BIG DATA: LA REVOLUCIÓN DE LOS DATOS

Enrique Puertas
Escuela de Arquitectura, Ingeniería y Diseño
UNIVERSIDAD EUROPEA DE MADRID

Vivimos inmersos en la era del Big Data y la generación de datos, ya sean estructurados, como no estructurados. Estos sistemas son algo mucho más complejo que sólo grandes cantidades de datos, ya que además del volumen, pueden tener otras características como la velocidad con que se generan, su variedad, el tener que garantizar la veracidad y el valor que aportan al negocio. Todos estos atributos hacen del Big Data un problema complejo y difícil de tratar. Este artículo expone qué es el Big Data, sus características, los orígenes de esta tecnología y cómo funcionan las técnicas para conseguir extraer valor de los datos.

PALABRAS CLAVES •
big data, datos, hadoop, hdfs

CÓMO CITAR ESTE ARTÍCULO •
Puertas, Enrique. 2020. "Big Data: La revolución de los datos" en :: UEM STEAM Essentials
Enlace web UEM :: http://projectbasedschool.universidadeuropea.es/escuela/escuela/steam_essentials

INTRODUCCIÓN

Cuenta una fábula de Esopo escrita en el siglo VI a.c., como un labrador, a punto de morir, reúne a sus siete hijos, les enseña un haz con siete varas de leña atadas fuertemente con un cordel, y les dice: hijos, dejaré toda mi fortuna a aquel de vosotros que sea capaz de romper este haz de varas de madera. Uno tras otro los hermanos intentan romper el fajo con todas sus fuerzas, pero ninguno lo consigue. Entonces, el padre agarra el fajo, desata el cordel, les da una de las varas a cada uno de sus hijos y les pide que las rompan, cosa que hacen todos sin problema.

Aunque esta fábula tenía su propia moraleja, nosotros lo vamos a utilizar para ejemplificar, de manera muy sencilla, cómo funcionan los sistemas Big Data. Supongamos que ese haz de varas de madera es un conjunto de datos muy grande y que cada uno de los hijos del labrador es un ordenador. Ninguno de los ordenadores, de forma individual, tiene la fuerza (capacidad de cómputo y memoria) necesaria para poder romper (procesar) los datos. Sin embargo, si creamos un sistema en el que los siete hermanos trabajan de forma conjunta, con los datos distribuidos (cada hermano tiene una sola vara) vemos que la tarea "romper varas de madera" sí es abordable, y el resultado final es justo el que se perseguía (el haz de varas de madera acaba roto). Eso es justo lo que se consigue con los sistemas Big Data: resolver problemas que de otra forma son inabarcables por el tamaño y complejidad de los datos.

Vivimos inmersos en la era del Big Data y la generación de datos. En un solo minuto de tiempo, se hacen 4,5 millones de búsquedas en Google, se ven casi 5 millones de videos en Youtube, se envían medio millón de tweets y se publican 55 mil fotos en Instagram (DOMO, n.d.); y eso es solo la punta del iceberg. A pesar de la enorme cantidad de datos que generan las Redes Sociales, no son ni de lejos la industria que más datos genera. Para hacernos una idea, en el colisionador de hadrones del European Organization for Nuclear Research (CERN), cada colisión genera un Petabyte[1] de información por segundo (el equivalente al contenido en texto de todo internet) (Abelev et al., 2014), aunque

solo se registra una pequeña cantidad de esos datos porque hoy en día no contamos con sistemas almacenamiento de información lo suficientemente grandes y rápidos como para poder guardarlo todo (de momento).

QUÉ ES EL BIG DATA

Existen multitudes de definiciones de lo qué es el Big Data. Veamos tres que quizás son las más habituales en la literatura:

"Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers".

> **Forrester** (THE PRAGMATIC DEFINITION OF BIG DATA, N.D.)

"Big Data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making".

> **Gartner** (BIG DATA, N.D.)

"Big Data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it".

> **O'Reilly** (WHAT IS BIG DATA? – O'REILLY, N.D.)

figura 01 » Las 5 Vs del Big Data (Fuente: propia)

LAS 5 Vs DEL BIG DATA

Los sistemas Big Data son algo mucho más complejo que procesar grandes cantidades de datos, ya que tiene otras características que hacen que se enfrenten a múltiples desafíos. Esas características son las conocidas como 5 Vs del Big Data: Volumen, Velocidad, Variedad, Veracidad y Valor (Figura 1). Estos 5 atributos provocan que sea una tarea compleja el extraer datos reales y de calidad, de conjuntos de datos masivos, cambiantes y complicados.

a » **Volumen:** El volumen hace referencia a la cantidad de datos que se generan en nuestro entorno. Es la característica que la gente suele asocial al Big Data, ya que hace referencia a la cantidad masiva de datos que son almacenados con el objetivo de ser procesados, transformando los datos en acciones. Muchas empresas se encuentran inmersas en un proceso de transformación digital, por lo que la cantidad de datos que generan es muy grande. Por ejemplo, una empresa del sector retail que vende sus productos a través de un canal online, necesita implantar tecnologías Big Data para procesar toda la información recogida en su página web, rastreando todas las acciones que lleva a cabo el cliente; conocer en qué enlaces y productos hace click más veces, cuántas veces pasa por el carrito de la compra, cuáles son los productos más vistos, las páginas más vistas, etc.

b » **Velocidad:** La velocidad hace referencia a los datos en movimiento por las constantes operaciones que realizamos, es decir, a la rapidez con la que los datos son creados, almacenados y procesados en tiempo real. En aquellos procesos en los que el tiempo es un factor clave, como por ejemplo la detección de fraude en una transacción bancaria, este tipo de datos debe analizarse en tiempo real para que resulten de utilidad para el negocio y se puedan tomar decisiones que aporten valor.

c » **Variedad:** La variedad se refiere a los distintos tipos y fuentes de datos de un sistema. En una empresa es muy habitual encontrar que trabajan con bases de datos relacionales, archivos Word y Excel, sistemas de información tipo CRM, etc. Los datos pueden ser estructurados y fáciles de manejar, como por ejemplo las bases de datos, o los datos no estructurados, entre los que se incluyen audios, vídeos, imágenes o documentos de texto. Este último tipo datos no estructurados requieren de herramientas específicas, debido a que el tratamiento de la información es muy diferente con respecto a los datos estructurados. Por este motivo, las empresas necesitan integrar, observar y procesar los datos que son recogidos y procesarlos para lidiar con esa heterogeneidad.

d » **Veracidad:** Cuando hablamos de veracidad hablamos de al grado de fiabilidad e incertidumbre de la información. Es necesario dedicar tiempo para conseguir que los datos

sean de calidad, aplicando procesos que garanticen que los datos son actuales, no están repetidos, y que tengan coherencia y consistencia en el tiempo. Para esta tarea es clave la figura del responsable de los datos de una empresa, que es quien debe velar por que se mantenga esta calidad de los datos en sus sistemas de información.

e » **Valor:** El dato, por sí solo, no tiene valor. Tampoco aporta valor el hecho de recopilar y almacenar gran cantidad de información. El valor se obtiene de los datos que se transforman en información y ésta a su vez se convierte en conocimiento que permite tomar decisiones y realizar acciones. El valor de los sistemas Big Data está en que los responsables del negocio puedan tomar decisiones (las mejores decisiones) en base a los datos.

LOS ORÍGENES DEL BIG DATA

El término Big Data fue usado por primera vez en el año 1997 por *Cox* y *Ellsworth* en la publicación: "Application controlled demand paging for out-of-core visualization" (Cox & Ellsworth, 1997) en la que indicaban que "el ritmo al que crecen los datos empieza a ser un problema para los sistemas informáticos actuales". Esto es lo que denominaron el "problema del Big Data". Y es que en un estudio realizado en esos años (Coffman & Odlyzko, 1998) se constató que en 1998 el tráfico en internet estaba creciendo a un ritmo en el que se duplicaba cada año.

En 1998, cuando Google comenzó a ofrecer un servicio de búsqueda en Internet, recibía sólo 10.000 consultas de búsqueda por día. En el año 2004, cuando Google empezó a cotizar en los mercados, ya estaba recibiendo 200 millones de consultas diarias. Para el año 2006, los usuarios de Google estaban enviando unas 10.000 consultas por segundo a este popular motor de búsqueda. En ese momento, mil ordenadores eran capaces de procesar una búsqueda en sólo 0,2 segundos. Pero dado la tasa de crecimiento que se estaba dando en internet, era obvio que Google tenía un problema, y que a ese ritmo pronto no podría dar respuesta a todas las búsquedas de sus usuarios en un tiempo razonable. El problema además era doble; por un lado, estaba el almacenamiento de la información de las páginas web, para lo que necesitaba servidores con una capacidad de almacenamiento enorme, es decir, necesitabas máquinas con discos duros cada vez más grandes. Por otro, esa información se guardaba en bases de datos que tardaban mucho en recuperar y procesar la información. Según crecía la complejidad y el tamaño de los datos almacenados, las bases de datos tradicionales requerían una elevada inversión en servidores más potentes que podían llegar a costar varios cientos de miles de euros cada unidad. Cuando el tamaño de los datos crecía mucho era necesario invertir en poner más memoria RAM y procesadores más potentes a los servidores, lo que se conoce como "escalado vertical". Ese escalado era muy costoso ya que el mercado de servidores empresariales estaba copado por unas pocas marcas (IBM, HP, Dell, etc.) que vendían los componentes para sus máquinas a unos precios muy altos, muy alejados de los precios de los componentes de ordenadores personales de aquella época, aprovechando el uso de hardware propietario que creaba una dependencia de la marca. Si usabas servidores de la compañía determinada, por lo general sólo podías actualizarlos con memoria, discos y procesador proporcionados por esa misma compañía.

A todo este problema de sostenibilidad de las bases de datos existentes había que sumar el que éstas sólo podían guardar información estructurada, es decir datos en forma de tabla. Por lo que no eran adecuadas para almacenar y procesar información no estructurada como videos, imágenes, textos en lenguaje natural, etc.

Y entonces llegaron las dos publicaciones que lo cambiaron todo. En 2004 Google publicó un artículo científico titulado "*MapReduce: Simplified Data Processing on Large Clusters*" (Dean & Ghemawat, 2008) y en 2006 publicó "*Bigtable: A Distributed Storage System for Structured Data*" (Chang et al., 2006), basado en otro publicado en 2003 titulado "*The Google File System*" (Sanjay Ghemawat and Shun-Tak Leung, 2003).

Estos artículos establecieron las bases de las técnicas de Big Data que usamos hoy en día. En ellos se describían un nuevo paradigma de programación llamado **MapReduce**, que permite el procesamiento de enormes cantidades de datos.

El segundo artículo establecía que es posible construir un sistema de almacenamiento distribuido para guardar datos estructurados usando ordenadores convencionales, lo que se conocía como **"commodity servers"**, y que resultaban muchísimo más baratos que los servidores propietarios (hasta 100 veces más baratos en algunos casos). Además, este nuevo sistema permitía incrementar las capacidades almacenamiento simplemente añadiendo más ordenadores baratos, lo que se conoce como escalado horizontal.

LA APARICIÓN DE HADOOP

Doug Cutting y Mike Caffarella son dos ingenieros que en el año 2002 empezaron a trabajar en un proyecto llamado "Nutch", que tenía como objetivo construir un buscador de páginas web libre y de código abierto (khare et al., 2005). En 2004, tras la publicación de los artículos de Google sobre GFS y MapReduce, añadieron a Nutch un sistema de archivos distribuido similar al descrito por Google, y usaron MapReduce para implementar las búsquedas, lo que mejoró el rendimiento de forma muy notable. A estas

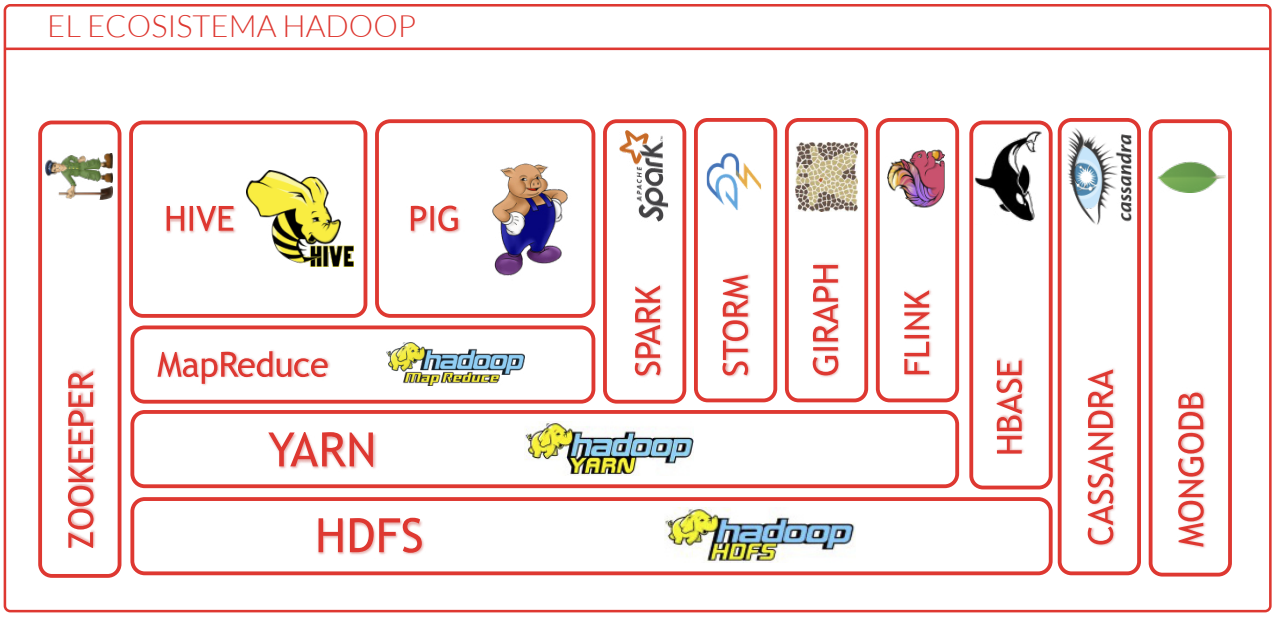

figura 02 » Ecosistema Hadoop (Fuente: propia)

nuevas características las llamaron **"HADOOP"**, que era el nombre del elefante de juguete del hijo de Cutting.

En 2006 Doug Cutting empieza a trabajar en un equipo de trabajo de Yahoo! que investigaba cómo construir nuevas arquitecturas informáticas para mejorar el rendimiento de las búsquedas en internet. Este nuevo equipo sacó Hadoop del proyecto Nutch, le dio entidad propia, y puso al frente de su desarrollo a Doug Cutting. De ahí surgió el framework HADOOP, basado en lo que llamaron "Hadoop Distributed Filesystem (HDFS)" para el almacenamiento de datos, y el paradigma de programación MapReduce como modelo de computación (BORTHAKUR, 2007). Este framework se creó como un proyecto Open-Source, libre y abierto para que pudiera utilizarlo cualquier persona.

Desde ese momento la popularidad de Hadoop creció exponencialmente. Fue adoptado por multitud de empresas para empezar a almacenar y procesar datos que hasta ese momento se descartaban. Por ejemplo, la compañía de tarjetas de crédito Visa fue capaz de reducir el tiempo que tardaba en procesar, los registros de 2 años, esto es 73.000 millones de transacciones, de un mes a 13 minutos escasos (MAYER-SCHÖNBERGER, VIKTOR & CUKIER, 2013), lo que les permitió, por ejemplo, detectar operaciones fraudulentas por valor de varios miles de millones de dólares.

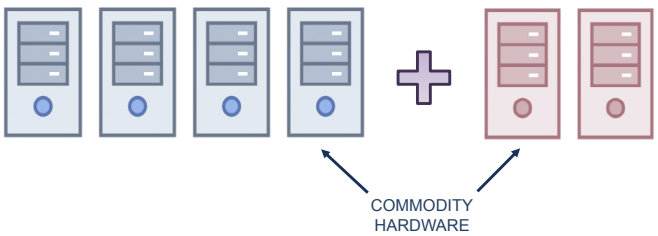

figura 03 » Escalado Horizontal (Fuente: propia)

Además, el hecho de ser un proyecto abierto permitió que surgieran muy rápidamente nuevas herramientas que mejoraban y complementaban las funcionalidades de Hadoop (Figura 2): bases de datos avanzadas como Cassandra, HBase o MongoDB, gestores de recursos como "Yarn", frameworks de programación como "Spark" o "Storm", que mejoran el rendimiento y las capacidades de MapReduce, etc.

HADOOP DISTRIBUTED FILESYTEM (HDFS)

Como se puede ver en la figura 2, el sistema de archivos distribuido HDFS es la base del resto de herramientas que conforman el ecosistema Hadoop. HDFS trocea los archivos y distribuye los trozos (conocidos como "chunks") para almacenarlos en varios ordenadores interconectados entre sí, lo que se conoce como un **cluster.** Aunque los datos están repartidos en varias máquinas, HDFS se encarga de gestionar la complejidad de trocear, repartir, balancear y mantener la coherencia de los datos distribuidos, todo de una forma transparente para el usuario del sistema de archivos, que accede a los datos para leerlos o escribirlos como si estuvieran alojados en una única máquina. La forma en que HDFS estructura y almacena la información proporciona dos capacidades básicas para procesar grandes cantidades de datos: escalabilidad y tolerancia fallos. La escalabilidad se consigue mediante un sistema que permite añadir nuevos ordenadores a nuestro cluster y automáticamente las capacidades de almacenamiento de los nuevos equipos se añaden al sistema (Figura 3).

Además, HDFS está diseñado para recuperarse ante fallos en los ordenadores de un cluster, usando redundancia de bloques. Cada bloque de datos en los que se divide un

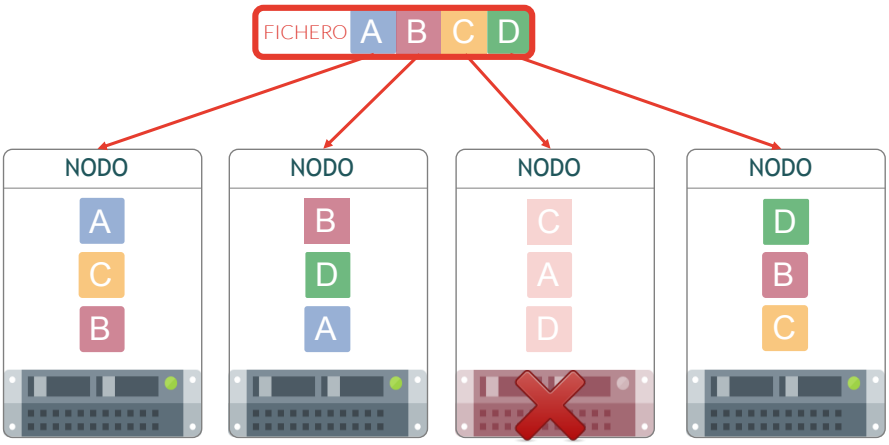

figura 04 » Redundancia de bloques en HDFS (Fuente: propia)

archivo, se copia varias máquinas distintas (por defecto 3), de forma que, si una máquina falla, los datos no se pierden porque se pueden recuperar usando las copias redundantes.

En la figura 4 podemos observar cómo, aunque el nodo 3 falle, el fichero se puede seguir leyendo y recuperando ya que existen copias de los bloques A C y D en otros nodos.

MAPREDUCE

En MapReduce se procesan los datos principalmente en dos fases, conocidas como la fase **"map"** y la fase **"reduce"**. En la fase map se leen los datos de entrada y se crea un listado de pares (clave, valor). Para cada valor de entrada k, se le aplica una función que devuelve un valor v y se añade a lista de salida de la fase map la tupla (k, v). En la fase reduce, se mezclan y juntan todos los pares (k, v) que tienen una clave k común, y se realiza una función de agregación sobre los valores. Por ejemplo, supongamos que queremos contar el número de apariciones de cada una de las palabras que aparecen en el texto de "El Quijote" (Cervantes Saavedra. 1547-1616, 1966). Con MapReduce, para cada palabra del texto le asignamos un valor "1", que es el valor de "aparición" inicial de cada palabra, y devolvemos la tupla (palabra, 1): ["(en,1); (un,1); (lugar,1); (de,1); (la,1); (Mancha,1)", ...] A continuación, se juntan todas las tuplas que tienen como clave la misma palabra y, por último, se realizaría la fase reduce aplicando la función de agregación ("sumar") que devolvería los resultados finales. La figura 5 muestra un ejemplo simplificado de este contador de palabras.

La ventaja con respecto al procesamiento tradicional es que tanto la fase map como la fase reduce se hacen de forma distribuida, repartiendo la carga de trabajo entre las distintas máquinas que tengamos disponible, reduciendo el tiempo de proceso de forma muy significativa.

Notas
01 » 1 Petabyte = 1015 bytes (1.000.000.000.000.000 de bytes)

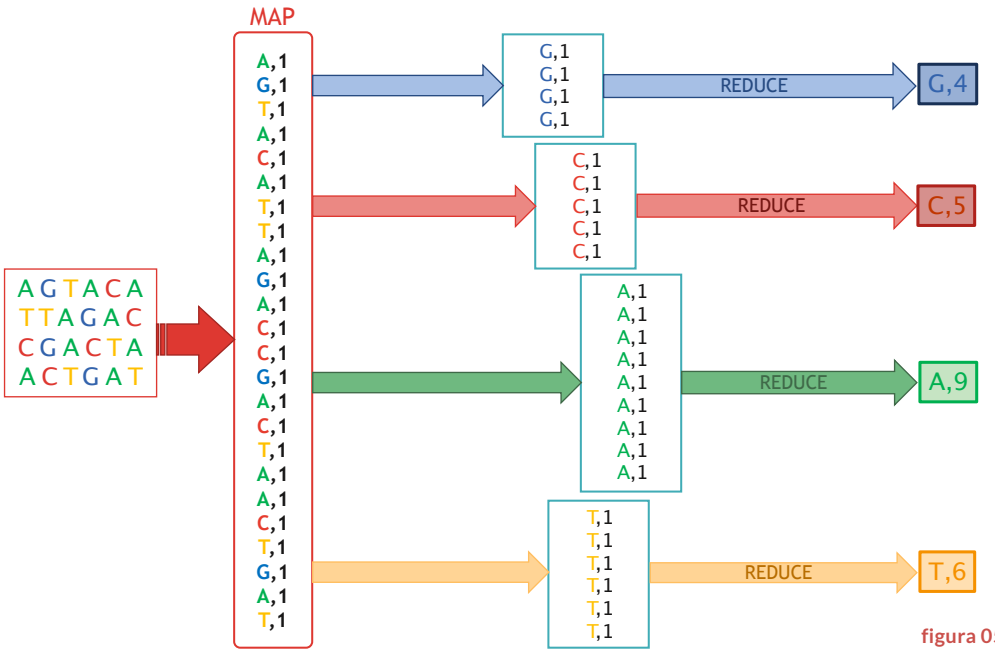

figura 05 » Contador de palabras con MapReduce (Fuente: propia)

CONCLUSIÓN

Aunque estamos viviendo una época de generación de información, en el que se producen petabytes de datos cada día, esa enorme cantidad de datos no es relevante. Lo que importa es lo que las organizaciones pueden hacer con los datos para obtener ideas que lleven a la toma de mejores decisiones y movimientos de negocio adecuados. Para una toma de decisiones correcta, la información es fundamental, y mucho más cuando podemos manejar toda la información que se genera cada día. Con las técnicas de Big Data podremos llevar a cabo planes de acción inteligentes y veloces que ayuden a favorecer el nuestro negocio usando grandes cantidades de datos que hace años era imposible tratar por su volumen y complejidad. Así por ejemplo, ahora es posible analizar los resultados de una campaña de marketing a los pocos segundos de haberse lanzado un anuncio, se pueden crear nuevos fármacos analizando y combinando millones de estructuras moleculares de forma automática, una tarea que habitualmente puede llevar años con las tecnologías tradicionales. En banca se está usando el Big Data para la detección de fraudes y el control de riesgos, o para mejorar la experiencia del cliente analizando compras y comportamientos. En salud se pueden analizar millones de imágenes médicas para la detección automática de patologías o para optimizar los recursos de hospitales y centros de Salud.

Además, el uso de Big data supone un considerable ahorro de costes en servidores informáticos, ya que se basa en el uso de ordenadores convencionales baratos (comparados con los servidores propietarios de grandes marcas) que se pueden añadir o quitar dinámicamente para ajustar las capacidades de almacenamiento y proceso de nuestro sistema.

REFERENCIAS BIBLIOGRÁFICAS

» Abelev, B., Abramyan, A., Adam, J., Adamová, D., Aggarwal, M. M., Agnello, M., Agostinelli, A., Agrawal, N., Ahammed, Z., Ahmad, N., Ahmad Masoodi, A., Ahmed, I., Ahn, S. U., Ahn, S. A., Aimo, I., Aiola, S., Ajaz, M., Akindinov, A., Aleksandrov, D., … Zyzak, M. (2014). Performance of the ALICE experiment at the CERN LHC. In ***International Journal of Modern Physics A*** (Vol. 29, Issue 24). World Scientific Publishing Co. Pte Ltd. https://doi.org/10.1142/S0217751X14300440

» ***Big Data.*** (n.d.). Retrieved June 18, 2020, from https://www.gartner.com/en/information-technology/glossary/big-data

» Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. ***Hadoop Project Website***. https://doi.org/10.1109/MSST.2010.5496972

» Cervantes Saavedra 1547-1616, M. de. (1966). ***El ingenioso hidalgo Don Quijote de La Mancha***. Madrid : Espasa-Calpe, 1966. https://search.library.wisc.edu/catalog/999703820802121

» Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., & Gruber, R. E. (2006). BigTable: A distributed storage system for structured data. ***OSDI 2006 - 7th USENIX Symposium on Operating Systems Design and Implementation***.

» Coffman, K. G., & Odlyzko, A. M. (1998). ***The size and growth rate of the Internet***.

» Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. ***Proceedings of the IEEE Visualization Conference***, 235–244. https://doi.org/10.1109/visual.1997.663888

» Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. ***Communications of the ACM***. https://doi.org/10.1145/1327452.1327492

» DOMO. (n.d.). ***Data Never Sleeps*** 7.0. Retrieved June 18, 2020, from https://www.domo.com/learn/data-never-sleeps-7

» Khare, R., Cutting, D., Sitaker, K., & Rifkin, A. (2005). Nutch : A Flexible and Scalable Open-Source Web Search Engine. ***14th International Conference on World Wide Web (WWW 2005)***. https://doi.org/10.1101/gad.11.7.926

» Mayer-Schönberger, Viktor & Cukier, K. (2013). Big Data. La revolución de los datos masivos. In ***Latin Trade***.

» Sanjay Ghemawat and Shun-Tak Leung, H. G. (2003). The google filesystem. ***Proceedings of the 19th ACM Symposium on Operating Systems Principles.***

» ***The Pragmatic Definition Of Big Data.*** (n.d.). Retrieved June 18, 2020, from https://go.forrester.com/blogs/12-12-05-the_pragmatic_definition_of_big_data/

» ***What is big data?*** – O'Reilly. (n.d.). Retrieved June 18, 2020, from https://www.oreilly.com/radar/what-is-big-data/

BIOGRAFÍA

Enrique Puertas es Ingeniero Informático y Doctor en Tecnologías de la Información Aplicadas. Especialista en Inteligencia Artificial y Big Data, desde el año 2016 dirige el Máster Universitario en Big Data Analytics de la Universidad Europea de Madrid, labor que compagina con la docencia como profesor de grado en las asignaturas de Inteligencia Artificial y Grandes Volúmenes de Datos. Miembro del grupo de investigación de "Sistemas Inteligentes", y co-director grupo "Machine Learning Salud-UE", centrado en el desarrollo y aplicación de herramientas de Inteligencia Artificial a la práctica clínica y de gestión sanitaria. Es además autor del „Manual práctico de Inteligencia Artificial en entornos sanitarios" (Elsevier 2020, ISBN: 9788491138013). Como investigador ha participado en más de una veintena proyectos y contratos de investigación, nacionales e internacionales, y es autor de más de treinta publicaciones en revistas y congresos internacionales en temas relacionados con Big Data, Minería de Datos, e Inteligencia Artificial.