Ciencias de la Información ISSN: 1606-4925 Vol. 49, No.2 , mayo-agosto, pp. 37 - 45, 2018

Una revisión de la tecnología «BigData» para grandes volúmenes de datos

A review of «BigData» technology for large volumes of data

Ireimis Leguen de Varona
Yoan Martínez López
Julio Madera

Resumen

Big Data es una tecnología que se ocupa del manejo de grandes volúmenes de datos, usando procedimientos para identificar patrones recurrentes dentro de esos datos. Debido al aumento de las investigaciones en esta área, se trazó el objetivo de este trabajo, el cual consiste en presentar una revisión bibliográfica del estado del arte de la tecnología Big Data, en la que se tiene en cuenta la definición, arquitectura, tipos de datos, así como los diferentes sistemas para el trabajo con esta tecnología. Se muestran distintas aplicaciones de esta tecnología en los diferentes campos de la ciencia y los tipos de roles de trabajo con Big Data. Asimismo, se tratan los problemas y retos a los que se enfrenta, demostrando la importancia de su uso y cuáles son sus proyecciones para el futuro.

Palabras clave: *arquitectura; Big Data; Hadoop; sistemas distribuidos; volúmenes de datos*

Abstract

Big Data is a technology that deals with the handling of large volumes of data, using procedures to identify recurring patterns within that data. Due to the increase in research in this area, the objective of this work was outlined, which consists of presenting a bibliographic review of the state of the art of Big Data technology, which takes into account the definition, architecture, types of data, as well as the different systems for working with this technology. Different applications of this technology are shown in the different fields of science and the types of work roles with Big Data. It also addresses the problems and challenges it faces, demonstrating the importance of its use and what its projections are for the future.

Keywords*: architecture; Big Data; Hadoop; distributed systems; data volumes*

Introducción

Los seres humanos están creando y almacenando información constantemente y en grandes cantidades, girando todo el mundo en torno a los datos, por ejemplo, en torno a la ciencia (bases de datos de astronomía, genómica, datos medio-ambientales), ciencias sociales y humanísticas (libros escaneados, documentos históricos, datos sociales), negocios y comercio (ventas de corporaciones, transacciones de mercados, censos), medicina (datos de pacientes, datos de escáner, radiografías), industria (industria automovilística, textil). Muchas son las fuentes de donde provienen estos grandes volúmenes de datos, como son las compañías que mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. En muchos países se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, etc., y si a todo esto se añaden transacciones financieras realizadas en línea o por dispositivos móviles, ubicación geográfica mediante coordenadas GPS, análisis de redes sociales (Facebook, Twitter, Flickrs, Youtube) y de multimedias, estos datos crecen enormemente. Pero la gestión de los mismos genera grandes dificultades en su manejo; las mismas se centran en la captura, almacenamiento, búsqueda, compartición, análisis y visualización (Schönberger, 2013). Por lo que se hace necesario incorporar la tecnología Big Data para el tratamiento y manejo de grandes volúmenes de datos.

Se puede definir Big Data como una colección de datos grandes, complejos, muy difíciles de procesar a través de herramientas de gestión y procesamiento de datos tradicionales. Son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y

conocimiento oculto en ellos (Pentland, 2012). Para describir mejor lo que representa Big Data, frecuentemente se habla de las cinco Vs. La *International Business Machines Corporation* (IBM, por sus siglas en inglés) fue la que empezó definiendo tres Vs y luego se han añadido las otras, dependiendo de la fuente, que definen perfectamente los objetivos que este tipo de sistemas buscan conseguir (Zikopoulos, Parasuraman,Deutsch, Giles, & Corrigan, 2012): 1) Volumen: un sistema Big Data es capaz de almacenar una gran cantidad de datos mediante infraestructuras escalables y distribuidas. En los sistemas de almacenamiento actuales empiezan a aparecer problemas de rendimiento al tener cantidades de datos; 2) Velocidad: una de las características más importantes es el tiempo de procesado y respuesta sobre esos grandes volúmenes de datos, obteniendo resultados en tiempo real y procesándolos en un intervalo de tiempo reducidos. Las fuentes de datos pueden llegar a generar mucha información cada segundo, obligando a tener que almacenar dicha información de manera veloz; 3) Variedad: las nuevas fuentes de distintos tipos y formatos de información a los ya conocidos hasta el momento como datos no estructurados, que pueden ser representados de diversos formatos y estructuras: como texto, numéricos, imágenes, audio, video, secuencias, series temporales; obtenidos de diferentes fuentes como: dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, y que la tecnología Big Data es capaz de almacenar y procesar sin tener que realizar un pre-procesamiento para estructurar o indexar la información; 4) Variabilidad: las tecnologías que componen una arquitectura Big Data deben ser flexibles a la hora de adaptarse a nuevos cambios en el formato de los datos tanto en la obtención como en el almacenamiento y procesado. Se podría decir que la evolución es una constante en la tecnología de manera que los nuevos sistemas deben estar preparados para admitirlos; 5) Valor: el objetivo final es generar valor de toda la información almacenada a través de distintos procesos de manera eficiente y con el coste más bajo posible.

La tecnología Big Data representa una oportunidad para tomar decisiones basadas en el uso intensivo de los datos y constituye un reto para manejar inconsistencias, datos incompletos, escalabilidad, corriente continua de datos, problemas de seguridad (Chen, 2014). Asimismo, la competitividad en la productividad de los negocios y las tecnologķas conllevan al uso de la tecnología Big Data (Kouzes, 2009). Big Data permite romper con el enfoque relacional de las bases de datos (NoSQL) debido a la naturaleza y dimensiones de los datos, lo cual no significa no usar SQL, sino es romper con los enfoques rígidos normalizados. Los enfoques NoSQL son útiles cuando trabajan con grandes cantidades de datos, y la naturaleza de los datos no requiere el modelo relacional para su estructura, tampoco necesitan esquemas. Entre los enfoques NoSQL más usados se encuentran Apache Cassandra (Hewitt, 2010), SimpleDB (Murty, 2009), Google BigTable, Apache Hadoop (White, 2009), MapReduce (White, 2009). También Big Data permite extraer conocimiento desde fuentes distribuidas y puede generar modelos de decisión dependientes de ese contexto; para ello es necesario usar mecanismos de intercambio y fusión de información que garantice que trabajando con información distribuida se encuentre un modelo global. Ademįs, obliga a romper con los conceptos clásicos de seguridad de los datos (Chen, 2014 & Kouzes, 2009). Esta tecnología tiene aplicaciones en muchos sectores de la sociedad, como es el caso del empresarial, el hotelero, el financiero, la educación, la medicina, la alimentación, el marketing, seguridad de redes, redes sociales y multimedias, redes de sensores, dispositivos móviles e instrumentos científicos, entre otros.

Material y métodos

Se realizó una investigación de tipo descriptiva con análisis documental, como método usado para la recopilación bibliográfica y la revisión de la literatura. El método histórico-lógico posibilitó analizar diferentes criterios y profundizar en la evolución y el desarrollo de la metodología Big Data.

Las bases de datos utilizadas fueron:
Scielo: http://www.scielo.org/php/index.php
Science Direct: https://www.sciencedirect.com/
Elsevier: https://www.elsevier.com/about/open-science
Springer: http://www.springer.com/gp/
IEEE: https://www.ieee.org/index.html
ACM: https://www.acm.org/

La búsqueda se limitó a los artículos originales publicados entre 2005-2015, escritos en español o inglés cuyos registros presentaran los resúmenes. En la base de datos Google Scholar en español que no tiene descriptores se usaron las siguientes palabras clave: «Arquitectura», «Big Data» «Hadoop», «sistemas distribuidos», «volúmenes de datos». En esta estrategia de búsqueda sólo se utilizó la combinación de las palabras clave con el operador booleano «OR» para recuperar toda la literatura existente en español sobre el tema.

Una vez identificados todos los artículos se seleccionaron los que estudiaban la salud laboral. Con respecto a la estrategia de búsqueda en la base de datos Google Scholar en inglés se emplearon los términos del Big Data Headings (BDH).

Se utilizaron los siguientes bloques:

a) Descriptores de Big Data: Big Data and Hadoop and Volume and Datos, MapReduce and *Arquitectura*, *Architecture*.

b) Descriptores sobre estado de Big Data o Grandes Volumen de Datos: Big Data; Hadoop; sistemas distribuidos; volúmenes de datos.

c) Descriptores para la localización de lugares e instituciones de Cuba, Estados Unidos, América Latina y Europa. La estrategia de búsqueda realizada utilizó la siguiente combinación (Bloque a) AND (Bloque b) AND (Bloque c).

Resultados y discusión

Según el sitio Google Scholar (scholar.google.com) a partir del año 2005 hasta principios del año 2015, van en aumento las publicaciones relacionadas con la tecnología Big Data, como se muestra en la figura 1.

Figura 1. *Cantidad de publicaciones de Big Data desde el año 2005 hasta principios del 2015*
Fuente*: http://scholar.google.com*

Un total de 133000 publicaciones fueron encontradas en la búsqueda en este sitio hasta principios del año 2015.

Diferentes autores han publicado importantes artículos sobre esta temática, entre los que se encuentran (Kouzes, 2009), (Lam, 2010), (Shvachko, Kuang, Radia, & Chansler, 2010), (Abuín, 2010), (Manyika, 2011), (Rogers,2011), (Mcafee, 2012), (Pentland, 2012), (Zikopoulos, Parasuraman,Deutsch, Giles, & Corrigan, 2012), (Barranco, 2012), (Schönberger, 2013), (Serrat, 2013), (Russom, 2013), (Liu, 2013),(Haines, 2013), (Moreno,2013), (Sabater, 2013), (Fernández, 2013), (Chambi, 2013) (Chen, 2014 a, Chen 2014 b), (Casado, 2014), (Manaure, 2014), (Rooney, 2014), (Kosner, 2012). Estos autores plantean que, lo que hace que Big Data sea tan útil para muchas empresas es el hecho de que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabían que tenían. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.

Big Data para el trabajo con grandes volúmenes de datos. Arquitectura de Big Data

La arquitectura Big Data está compuesta generalmente por cinco capas: recolección de datos, almacenamiento, procesamiento, visualización y administración. Esta arquitectura no es nueva, sino que ya es algo generalizado en las soluciones de Inteligencia de Negocios que existen hoy en día. Sin embargo, debido a las nuevas necesidades cada uno de estos pasos ha ido adaptándose y aportando nuevas tecnologías a la vez que abriendo nuevas oportunidades.

En la figura 2 se muestra el flujo que la información tendría en una arquitectura Big Data, con orígenes de datos diversos, bases de datos o documentos que se reciben y almacenan a través de la capa de recolección de datos, con herramientas específicamente desarrolladas para tal función. Los datos recibidos pueden procesarse, analizarse y/o visualizarse tantas veces como haga falta y lo requiera el caso de uso específico (Serrat, 2013).

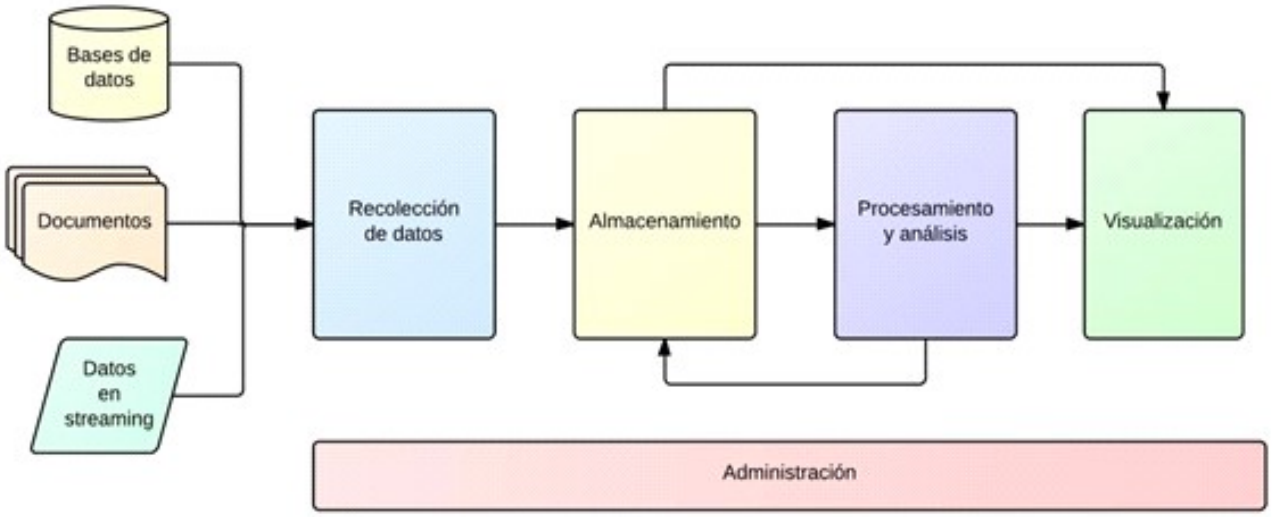

Figura 2 *Arquitectura Big Data por capas.*
Fuente: *Serrat, (2013)*

Capa Recolección de datos

En esta capa es donde se lleva a cabo la obtención de datos legales (cumplen la ley de protección de datos), considerados que tras un procesamiento adecuado pueden aportar información de valor para la empresa. Además, se envían los datos a la etapa de almacenamiento, donde se guardarán todos los datos recolectados para su futuro procesamiento. En esta etapa, los datos pueden sufrir algún tipo de proceso o cambio si la aplicación así lo requiere, por ejemplo, el filtrado de información no deseada o el formato con el que se guardará finalmente en el sistema de almacenamiento.

Capa de Almacenamiento

En esta capa se encuentran todas las herramientas que permiten almacenar información de gran volumen y de formato variable. No sólo se almacenan los datos recolectados en la etapa de recolección, sino que se suelen guardar los resultados obtenidos al procesar un conjunto de datos en la etapa de procesamiento. La capa de almacenamiento tiene, a grandes rasgos, dos elementos básicos: el sistema de ficheros y la base de datos. Hasta hace poco los sistemas de tratamiento de la información se centraban principalmente en las bases de datos, pero, debido a que en los sistemas Big Data se busca la mayor variedad posible las bases de datos acostumbran a ser poco flexibles, los sistemas de ficheros han cobrado mayor importancia.

Procesamiento y análisis

En la capa de procesamiento es donde se llevan a cabo todos los análisis, y procesamientos previos de los datos para el futuro análisis. Entre la capa de almacenamiento y la capa de procesamiento los datos fluyen en ambas direcciones ya que o bien se obtiene un conjunto de los datos almacenados para poder procesarlos y analizarlos, o bien se ha hecho ya el procesamiento de los datos y es necesario almacenarlos para poder visualizarlos más adelante o hacer otros procesamientos más complejos que requerían previamente preparar los datos.

Visualización

La capa de visualización es la etapa en la que se muestran los resultados de los análisis que se han realizado sobre los datos almacenados. La visualización de los resultados es normalmente bastante gráfica, de manera que se permita una adquisición rápida de conclusiones para poder decidir cuanto antes cómo actuar o qué estrategias se van a seguir, con el objetivo de poder ganar la máxima ventaja o evitar un problema mayor.

Tipos de datos

En la actualidad existe una amplia variedad de tipos de datos a analizar, una buena clasificación ayudaría a entender mejor su representación, aunque es muy probable que estas categorías puedan extenderse con el avance tecnológico. La figura 3, tomado de Barranco, (2012) muestra los tipos de datos de Big Data.

Estos se clasifican en:

Web and Social Media: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc.

Machine-to-Machine (M2M): Se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas, etc.) los cuales transmiten a través de redes inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

Big Transaction Data: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas, etc. Estos datos transaccionales están disponibles en formatos tanto semi-estructurados como no estructurados.

Biometrics: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc.

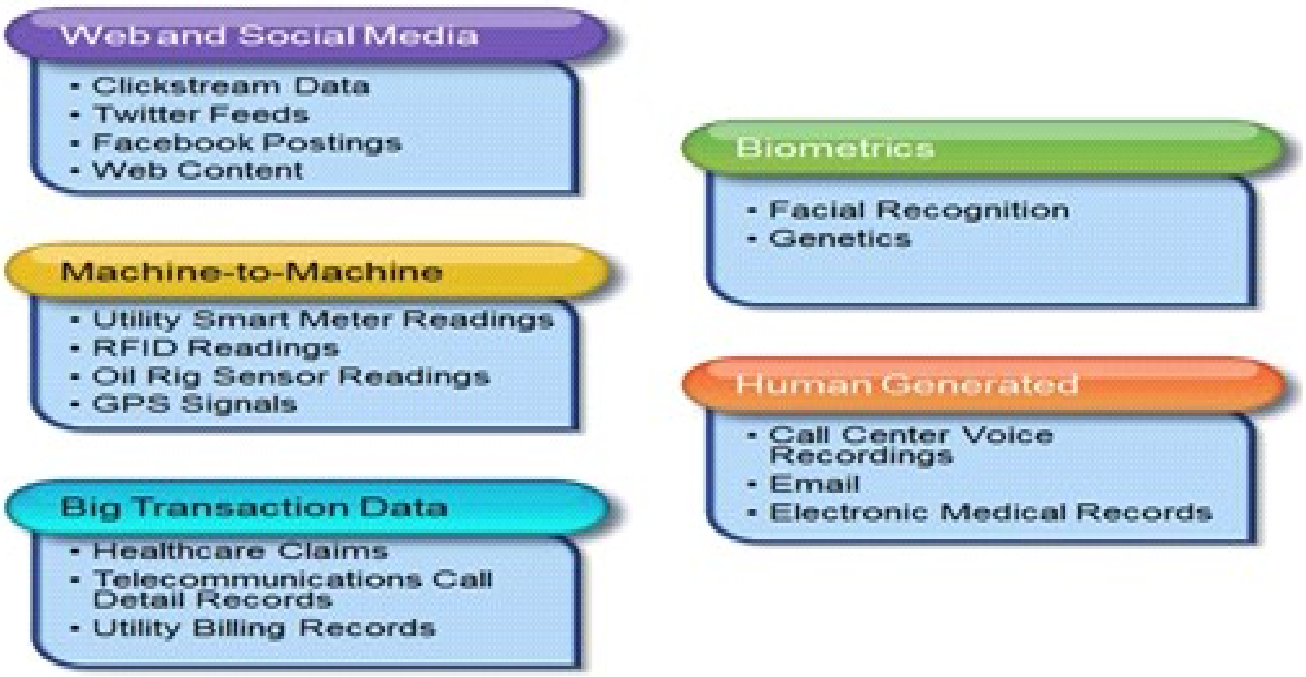

Figura 3. *Tipos de datos de Big Data.*
Fuente: *Barranco, (2012)*

Figura 4. *Cantidad de publicaciones de sistemas de Big Data desde el año 2010 hasta principios del 2015.*
Fuente: *http://scholar.google.com*

En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación. *Human Generated:* Las personas generan diversas cantidades de datos como la información que guarda un centro de llamadas al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

Sistemas distribuidos para el trabajo con Big Data
Sistemas distribuidos

Un sistema distribuido se define como una colección de computadoras separadas físicamente y conectadas entre sí por una red de comunicaciones distribuida; cada máquina posee sus componentes de hardware y software que el usuario percibe como un solo sistema. El usuario accede a los recursos remotos de la misma manera en que accede a recursos locales, o un grupo de computadores que usan un software para conseguir un objetivo en común.

Los sistemas distribuidos deben ser muy confiables, ya que, si un componente del sistema se descompone en otro, debe ser capaz de reemplazarlo. Esto se denomina tolerancia a fallos (Berman, Fox, & Hey, 2003; Foster & Kesselman, 1999). Entre los diferentes sistemas distribuidos se encuentran HBase (George, 2011), que es una herramienta no relacional para el modelado de Google con Big Table (inscripción para Hbase , sistema construido para Google File System), Apache Cassandra (Hewitt, 2010), que es un sistema de software que soporta grandes cantidades de datos; desarrollados por Facebook pero actualmente mantenido por la empresa Apache Software Fundation, MongoDB (Banker, 2011 & Hawkins, 2010), que es un sistema de base de datos NoSQL orientado a documentos, desarrollado bajo el concepto de código abierto, Spark (Zaharia, 2011) (Matei, 2011), es framework (modelo de programación) para clúster de computadoras que trabaja con los paradigmas MapReduce, usualmente se usa para el desarrollo de algoritmos de aprendizaje automático y Hadoop (White, 2009), que es un framework de código abierto para procesar conjuntos de datos de un determinado tipo para problemas concretos.

Una búsqueda en Google scholar muestra las cantidades de publicaciones relacionadas con estos sistemas, hasta principio del año 2015, las cuales son visualizadas en la figura 4.

Figura 5. . Cantidad de publicaciones de sistemas de Hadoop desde el año 2005 hasta principios del 2015.
Fuente: *http://scholar.google.com*

Hadoop

Apache Hadoop es un framework (modelo de programación)de software que soporta aplicaciones distribuidas bajo una licencia libre (Cloud, 2013). Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en los documentos Google para MapReduce y Google File System (GFS). Apache es un proyecto de alto nivel que está siendo construido y usado por una comunidad global de contribuyentes, mediante el lenguaje de programación Java. Yahoo! ha sido el mayor contribuyente al proyecto y lo usa extensivamente en su negocio.

Hadoop se puede ejecutar de tres formas distintas:

• Modo Local / Standalone
Por defecto, Hadoop está configurado para ejecutarse en modo no distribuido como un proceso Java aislado. No hay «daemons Hadoop» en ejecución y todo se ejecuta en una única Máquina Virtual de Java. Esto es útil para la depuración (Cloud, 2013).

• Modo Pseudo-distribuido
Hadoop puede ejecutarse en un modo pseudo-distribuido. Cada «daemon Hadoop» se ejecuta en un proceso Java diferente. Los «daemons Hadoop» se ejecutan en el equipo local, simulando así un clúster o sistema distribuido de pequeña escala. También es apropiado para la ejecución de programa MapReduce (Cloud, 2013).

• Distribuido (Clúster)
En este modo, los «daemons Hadoop» se ejecutan en un clúster de máquinas. Es la forma de aprovechar toda la potencia de Hadoop, se maximiza el paralelismo de procesos y se utilizan todos los recursos disponibles del clúster en el que se va a configurar. Hadoop permite la creación de aplicaciones para procesar grandes volúmenes de información distribuida a través de un modelo de programación sencillo. Está diseñado para ser escalable puesto que trabaja con almacenamiento y procesamiento local (pero distribuido), de manera que funciona tanto para clústeres de un solo nodo como para los que estén formados por miles. Detecta errores a nivel de aplicación, pudiendo gestionar los fallos en los distintos nodos y ofreciendo un buen nivel de tolerancia a errores (Serrat, 2013).

Muchas son las publicaciones relacionadas con Hadoop, según Google scholar, entre los años 2005 y principios del 2015, se publicaron 47600 artículos acerca de este sistema distribuidos, como se muestra en la figura 5.

Componentes de Hadoop

El proyecto Hadoop se ha convertido en la solución Big Data más utilizada por las empresas. Grandes compañías como Oracle, IBM y Microsoft han apostado por Hadoop como la solución general a los problemas empresariales que requieren de tecnología Big Data.

Hadoop es una implementación de código abierto, inspirado en el proyecto de Google File System (GFS) y en el paradigma de programación MapReduce, el cual consiste en dividir en dos tareas (mapper – reducer) para manipular los datos distribuidos a nodos de un clúster logrando un alto paralelismo en el procesamiento (Russom, 2013). Hadoop está compuesto de dos componentes: Hadoop Distributed File System (HDFS) y Hadoop MapReduce. (Lam, 2010; Shvachko, Kuang, Radia, & Chansler, 2010).

HDFS es un sistema de archivos distribuido, escalable y portátil escrito en Java para el framework Hadoop. Los datos en el clúster de Hadoop son divididos en pequeñas piezas llamadas bloques y distribuidas a través del clúster; de esta manera, las funciones map y reduce pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes.

MapReduce es el núcleo de Hadoop. El mismo es un framework utilizado por Google para dar soporte a la computación paralela (es una forma de cómputo en la que muchas instrucciones se ejecutan simultáneamente, operando sobre el principio de que problemas grandes se pueden dividir en unos más pequeños, que luego son resueltos simultáneamente) sobre grandes colecciones de datos en grupos de computadoras (Liu, 2013).

No todos los procesos pueden ser abordados desde el framework MapReduce. Concretamente son abordables sólo aquellos que se pueden disgregar en las operaciones de ***map*** y de ***reduce***. Estas funciones están definidas ambas con respecto a datos estructurados en tuplas del tipo (clave, valor), estos procesos los ejecuta Hadoop por separado (Dean & Ghemawat, 2008).

Roles de Big Data

Dentro de las empresas, las personas con nuevos conocimientos y habilidades para poder llevar a cabo las mejoras del trabajo con Big Data, necesitan incorporar nuevos roles como (Chambi, 2013):

Big Data Administrador

Un perfil con muchas de las habilidades de un administrador de sistema clásico, pero con la capacidad de operar sobre infraestructuras OnCloud, instalar y configurar soluciones y componentes estándar Big Data, definir la arquitectura Big Data utilizada, y ser responsable del mantenimiento, escalabilidad y disponibilidad.

Big Data Architect

Un perfil con una capacidad funcional respecto de las soluciones Big Data. Tiene que ser capaz de conocer las tecnologías de referencia de los componentes Big Data, con experiencia en la definición de arquitecturas, y con una visión global de la solución.

Big Data Developer

Un desarrollador o ingeniero de software especializado en alguno o varios de los componentes habitualmente usados en una arquitectura Big Data. Debe tener conocimientos de procesamiento distribuido.

Data Scientist

Es el perfil estrella de Big Data. Combina una enorme base matemática y estadística con buenas capacidades para la programación y el desarrollo de software. Es imprescindible que conozca el negocio de la empresa, ya que será el encargado de encontrar los algoritmos y patrones que resultarán en información muy valiosa para la empresa.

Data Scientist = Estadístico + Programador + Coach + Artista.
Actualmente, formar equipos con estos perfiles dentro de las

empresas es bastante complicado. La tecnología Big Data está comenzando a llegar a las empresas; y, aunque algunas de ellas tienen suficiente madurez, otras muchas están todavía en sus inicios. Esto implica que es muy complicado encontrar perfiles con las habilidades solicitadas y con alguna experiencia.

Aplicaciones de tecnologías Big Data

En Inteligencia de Negocio, es una nueva forma de visualizar resultados concretos en grandes cubos de cientos de dimensiones, de una forma rápida, barata y dinámica, sin necesidad de procesos de agregación, ya que se pueden aplicar técnicas de Minería de Datos (Fernández, 2013), Inteligencia de Negocios o Análisis Predictivos en tiempo real y en entornos distribuidos (Moreno, 2013). Asimismo, IBM creó una plataforma con el uso de Big Data para recopilar información y analizar la propagación de la enfermedad del Ébola desde mensajes de texto y llamadas telefónicas; y de esta forma, los funcionarios de salud puedan tener una mejor imagen de cómo se expande la enfermedad, la cual, según la Organización Mundial de la Salud, ha infectado a 10,000 personas en la región de Sierra Leona y ha matado a 4,912, expandiéndose rápidamente por el resto del mundo (Rooney, 2014).

En el sector empresarial son múltiples las empresas y compañías que apuestan por el Big Data, tal es el caso de LinkedIn, que lo usa para encontrar el talento más deseado. La compañía presenta dos herramientas que optimizan el proceso de contratación de candidatos. La primera de ellas es una herramienta de visualización de datos que permite a los reclutadores elegir entre una muestra más amplia de candidatos atendiendo a diversos criterios como la localización, el salario o el nivel de experiencia. La segunda analiza tendencias y patrones de contrataciones exitosas llevadas a cabo con anterioridad. La idea es descubrir lo que estos empleados tienen en común para conseguir exitosos fichajes en el futuro (Abuín, 2010).

En el sector hotelero el uso de las tecnologías Big Data también es muy importante. A través del uso de estas tecnologías va a ser posible analizar gran cantidad de información continuamente proveniente de múltiples fuentes tanto internas a la organización (históricos de reservas, flujos de ventas, etc.) como externas (competencia, previsiones demanda, reputación online, etc.) en tiempo real y poner de relevancia datos importantes para tomar las mejores decisiones comerciales, orientar las estrategias de marketing o establecer las políticas de calidad (Abuín, 2010).

En la detección del fraude es utilizado Big Data para construir modelos de predicción más complejos, que investiguen el comportamiento individual de cada individuo y trabajen con conjuntos de datos mucho más grandes a la vez que se mantenga un tiempo de procesamiento pequeño para poder detectar los casos de fraude lo más rápido posible.

Además, Big Data es utilizado para analizar en tiempo real las tendencias de mercado potenciales, y entender las posibilidades futuras para reducir el riesgo al tomar una posición financiera o modificar la cartera de inversiones.

También se trabaja con Big Data en el procesamiento de información WEB. Un servidor web puede guardar en un fichero de registro todas las interacciones que observa entre un usuario o navegador y la aplicación web. Los datos guardados se pueden analizar para poder comprobar cómo se accede a la página web, si hay partes de la página que se acceden muy poco, si al rellenar un formulario los usuarios suelen dejar algún campo en concreto, etc. El resultado de un buen análisis de toda la información recogida puede permitir optimizar la página web para facilitar e incrementar su uso (IRG, 2011).

Problemas con el uso de la tecnología Big Data

Big Data presenta problemas comunes como:

• Capacidad de las personas: al ser una tecnología en desarrollo, la cantidad de personas que tengan el conocimiento para poder procesar de manera correcta el volumen de información es relativamente poco, lo que dificulta el desarrollo de proyectos.

• Estructura de datos: otro gran inconveniente es la manera en la que se almacenan los datos. La forma misma en que está concebida la idea de cómo guardar los datos, en la actualidad, presenta un reto enorme para Big Data. El desafío de hoy es que la mayoría de los almacenes de datos empresariales ven un cliente o una entidad que trabaja con una fila de datos en lugar de una columna. Esa fila se rellena y se actualiza quizá a diario. Al realizar esta actualización, se está perdiendo la información recolectada, lo que conlleva a menor capacidad de predicción o información a procesar.

• La tecnología: el problema de que muchos de sus tecnologías no son en tiempo real o muy dinámica en absoluto. La ejecución de consultas en un clúster suele tener una gran latencia ya que hay que distribuir cada consulta de manera individual, luego, hacer su etapa de reducción, que está trayendo todos los datos de nuevo juntos. Así que algunas tecnologías son de alto rendimiento, pero de alta latencia.

• Privacidad: junto con la obtención de volúmenes de datos incalculables, viene una cantidad de datos que se puede considerar intrusiva, con esta capacidad de Big Data de intentar analizar absolutamente todo, podría darse un examen inapropiado de los datos de usuarios, conllevando rupturas en la privacidad de los datos de los mismos.

• Volumen, variedad y velocidad: la forma de encontrar un equilibrio entre todas ellas depende de la capacidad de plantear un desarrollo sustentable y un plan acorde a las posibilidades tecnológicas de la empresa que desarrolla con esta tecnología.

Retos que enfrenta la tecnología Big Data

El ruido en grandes volúmenes de datos dificulta el procesamiento y análisis. Se vuelven aún más grandes rápidamente, se accede a ellos con poca frecuencia para ayudar al procesamiento asociado con objetivos de nivel de servicio relajados y sin valor probado. Las empresas tienen que capturar volúmenes de datos cada vez más grandes en los que la señal útil está acompañada por un volumen aún mayor de datos que suponen ruido para la mayor parte de las compañías, que buscan modelos rentables de almacenamiento y procesamiento de datos.

Además, otro de los retos que se enfrenta Big Data está el de los

datos multi-estructurados. Los datos de transacciones y eventos que se han ido almacenando, integrando y analizando en los Almacenes de Datos tradicionales y en aplicaciones de Inteligencia de Negocios durante las tres últimas décadas están en gran parte orientados a dejar constancia de lo ocurrido y se definen en términos de esquema explícito; no siempre se puede decir lo mismo de las nuevas fuentes de Big Data. Los datos sociales y de registro de usuarios se caracterizan por su volatilidad: el modelo de información que se usa para entenderlos puede ser implícito en lugar de explícito, puede ser orientado a documento, pudiendo o no incluir algún nivel de organización jerárquica, puede cambiar continuamente o puede que se quieran aplicar diferentes interpretaciones a los datos en tiempo real en función de cada uso y aplicación. Los enfoques tradicionales hacia la integración de datos suelen dar problemas con la captura de datos multiestructurados y tienen aún más dificultades en estos escenarios debido al tiempo y coste que hay entre los datos científicos y el acceso a los nuevos datos. Se ha estimado que los costes de adquisición, normalización e integración de datos representan hasta el 70% del coste total de implementar una base de datos analítica y aun así es más barato que las alternativas (Teradata, 2014).

Conclusiones

La tecnología Big Data, se puede considerar como el futuro de las tecnologías de la información y las comunicaciones. No solamente porque se adapta mejor a los cambios y evolución que está sufriendo la sociedad tanto en tecnología como en necesidades, sino también, porque el análisis de estos grandes volúmenes de datos se convierte en la clave de la competitividad empresarial.

Las empresas pueden obtener grandes beneficios con la implantación de la tecnología Big Data, ganando una notable ventaja al diseñar algoritmos más complejos, teniendo en cuenta grandes conjuntos de datos; permitiendo tomar decisiones en un menor tiempo y más acertada, mejorando las capacidades del servicio y reduciendo los costes. Empresas como Google, Amazon, Facebook, Walmart, Ebay, Apple utilizan las herramientas de la tecnología Big Data dentro de sus procesos de toma de decisiones, las cuales se pueden adaptar a las diferentes investigaciones y proyectos del tema.

Referencia

Abuín, L. http://www.siliconnews.es/2010/12/14/linkedin-desvela-los-adjetivos-mas-comunes-de-los-perfiles-profesionales/?PageSpeed=noscript,.Retrieved 26 de octubre, 2014.

Barranco, R. ¿Qué es Big Data?, 2012, http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/, Retrieved 26 de octubre, 2014.

Banker, Kyle, *MongoDB in Action* (1st edición), 2011, Manning, pp. 375, ISBN 9781935182870

Berman, F.; Fox,G.; Hey, A. J.(2003) Grid computing: making the global infrastructure a reality , Vol. 2, JohnWiley and sons.

Casado, P, El Big Data son sólo datos. Lo importante es lo que se puede hacer con ellos, 2014, Retrieved 26de octubre, 2014, from http://bigdata.ticbeat.com/entrevista-pablo-casado-incubio/.

Chambi, J, .La Evolución del Estadístico: ¿Data Scientist?, 2013, Retrieved 28 deoctubre,2014, from http://www.peruanalitica.com/2013/12/la-evolucion-del-estadistico-data-scientist/.

Chen, Min; Mao, Shiwen; Liu, Yunhao(2014). Big data: a) A survey. *Mobile Networks and Applications*, vol. 19, no 2, p. 171-209.

Chen, C.L. Philip; Chun-Yang Zhang (2014). b) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, 275, 314–347.

Cloud Partners. Uses cases for Hadoop, 2013, Retrieved 28de octubre, 2014, from http://www.cloudpartnerstm.com/wp-content/uploads/2012/09/Use-Cases-for-Hadoop.pdf.

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

Fernández, L. Big data e inteligencia artificial para un aprendizaje individualizado., 2013, Retrieved28 de octubre, 2014, fromhttp://www.euskadinnova.net/es/innovacion-social/noticias/data-inteligencia-artificial-para-aprendizaje-individualizado/11062.aspx.

Foster, I.; Kesselman, C(1999). The Grid2: Blue print for a New Computing Infrastructure, Morgan Kaufmann Publishers.

George, L(2011). HBase: the definitive guide:»O'Reilly Media, Inc.».

Haines, S., Big Data Analysis with MapReduce and Hadoop, 2013, from http://www.informit.com/articles/article.aspx?p=2008905.

Hawkins, Tim; Plugge, Eelco; Membrey, Peter(2010). The Definitive Guide to Mongo DB: The NoSQL Database for Cloud and Desktop Computing (1st edición), ,A press, pp. 350, ISBN 9781430230519

Hewitt, E.(2010) Cassandra: the definitive guide:»O'Reilly Media, Inc.»..

IBM. IBM analytics–conversations: social sentiment index, 2012, Retrieved 26 de octubre2014, from http://www.ibm.com/analytics/us/en/conversations/social-sentiment.html.

IRG. The Enterprise Value of Hadoop., 2011, www.greenplum.com/sites/default/files/IRG_2011-The_Enterprise_Value_of_Hadoop.pdf

Kouzes, Richard T.; Anderson, Gordon A., Elbert, Stephen T.,; Gorton, Ian, Gracio, Deborah K.(2009). The changing paradigm of data-intensive computing, Computer, 42 (1), 26–34.

Kosner, A. You Tube Turns Seven Today, NowUploads 72 Hours of Video Per Minute, 2012, Retrieved 28 de octubre, 2014, from http://www.forbes.com/sites/anthonykosner/2012/05/21/youtube-turns-seven-now-uploads-72-hours-of-video-per-minute/

Lam, C. (2010), Hadoop in action: Manning Publications Co.

Liu, J. (2013), MapReduce is Good Enough? Big Data , 1, 28-37.

Mcafee, Andrew, et al. (2012). Big data. *The management* revolution. Harvard Bus Rev , vol. 90, no 10, p. 61-67.

Manyika, James, et al. (2011). Big data: The next frontier for innovation, competition, and productivity..

Matei, Zaharia (2011). *Spark: In-Memory Cluster Computing for Iterative and Interactive Applications*. Invited Talk at NIPS, Big Learning Workshop: Algorithms, Systems, and Tools for Learning at Scale.

Manaure, A. Lexmark apuntala su estrategia en Big Data, 2013, Retrieved 26 de octubre,2014, from http://www.cioal.com/2013/03/13/lexmark-adquiere-twistage-y-accessvia/

Moreno, A., Big Data: el término de moda en el mundo de la informática, 2013, Retrieved26 deoctubre,2014, from http://www.ibermatica.com/sala-de-prensa/opinion/big-data-el-termino-de-moda-en-el-mundo-de-la-informatica.

Murty, J(2009) Programming amazon web services: S3, EC2, SQS, FPS, and SimpleDB. O'Reilly Media Inc.

ORACLE, Financial services data management: Big data technology in financial services, 2013, From http://www.oracle.com/us/industries/financial-services/bigdata-in-fs-final-wp-1664665.pdf

Pentland, A. Reinventing society in the wake of big data, 2012, Edge.org Conversation, August.

Rogers, S (2011). Big Data is Scaling BI and Analytics. Information Management, Retrieved28 de octubre, 2014, from http://www.information-management.com/issues/21_5/big-data-is-scaling-bi-and-analytics-10021093-1.html.

Rooney, B., IBM usa el 'Big Data' para luchar contra el ébola., 2014, From http://www.cnnexpansion.com/tecnologia/2014/10/27/ibm-usa-el-big-data-para-luchar-contra-el-ebola.

Russom, P (2013), Integrating Hadoop into Business Intelligence and DataWarehousing, Second Quarter.

Sabater, J, Big Data. (Grado en ingeniería informática Bachelor thesis), 2013,Universitat Politècnica de Catalunya, España. Retrieved from http://hdl.handle.net/2099.1/20144.

Shvachko, K., Kuang, H., Radia,S., Chansler, R.(2010), The hadoop distributed file system. Paper presented at the Mass Storage Systems and Technologies (MSST), IEEE 26th Symposium on.

Schönberger, V.M; Cukier, Kenneth (2013). Big data: A www.computerworld.es/sociedad-de-la-informacion/a-que-retos-se-enfrenta-el-big-data.

White, T. (2009). Hadoop: the definitive guide: the definitive guide:» O'Reilly Media , Inc.».

Zikopoulos, P.,Parasuraman, K.,Deutsch,T., Giles ,J., Corrigan, D.(2010) Harness the Power of Big Data The IBM Big Data Platform: McGraw Hill, Professional.

Recibido: 2 de abril de 2018
Aprobado en su forma definitiva:
12 de agosto de 2018

Ireimis Leguen de Varona
Universidad de Camagüey Ignacio Agramonte
Camagüey, Cuba.
Correo-e.: ireimis.leguen@reduc.edu.cu

Yoan Martínez López
Universidad de Camagüey Ignacio Agramonte
Camagüey, Cuba.
Correo-e.: yoan.martinez@reduc.edu.cu

Julio Madera
Universidad de Camagüey Ignacio Agramonte
Camagüey, Cuba.
Correo-e.: julio.madera@reduc.edu.cu