor), y la función definida para el paso *reduce* procesa de manera agrupada (según la función de agregación que se necesite) los valores de los datos intermedios correspondientes a una misma clave para producir el resultado final. La Fig.1 detalla un ejemplo ilustrativo de este modelo de programación.

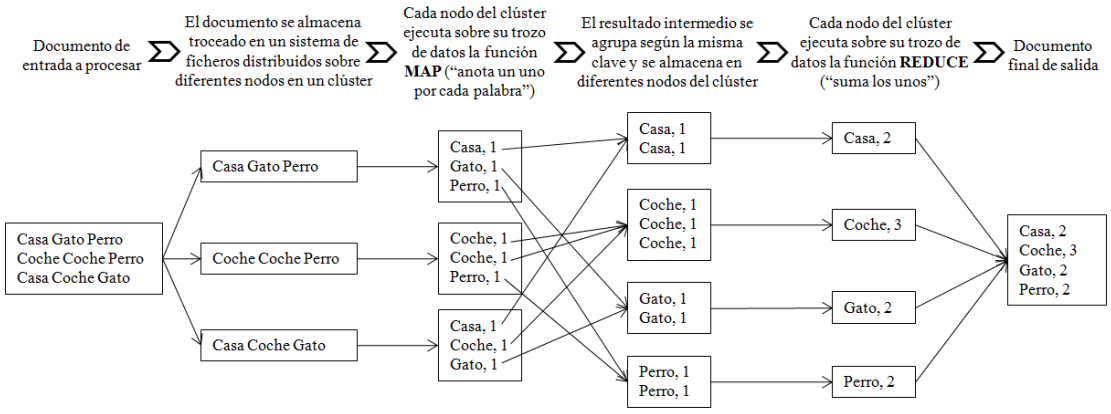


*Fig. 1. Ejemplo de proceso usando el modelo MapReduce para contar las apariciones de las palabras de un documento*

La divulgación por parte de Google de los detalles de su sistema de ficheros distribuidos y el modelo de programación MapReduce sirvió de inspiración a otros proyectos que buscaban resolver problemas similares. En particular, los artículos de Google sirvieron a Doug Cutting [15] para mejorar el proyecto de desarrollo de un motor de búsqueda web en el que se encontraba inmerso en aquella época. Este proyecto sirvió de germen para el trabajo que posteriormente Cutting desarrolló al incorporarse a Yahoo, construyendo un sistema que implementase MapReduce con la capacidad de procesar de manera distribuida el enorme volumen de datos requerido por un gran motor de búsqueda a nivel global. Así nació el sistema de código abierto **Apache Hadoop**, cuyos dos módulos principales eran el **Hadoop Distributed File System** (**HDFS**, la implementación en software libre del sistema distribuido de ficheros descrito por Google años antes) y **Hadoop MapReduce** (implementado sobre el mencionado HDFS).

## 3. DESARROLLO POSTERIOR DE LAS TECNOLOGÍAS *BIG DATA*

La disponibilidad de una solución de código abierto como Apache Hadoop facilita la adopción de estas tecnologías *Big Data*, favoreciendo al mismo tiempo la creación de herramientas adicionales sobre dicha plataforma que permitan potenciar su funcionalidad. De hecho, durante los años siguientes a la divulgación del modelo MapReduce se repite el mismo patrón de innovación, en el que los desarrollos divulgados por Google en torno al procesamiento masivo de datos sirven de inspiración para crear soluciones de código abierto que aborden los mismos problemas.

Así, en 2005 Google divulga **Sawzall** [16], un lenguaje para expresar y programar las tareas más usuales dentro de la consulta y análisis de grandes estructuras de datos sobre un modelo MapReduce. Esto inspira el desarrollo de la herramienta de código abierto **Apache Pig** y su lenguaje Pig Latin, para su uso sobre Apache Hadoop. Así como Apache Pig es inicialmente desarrollado y promovido por Yahoo (al igual que con Hadoop), desde la empresa Facebook se desarrolla inicialmente una herramienta de código abierto para fines similares, **Apache Hive**, aunque en este caso con un modo de uso más cercano al conocido lenguaje SQL para interrogar bases de datos. Del mismo modo, en 2006



DYNA New Technologies
c) Mazarredo nº69 -4º 48009-BILBAO (SPAIN)
Tel +34 944 237566 – www.dyna-newtech.com - email: info@dyna-newtech.com

Google divulga su sistema *NoSQL* de almacenamiento de datos, **Bigtable** [17], utilizado desde entonces en la mayoría de aplicaciones de Google que necesitan almacenar y procesar grandes volúmenes de datos. Esto origina el desarrollo de su equivalente de código abierto **Apache HBase** para ser utilizado con HDFS/Hadoop. Dentro de este mismo campo de las bases de datos *NoSQL*, otra de las grandes empresas que impulsa el desarrollo de tecnologías *Big Data* como es Amazon divulga su sistema de almacenamiento de datos masivos **Dynamo** [18]. Los elementos clave de dicho sistema, junto con los de Bigtable, sirven de inspiración para otro de los desarrollos de código abierto destacado en este campo, **Apache Cassandra**, que junto con **MongoDB** y el propio HBase constituyen la terna de sistemas *NoSQL* más representativos en el mercado actual [19].

Este mismo patrón se repite a finales de la pasada década con **Pregel** [20] y **Apache Giraph** para el procesamiento de datos almacenados como grafos, o con **Dremel** [21] e **Impala** para hacer más eficientes las consultas a grandes volúmenes de datos. Estos desarrollos en código abierto van conformando un "ecosistema" de herramientas *Big Data* en torno a Apache Hadoop, junto con otros componentes como por ejemplo Mahout (aprendizaje automático), Sqoop (transferencia de datos entre Hadoop y otros sistemas), Oozie (programación y monitorización de trabajos sobre Hadoop), ZooKeeper (coordinación de procesos distribuidos), etc.

En paralelo a estos avances, cada vez gana más protagonismo el impulso del *Big Data* desde las compañías tecnológicas que surgen con la eclosión de la web 2.0 y las redes sociales, como Facebook o Twitter. Estas empresas parten de una necesidad de negocio similar a la que motivó a Google a iniciar el desarrollo de tecnologías *Big Data*, aunque necesitan de herramientas específicas para sus aplicaciones particulares, como por ejemplo el manejo de datos masivos almacenados como grafos, para procesar las conexiones de usuarios en una red social, o el procesamiento de flujos ("*streams*") de datos masivos al mismo tiempo que se van generando en la red, de cara a ofrecer un análisis en tiempo real. Para este último caso de uso surgen herramientas como **Apache Storm**, creada por Nathan Marz [22] inicialmente en la empresa BackType. Esta empresa fue adquirida en 2011 por Twitter, desde donde se dio a conocer la herramienta. Marz ideó Storm utilizando una serie de abstracciones similares a las ideas de "*map*" y "*reduce*" que se utilizan en Apache Hadoop, pero en este caso enfocadas al procesamiento distribuido de *streams* de datos en tiempo real. El conocimiento práctico de estas herramientas también le sirve a Marz para concebir una aproximación genérica al diseño de sistemas *Big Data*, que bautiza como "*Lambda Architecture*" (Arquitectura Lambda) [23]. En ella se identifican los principios y elementos fundamentales con los que debe contar un sistema de este tipo, y cómo se interrelacionan para responder a diferentes necesidades de procesamiento de datos (tanto en diferido como en tiempo real) y al mismo tiempo asegurar objetivos como la escalabilidad o la tolerancia a fallos.

A medida que se van ampliando los campos de aplicación para las tecnologías *Big Data*, empieza a quedar patente que el modelo MapReduce presenta limitaciones en su eficiencia ante determinados casos de uso. Esto da pie a diferentes líneas de trabajo, unas para mejorar la estructura interna de las implementaciones del modelo MapReduce, y otras para explorar modelos alternativos de procesamiento de *Big Data* que mantengan muchos de los principios de diseño del MapReduce pero que al mismo tiempo mejoren sus prestaciones. En el primer grupo destaca la evolución de Apache Hadoop, implantada a partir de noviembre de 2011 [24], en la que se mejora su estructura interna y en particular su gestor de tareas y recursos, sustituyendo la versión anterior por una nueva llamada **YARN** [25], para ganar en eficiencia y optimización del uso del conjunto de máquinas utilizadas en el proceso de los datos. En el segundo grupo sobresale el desarrollo de **Apache Spark** [26], la pieza nuclear para el procesamiento de *Big Data* dentro de lo que se conoce como la *Berkeley Data Analytics Stack* (BDAS) [27]. Spark se basa en un uso preferente de la memoria principal en detrimento del almacenamiento en disco, y en el empleo de la abstracción de datos llamada ***Resilient Distributed Datasets* (RDD)** [28], para ofrecer mejores prestaciones que Hadoop en casos de uso como procesos iterativos o intensivos en consultas.

La Fig.2 recoge a modo esquemático los principales hitos mencionados en este repaso por los antecedentes, origen y desarrollo de las tecnologías *Big Data*.



DYNA New Technologies
c) Mazarredo nº69 -4º 48009-BILBAO (SPAIN)
Tel +34 944 237566 – www.dyna-newtech.com - email: info@dyna-newtech.com

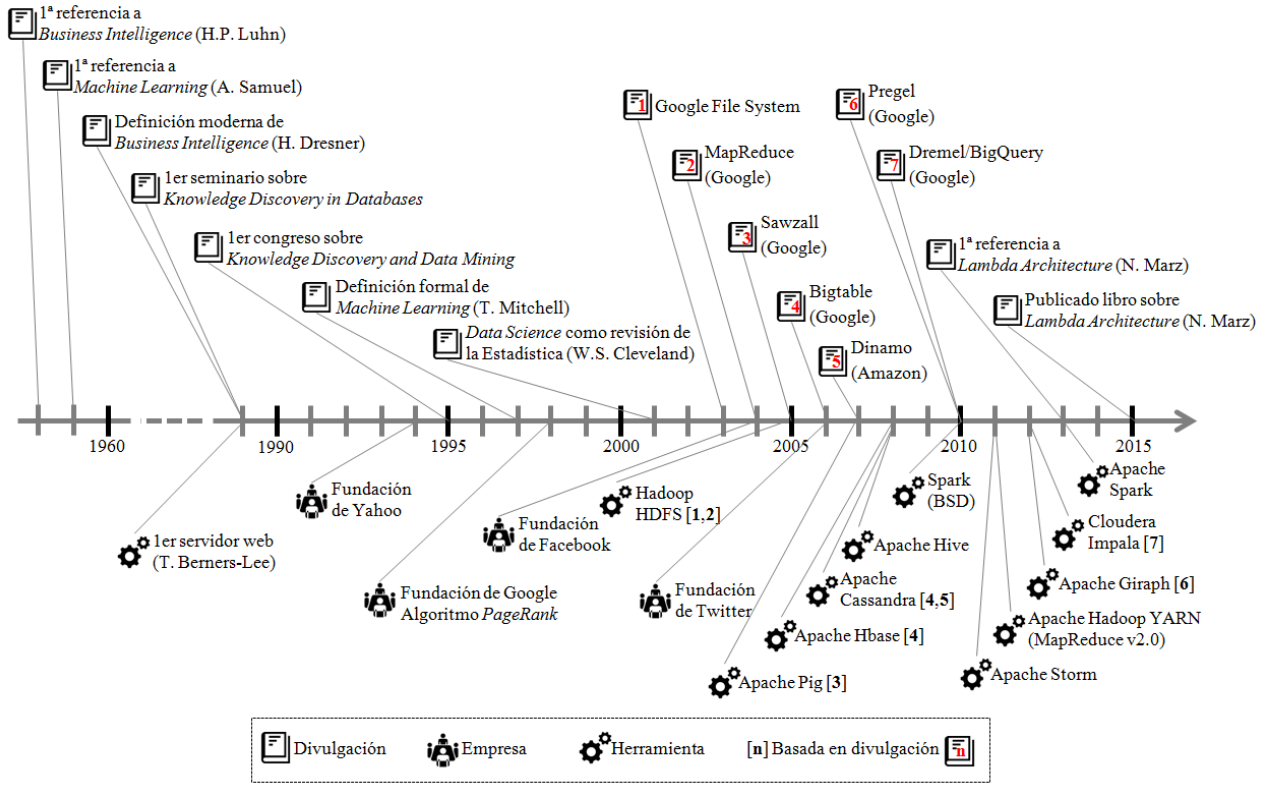


*Fig. 2. Cronología de los hitos citados en relación a los antecedentes y desarrollo de las tecnologías Big Data [29]*

## 4. CONTEXTO ACTUAL: LA POPULARIZACIÓN DEL *BIG DATA*

Varios factores explican la creciente adopción de las tecnologías *Big Data* en los últimos años. Por un lado tenemos una sensible reducción en la barrera económica de entrada a esta tecnología, gracias al amplio conjunto de herramientas de código abierto que implementan los avances en este campo, y a la disponibilidad cada vez más asequible a plataformas de ***Cloud Computing*** *(computación en la nube)* que nos facilitan el acceso a grandes conjuntos ("clústers") de máquinas en modo de servicio o alquiler dinámico, pagando sólo por el gasto en recursos que estemos haciendo en cada momento. A esto se une también el auge de conceptos como ***Internet of Things*** *(Internet de las cosas)*, que permiten la proliferación de todo tipo de dispositivos capturando y compartiendo datos para analizar diversos procesos en diferentes campos de aplicación. Todo lo anterior, junto a la enorme repercusión mediática que adquieren muchas de las innovaciones tecnológicas generadas en entornos como Silicon Valley, explica la popularización del término *Big Data* y el interés por parte del público en general, más allá de contextos especializados.

Por otra parte, los contextos que requieren de *Big Data* se han convertido en el campo de aplicación perfecto que el *Machine Learning* estaba esperando para desarrollar todo su potencial y hacer frente al tradicional problema de contar con muestras de datos de tamaño reducido. La capacidad de procesar un mayor volumen de datos (gracias a las tecnologías *Big Data*) nos dotará potencialmente de mayor flexibilidad al determinar la estrategia de uso de diferentes algoritmos predictivos.

Este foco de la aplicación del *Big Data*, centrado en su combinación con técnicas de *Data Mining* y herramientas de análisis predictivo [30], conlleva una frecuente confusión de términos al popularizarse estas tecnologías. Dado que el público en general ha empezado a conocer la minería de datos y la analítica predictiva de la mano de su aplicación en las empresas que desarrollan el *Big Data* (usos por tanto donde sí se requiere el empleo de tecnologías *Big Data* por la



DYNA New Technologies
c) Mazarredo nº69 -4º 48009-BILBAO (SPAIN)
Tel +34 944 237566 – www.dyna-newtech.com - email: info@dyna-newtech.com

gran cantidad de datos a procesar), en muchos casos los conceptos acaban entremezclándose y la expresión *Big Data* acaba usándose erróneamente para denominar a toda aplicación de la minería de datos.

## 5. CONCLUSIÓN

Como hemos visto, tanto la aplicación del análisis de datos a necesidades de negocio, como la creación de modelos predictivos a partir del uso de minería de datos, no son ideas en absoluto recientes, y su práctica lleva desarrollándose desde hace décadas. Lo que sí está claro es que la aparición del *Big Data* ha supuesto un renovado impulso durante la última década para la minería de datos y las diferentes técnicas implicadas en su puesta en práctica. Los sectores de aplicación se multiplican y ya no son sólo los negocios web los que se interesan por estas tecnologías, sino que otros muchos, como por ejemplo las empresas industriales [31], dedican recursos estratégicos a evaluar las posibilidades de negocio que se abren con la utilización de estas tecnologías. Esta aplicación de la minería de datos combinada con tecnologías *Big Data* también se extiende a otros sectores como la salud [32] o el transporte [33].

De todas maneras, a pesar del uso ya popularizado del término *Big Data*, debería quedar claro que **no toda la minería de datos es *Big Data***, y deberíamos diferenciar si el problema a resolver requiere o no de esas tecnologías específicas para procesar y analizar grandes cantidades de datos. También puede resultar engañoso el uso del adjetivo "*big*" y llevarnos a pensar que es únicamente el volumen de datos lo que implica la necesidad de una tecnología especial, cuando el uso o tipo de análisis que queramos realizar con esas cantidades masivas de datos es un factor mucho más determinante para descartar sistemas más tradicionales para la gestión de los datos. De hecho, la limitación principal de un sistema relacional de bases de datos para enfrentarse a estos problemas no está tanto en la capacidad de almacenamiento en sí, sino en que dichos sistemas están generalmente diseñados para realizar eficientemente transacciones de actualización y búsqueda de datos y consultas que recuperan cantidades limitadas de datos, y no tanto para un análisis profundo de datos masivos que extraiga patrones o modelos predictivos en un tiempo razonable [34].

Así pues, la necesidad o no de tecnologías *Big Data* dependerá del tipo de uso que se quiera hacer de los datos y del resultado que se espera de dicho análisis. Habrá muchos escenarios de negocio en los que los sistemas de bases de datos relacionales o el uso de "*data warehouses*", combinados con el abaratamiento en costes de almacenamiento y el aumento de las capacidades de discos y memorias, sean respuesta suficiente. En cualquier caso, aunque la mayoría de casos que nos rodean en diversos campos de aplicación pueden no tener esa necesidad de *Big Data* en particular, seguramente sí la tengan de una estrategia clara para la explotación de sus activos de datos a través del *Data Mining* que permita la generación de nuevos modelos de negocio. Sea cual sea la tecnología concreta más adecuada para ello, la oportunidad está servida.

## PARA SABER MÁS

[1] Luhn HP. "A Business Intelligence System". *IBM Journal of Research and Development*. October 1958. Vol.2-4. p.314-319. DOI: http://dx.doi.org/10.1147/rd.24.0314

[2] Martens C. "BI at age 7" [on-line]. *Computerworld*. 23 October 2006. http://www.computerworld.com/article/2554088/business-intelligence/bi-at-age-17.html [accessed: 24 September 2015]

[3] Schmarzo B. "Business Analytics: Moving From Descriptive To Predictive Analytics" [on-line]. *InFocus blog – EMC Global Services*. https://infocus.emc.com/william_schmarzo/business-analytics-moving-from-descriptive-to-predictive-analytics/ [accessed: 24 September 2015]

[4] Han J, Kamber M. *Data Mining: Concepts and Techniques*. 2nd Edition. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2006. 770p. ISBN: 978-1-55860-901-3

[5] Piatetsky-Shapiro G. "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop". *AI Magazine*. January 1991. Vol.11-5. p.68-70. DOI: http://dx.doi.org/10.1609/aimag.v11i4.873

[6] *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*. Fayyad UM, Uthurusamy R (eds.). AAAI Press, 1995. 359 p. ISBN: 978-0-929280-82-0

[7] Mitchell TM. *Machine Learning*. 1st Edition. New York, NY, USA: McGraw-Hill, 1997. 432p. ISBN: 978-0-07-042807-2

[8] Cleveland WS. "Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics". *International Statistical Review*. April 2001. Vol.69-1. P.21-26. DOI: http://dx.doi.org/10.2307/1403527

[9] Provost F, Fawcett T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. 1st Edition. Sebastopol, CA, USA: O'Reilly Media, 2013. 414p. ISBN: 978-1-449-36132-7

[10] Laney D. "3-D Data Management: Controlling Data Volume, Velocity and Variety" [on-line]. *Gartner Blog Network*. 6 February 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [accessed: 24 September 2015]

[11] Leskovec J, Rajamaran A, Ullman JD. *Mining of Massive Datasets*. 2nd Edition. Cambridge, UK: Cambridge University Press, 2014. 476p. ISBN: 978-1-107-07723-2

[12] Page L, Brin S, Motwani R et al. "The PageRank Citation Ranking: Bringing Order to the Web". *Technical Report*. January 1998. Stanford Digital Library Technologies Project. DOI: http://dx.doi.org/10.1.1.38.5427

[13] Ghemawat S, Gobioff H, Leung ST. "The Google file system". *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP'03)*. October 2003. p.29-43. DOI: http://dx.doi.org/10.1145/945445.945450

[14] Dean J, Ghemawat S. "MapReduce: Simplified Data Processing on Large Clusters". *Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI'04)*. December 2004. p.137-150

[15] Cutting D. "Joining Cloudera" [on-line]. *Free Search*. 10 August 2009. https://cutting.wordpress.com/2009/08/10/joining-cloudera/ [accessed: 24 September 2015]

[16] Pike R, Dorward S, Griesemer R et al. "Interpreting the data: Parallel analysis with Sawzall". *Scientific Programming Journal, Special Issue on Dynamic Grids and Worldwide Computing*. October 2005. Vol.13-4. p.277-298. DOI: http://dx.doi.org/10.1155/2005/962135

[17] Chang F, Dean J, Ghemawat S et al. "Bigtable: A Distributed Storage System for Structured Data". *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06)*. November 2006. p.205-218.

[18] DeCandia G, Hastorun D, Jampani M et al. "Dynamo: Amazon's Highly Available Key-value Store". *Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP'07)*. October 2007. p.205-220. DOI: http://dx.doi.org/10.1145/1294261.1294281

[19] Asay M. "MongoDB, Cassandra, and HBase -- the three NoSQL databases to watch" [on-line]. *InfoWorld*. 19 November 2014. http://www.infoworld.com/article/2848722/nosql/mongodb-cassandra-hbase-three-nosql-databases-to-watch.html [accessed: 24 September 2015]

[20] Malewicz G, Austern MH, Bik AJC et al. "Pregel: A System for Large-Scale Graph Processing". *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD'10)*. June 2010. p.135-146. DOI: http://dx.doi.org/10.1145/1807167.1807184

[21] Melnik S, Gubarev A, Long JJ et al. "Dremel: Interactive Analysis of Web-Scale Datasets". *Proceedings of the 36th International Conference on Very Large Data Bases (VLDB'10)*. September 2010. p.330-339

[22] Marz N. "History of Apache Storm and lessons learned" [on-line]. *Thoughts from the red planet*. 6 October 2014. http://nathanmarz.com/blog/history-of-apache-storm-and-lessons-learned.html [accessed: 24 September 2015]

[23] Marz N, Warren J. *Big Data: Principles and best practices of scalable real-time data systems*. 1st Edition. Shelter Island, NY, USA: Manning Publications, 2015. 328p. ISBN: 978-1-61729-034-3

[24] Murthy A. "Apache Hadoop 0.23 is Here!" [on-line]. *Hortonworks blog*. 15 November 2011. http://hortonworks.com/blog/apache-hadoop-is-here/ [accessed: 24 September 2015]

[25] Vavilapalli VK, Murthy AC, Douglas C et al. "Apache Hadoop YARN: Yet Another Resource Negotiator". *Proceedings of the 4th ACM Symposium on Cloud Computing (SoCC'13)*. October 2013. p.5:1-5:16. DOI: http://dx.doi.org/10.1145/2523616.2523633

[26] Zaharia M, Chowdhury M, Franklin MJ et al. "Spark: cluster computing with working sets". *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing (HotCloud'10)*. June 2010. p.10:1-10:7

[27] Konwinski A, Stoica I, Zaharia M. "The future of big data with BDAS, the Berkeley Data Analytics Stack" [on-line]. *O'Reilly Radar*. 18 febrero 2013. http://radar.oreilly.com/2013/02/the-future-of-big-data-with-bdas-the-berkeley-data-analytics-stack.html [accessed: 24 September 2015]

[28] Zaharia M, Chowdhury M, Das T et al. "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing". *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12)*. April 2012. p.15-28

[29] Niño M. "Chronology of antecedents, origin and development of Big Data" [on-line]. *Blog de Mikel Niño*. 24 September 2015. http://www.mikelnino.com/2015/09/cronologia-big-data.html [accessed: 24 September 2015]

[30] Levine P. "Machine Learning + Big Data" [on-line]. *Andreessen Horowitz blog*. 22 January 2015. http://a16z.com/2015/01/22/machine-learning-big-data/ [accessed: 24 September 2015]

[31] Niño M, Blanco JM, Illarramendi A. "Business Understanding, Challenges and Issues of Big Data Analytics for the Servitization of a Capital Equipment Manufacturer". *Proceedings of the 2015 IEEE International Conference on Big Data*. October 2015, p.1368-1377. DOI: http://dx.doi.org/10.1109/BigData.2015.7363897

[32] Sun Y, Xiong Y, Xu Q et al. "A Hadoop-Based Method to Predict Potential Effective Drug Combination". *BioMed Research International*. July 2014. Vol. 2014, Article ID 196858, 5 pp. DOI: http://dx.doi.org/10.1155/2014/196858

[33] Núñez A, Hendriks J, Li Z et al. "Facilitating maintenance decisions on the Dutch railways using big data: The ABA case study". *Proceedings of the 2014 IEEE International Conference on Big Data*. October 2014, p.48-53. DOI: http://dx.doi.org/10.1109/BigData.2014.7004431

[34] Jacobs A. "The Pathologies of Big Data". *Communications of the ACM*. August 2009. Vol.52-8. p.36-44. DOI: http://dx.doi.org/10.1145/1536616.1536632

[35] Grupo Spri, Gobierno Vasco. "¿Cómo puedo explotar el Big Data para crear nuevas oportunidades de negocio en mi sector?" [on-line]. *Portal Euskadinnova*. http://www.euskadinnova.net/es/enpresa-digitala/agenda/como-puedo-explotar-data-para-crear-nuevas-oportunidades-negocio-sector/9557.aspx [accessed: 24 September 2015]

## AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad (referencias FEDER/TIN2013-46238-C4-1-R y BES-2014-069367) y el Gobierno Vasco (referencia IT797-13). Los autores agradecen a la Cámara de Comercio de Gipuzkoa, a la Diputación Foral de Gipuzkoa y al Gobierno Vasco, entidades organizadoras del congreso sobre Big Data celebrado el 9 de junio de 2015 en el Parque Tecnológico de Donostia-San Sebastián [35], la invitación a Mikel Niño a impartir la ponencia que origina la elaboración del presente artículo. Los autores agradecen asimismo a los revisores sus comentarios para mejorar la calidad del trabajo.



DYNA New Technologies
c) Mazarredo nº69 -4º 48009-BILBAO (SPAIN)
Tel +34 944 237566 – www.dyna-newtech.com - email: info@dyna-newtech.com