ISSN 0121-1129
eISSN 2357-5328

Conociendo Big Data

Knowing the Big Data

Conhecendo Big Data

Fecha de Recepción: 1 de Octubre de 2014
Fecha de Aceptación: 1 de Diciembre de 2014

Juan José Camargo-Vega*
Jonathan Felipe Camargo-Ortega**
Luis Joyanes-Aguilar***

Resumen

Teniendo en cuenta la importancia que ha adquirido el término Big Data, la presente investigación buscó estudiar y analizar de manera exhaustiva el estado del arte del Big Data; además, y como segundo objetivo, analizó las características, las herramientas, las tecnologías, los modelos y los estándares relacionados con Big Data, y por último buscó identificar las características más relevantes en la gestión de Big Data, para que con ello se pueda conocer todo lo concerniente al tema central de la investigación.

La metodología utilizada incluyó revisar el estado del arte de Big Data y enseñar su situación actual; conocer las tecnologías de Big Data; presentar algunas de las bases de datos NoSQL, que son las que permiten procesar datos con formatos no estructurados, y mostrar los modelos de datos y las tecnologías de análisis de ellos, para terminar con algunos beneficios de Big Data.

El diseño metodológico usado para la investigación fue no experimental, pues no se manipulan variables, y de tipo exploratorio, debido a que con esta investigación se empieza a conocer el ambiente del Big Data.

Palabras clave: Big Data, Hadoop, MapReduce, NoSQL, Análisis de datos, Modelo de datos.

Abstract

Given the importance acquired by the term Big Data, the present investigation aims to study and analyze thoroughly the Big Data state of art. Moreover, a second objective is to study the features, tools, technologies, models and

* D.E.A. - M. Sc. Universidad Pedagógica y Tecnológica de Colombia (Tunja-Boyacá, Colombia). jjcamargovega@uptc.edu.co
** Universidad El Bosque (Bogotá-Cundinamarca, Colombia). jfcamargo@unbosque.edu.co
*** Ph. D. Universidad Pontificia de Salamanca (Madrid, España). luis.joyanes@upsam.es

standards related to Big Data. And finally it seeks to identify the most relevant features that manage Big Data, so it can be known everything about the focus of the investigation.

Regarding the methodology used in the development of the research, included to review the state of the art of Big Data, and show what is its current situation, to know the Big Data technologies, to present some of the NoSQL databases, which are those that allow to process unstructured data formats. Also display data models and the analysis technologies they offer, to end with some benefits from Big Data.

The methodology desing used in this investigation, was not experimental, because no variables are manipulated, neither exploratory ones, because with the present investigation, only begins to know the Big Data evirioment.

Keywords: Big Data, Hadoop, MapReduce, NoSQL, Data Analysis, Data Model

Resumo

Tendo em conta a importância adquirida pelo termo Big Data, a presente pesquisa buscou estudar e analisar de maneira exaustiva o estado da arte do Big Data; além disso, e como segundo objetivo, analisou as características, as ferramentas, as tecnologias, os modelos e os standards relacionados com Big Data, e por último buscou identificar as características mais relevantes na gestão de Big Data, para que com ele possa conhecer-se todo o concernente ao tema central da pesquisa.

A metodologia utilizada incluiu revisar o estado da arte de Big Data e ensinar sua situação atual; conhecer as tecnologias de Big Data; apresentar algumas das bases de dados NOSQL, que permitem processar dados com formatos não estruturados, e mostrar os modelos de dados e as tecnologias de análise deles, para terminar com alguns benefícios de Big Data.

O desenho metodológico usado para a pesquisa foi não experimental, pois não se manipulam variáveis, e sim de tipo exploratório, devido a que com esta pesquisa se começa a conhecer o ambiente do Big Data.

Palavras chave: Big Data, Hadoop, MapReduce, NoSQL, Análise de dados, Modelo de dados.

I. Introducción

El sector empresarial presenta gran desconocimiento sobre lo que significa Big Data; hoy las compañías no saben qué hacer con el gran volumen de datos e información almacenada en diferentes medios o bases de datos, los cuales pueden ser de gran importancia, principalmente en la toma de decisiones. Es por ello que la presente investigación se orientó a evidenciar la importancia de la Big Data y a mostrar que los datos se generan con cierta velocidad y variedad, ocasionando el crecimiento en volumen.

Como un problema de Big Data se puede contemplar la forma como hoy crecen los datos en volumen, velocidad y variedad; esto es debido al gran avance y uso de las tecnologías de información, y al uso diario que las personas hacen de ellas.

La presente investigación es útil para las personas que no tienen mayor conocimiento sobre lo que significa Big Data, sobre sus alcances, sus tecnologías y su aprovechamiento. De la misma forma, a las empresas, independientemente su tamaño, siempre y cuando desconozcan el uso de Big Data, de forma que puedan gestionar datos y convertirlos en conocimiento útil en sus labores diarias.

II. Estado del arte de Big Data

Para iniciar, se presentan algunas definiciones sobre el término Big Data, del cual existen innumerables definiciones, entre ellas se tienen:

Según [1], el término aplica a la información que no puede ser procesada o analizada mediante procesos tradicionales. Para [2], Big Data son "cantidades masivas de datos que se acumulan con el tiempo que son difíciles de analizar y manejar utilizando herramientas comunes de gestión de bases de datos", y para [3], Big Data se refiere "al tratamiento y análisis de enormes repositorios de datos, tan desproporcionadamente grandes que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales".

Por su parte, el analista Dan Kusnetzky, del Grupo Kusnetzky [4], señala que "La frase Big Data se refiere a las herramientas, procesos y procedimientos que permitan a una organización crear, manipular y administrar grandes conjuntos de datos e instalaciones de almacenamiento".

En [5],"Forrester define Big Data como las técnicas y tecnologías que hacen que sea económico hacer frente a los datos a una escala extrema. Big Data trata de tres cosas: 1) Las técnicas y la tecnología, lo que significa que la empresa tenga personal, el cual tenga gran representación y análisis de datos para tener un valor agregado con información que no ha sido manejada. 2) Escala extrema de datos que supera a la tecnología actual debido a su volumen, velocidad y variedad. 3) El valor económico, haciendo que las soluciones sean asequibles y ayuden a la inversión de los negocios".

Según [7], Big Data "se refiere a las herramientas, los procesos y procedimientos que permitan a una organización crear, manipular y gestionar conjuntos de datos muy grandes y las instalaciones de almacenamiento".

Gartner [8] define el Big Data como "un gran volumen, velocidad o variedad de información que demanda formas costeables e innovadoras de procesamiento de información que permitan ideas extendidas, toma de decisiones y automatización del proceso". Y [9] indica que "Big Data consiste en consolidar toda la información de una organización y ponerla al servicio del negocio".

A. Estado actual de Big Data

Las investigaciones en Big Data son bastantes en la actualidad; aquí se presentan algunas de ellas:

Una encuesta realizada por LogLogic encuentra que el 49% de las organizaciones están algo o muy preocupados por la gestión de Big Data, pero que el 38% no entiende lo que es, y el 27% dice que tiene una comprensión parcial [10]; además, la encuesta encontró que 59% de las organizaciones carecen de las herramientas necesarias para gestionar los datos de sus sistemas de TI [10]. Khera explicó que: "Big Data se trata de muchos terabytes de datos no estructurados", donde "La información es poder, y Big Data, si se gestiona correctamente, puede proporcionar una cantidad de conocimiento…" [10]. Según la encuesta, el 62% mencionó que ya había gestionado más de

un terabyte de datos; eso indica que el volumen de datos en el mundo está aumentando a un ritmo casi incomprensible.

Señala Beyer de Gartner y otros expertos que: "los grandes volúmenes de datos, o Big Data, requieren grandes cambios en el servidor, la infraestructura de almacenamiento y la arquitectura de administración de la información en la mayoría de las empresas" [11].

En [11], McKinsey dice que es necesario prepararse para contratar o reciclar personal, pues las empresas u organizaciones carecen de personas capacitadas en Big Data. Además, "proyecta que para el 2018, solo en Estados Unidos, se necesitarán entre 140 mil y 190 mil nuevos expertos en métodos estadísticos y tecnologías de análisis de datos, incluyendo el ampliamente publicitado papel de científico de datos". Señala Williams de Catalina, en [11], que "La gente que construyó las bases de datos del pasado no son necesariamente las personas que van a construir las bases de datos del futuro".

Según el estudio de Worldwide Big DataTechnology and Services 2013-2017 de IDC, "La tecnología y servicios de Big Data crecerá con una tasa anual de crecimiento compuesto del 27% hasta llegar a los 32.400 millones de dólares en 2017, unas seis veces la tasa de crecimiento del mercado general de tecnologías de la información y comunicaciones" [12].

B. Dimensiones de Big Data

Existen tres características o dimensiones: Volumen, Velocidad y Variedad.

1) Volumen: Cada día, las empresas registran un aumento significativo de sus datos (terabytes, petabytes y exabytes), creados por personas y máquinas. En el año 2000 se generaron 800.000 petabytes (PB), de datos almacenados y se espera que esta cifra alcance los 35 zettabytes (ZB) en el 2020. Las redes sociales también generan datos, es el caso de Twitter, que por sí sola genera más de 7 terabytes (TB) diariamente, y de Facebook, 10 TB de datos cada día. Algunas empresas generan terabytes de datos cada hora de cada día del año, es decir, las empresas están inundadas de datos [13].

***2) Variedad*:** Se puede mencionar que va muy de la mano con el volumen, pues de acuerdo con éste y con el desarrollo de la tecnología, existen muchas formas de representar los datos; es el caso de datos estructurados y no estructurados; estos últimos son los que se generan desde páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos o producto de sensores en diferentes actividades de las personas; un ejemplo [14] es el convertir 350 mil millones de lecturas de los medidores por año para predecir el consumo de energía.

3) Velocidad: Se refiere a la velocidad con que se crean los datos, que es la medida en que aumentan los productos de desarrollos de software (páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos, entre otros).

Las tres características tienen coherencia entre sí; por ejemplo [13], analizar 500 millones de registros de llamadas al día en tiempo real para predecir la pérdida de clientes.

C. Análisis de Big Data

El Big Data crece diariamente, como ya se mencionó,y una de las justificaciones es que los datos provienen de gran variedad de fuentes, tales como la Web, bases de datos, rastros de clics, redes sociales, Call Center, datos geoespaciales, datos semiestructurados (XML, RSS), provenientes de audio y video, los datos generados por los termómetros, datos de navegación de sitios web durante cierto tiempo, las RFID (Radio Frequency Identification - identificación por radiofrecuencia) [15].

Existen algunos beneficios del análisis de Big Data para las organizaciones, tal como se observó en el área de marketing, demostrados en la encuesta realizada por TDWI (The Data Warehousing Institute), cuando preguntó: "¿Cuál de los siguientes beneficios se produciría si la organización implementa alguna forma de análisis de Big Data?". El 61% respondió que influye de manera social; el 45%, que habrá más puntos de vista de negocio; el 37% se inclinó por las decisiones automatizadas en los procesos en tiempo real; el 29% mencionó que se mejoraría la

planificación y la previsión, y el 27%, que se entendería el comportamiento del consumidor [16].

Según la misma encuesta, se encontraron ciertos inconvenientes del análisis del Big Data, entre ellos: la falta de personal y de habilidades del recurso humano (46%), la dificultad en la arquitectura de un sistema de análisis de Big Data (33%), problemas con el Big Data utilizable para los usuarios finales (22%), la falta de patrocinio empresarial (38%) y la deficiencia de un argumento empresarial convincente (28%), la carencia de análisis de bases de datos (32%), problemas de escalabilidad de Big Data (23%), rapidez en las consultas (22%) y dificultad para cargar los datos lo suficientemente rápido (21%), entre otros [16].

Ante la pregunta sobre cada cuánto realizan análisis de Big Data, se halló que anualmente, el 15%; mensual, el 35%; semanal, el 14%; diario, 24%; cada pocas horas, 5%; cada hora, 4%; en tiempo real, 4%. Lo anterior fue el resultado de 96 entrevistados [16].

El objetivo del análisis de datos es examinar grandes cantidades de datos con una variedad de clases, con el fin de descubrir información que sea relevante y útil para la empresa, de manera que le permita tomar las mejores decisiones y obtener ventajas competitivas en comparación con otras de su clase.

El análisis de datos se realiza con tecnologías de bases de datos como NoSQL, Hadoop y MapReduce, las cuales soportan el procesamiento del Big Data.

III. Tecnologías de big data

Para el manejo de datos es necesario tener dos componentes básicos, tanto el hardware como el software; respecto al primero, se tienen tecnologías tales como arquitecturas de Procesamiento Paralelo Masivo (MPP), que ayudan de forma rápida a su procesamiento. Para el manejo de datos no estructurados o semiestructurados es necesario acudir a otras tecnologías; es aquí donde aparecen nuevas técnicas y tecnologías, como MapReduce o Hadoop, diseñado para el manejo de información estructurada, no estructurada o semiestructurada.

A. Apache Hadoop

Según [18],"Apache Hadoop es un marco de software de código abierto para aplicaciones intensivas de datos distribuidos originalmente creado por Doug Cutting para apoyar su trabajo en Nutch, una Web de código abierto motor de búsqueda. Hadoop es ahora una de las tecnologías más populares para el almacenamiento de los datos estructurados, semi-estructurados y no estructurados que forman Big Data. Hadoop está disponible bajo la licencia Apache 2.0".

Según [19], "es una biblioteca de software que permite el procesamiento distribuido de grandes conjuntos de datos a través de grupos de ordenadores que utilizan modelos sencillos de programación. Está diseñado para pasar de los servidores individuales a miles de máquinas, cada oferta local de computación y almacenamiento".

Según [20], Hadoop es un framework de código abierto, el cual permite escribir y ejecutar aplicaciones distribuidas que procesan grandes cantidades de datos. Tiene algunas características importantes:

- Fue diseñado para ejecutarse en grupos relativamente grandes de hardware, es decir, en clúster robustos.
- Es robusto, pues ante un mal funcionamiento del hardware puede superar tales situaciones sin mayor inconveniente.
- Tiene la ventaja de poder ser escalable, lo que indica que permite crecer o agregar nodos al clúster con relativa facilidad; por ejemplo, ante la forma vertiginosa como crecen las redes sociales, permite agregar más nodos con facilidad.
- Es simple, por lo que permite a los usuarios escribir código con eficiencia, para software distribuido.

Hadoop tiene sus inicios como un subproyecto de Nutch, que era a su vez un subproyecto de Apache Lucene; es una indexación de texto y de búsqueda bibliográfica, es decir, permite realizar búsquedas dentro de documentos. Nutch es un proyecto más ambicioso que Apache Lucene, lo que se busca es diseñar un motor de búsqueda para la web, el cual

contiene analizador para HTML, un rastreador web, una base de datos de link-gráfica y otros componentes adicionales necesarios.

Hoy en día, Hadoop muestra ventajas significativas frente a bases de datos SQL (Structured Query Language), que se presentan como un diseño para el manejo de información estructurada, donde los datos residen en tablas relacionales que tienen una estructura definida, pues fue diseñado para información no estructurada o semiestructurada, como documentos de texto, imágenes y archivos XML.

"Hadoop puede manejar todos los tipos de datos de sistemas dispares: estructurado, no estructurado, los archivos de registro, imágenes, archivos de audio, archivos de correo electrónico, las comunicaciones…, casi cualquier cosa que se pueda imaginar, sin importar su formato nativo" [21].

Según [20], Hadoop es un framework usado para escribir y ejecutar aplicaciones distribuidas que permite procesar grandes cantidades de datos. Hadoop está compuesto por dos módulos [19]: Hadoop Distributed File System (HDFS), y HadoopMapReduce.

1) Hadoop Distributed File System *(HDFS: Sistema de archivos distribuido Hadoop):* Es un sistema de archivos altamente tolerante a fallos, escalable y con una arquitectura distribuida; puede llegar a almacenar 100 TB en un solo archivo, lo cual no es tan fácil en otros tipos de sistemas de archivos. Además, brinda la apariencia de estar trabajando en un solo archivo, pero realmente lo que se tiene es que están distribuidos en varias máquinas para su procesamiento.

Lam menciona [20] que HDFS se diseñó para el procesamiento por lotes, en lugar de uso interactivo por los usuarios; pero realmente se diseñó para resolver dos problemas importantes que se presentan en el procesamiento de datos a gran escala: el primero es la capacidad de descomponer los archivos en varias partes y procesar cada una independientemente de las demás, y al final consolidar todas las divisiones del archivo en uno solo; el segundo problema era la tolerancia a fallos, tanto en el nivel de procesamiento de archivos como de forma general del software, al momento de realizar el procesamiento de datos distribuidos; lo que se busca es que el sistema pueda recuperarse de la falla que se pueda presentar sin afectar demasiado el proceso [20,21].

Cuando se creó HDFS se propusieron tres objetivos [21]:

- Permitir procesar archivos con tamaños de gigabytes (GB) hasta petabytes (PB).
- Poder leer datos a grandes velocidades.
- Capacidad para ser ejecutado en una máquina, sin solicitar hardware especial.

La arquitectura de HDFS está compuesta por un nodo principal (NameNode) y varios nodos esclavos (DataNodes).

El nodo principal es el servidor maestro, dedicado a gestionar el espacio del nombre de los archivos y controlar el acceso de los diferentes archivos de usuarios; además, el nodo maestro se encarga de gestionar las operaciones de abrir, cerrar, mover, nombrar y renombrar archivos y directorios.

Los nodos esclavos (DataNodes), como su nombre lo indica, representan los esclavos de la arquitectura HDFS. En un HDFS pueden existir miles de nodos esclavos y decenas de miles de clientes HDFS por clúster; esto se debe a que cada nodo esclavo puede ejecutar múltiples tareas de aplicaciones de forma simultánea. La función del nodo esclavo es la de gestionar tanto la lectura como la escritura de los archivos de los usuarios, y realizar la replicación de acuerdo a como lo indique el nodo maestro (NameNode) [21].

2) Hadoop MapReduce: Según [19],"Es un sistema basado en hilados para el procesamiento paralelo de grandes conjuntos de datos". Hadoop MapReduce es un marco de software creado con el fin de hacer aplicaciones que puedan procesar grandes cantidades de datos de forma paralela, en un mismo hardware. Cuando los datos entran para ser procesados se dividen de manera independiente, para su procesamiento, es decir, de manera distribuida en diferente hardware que exista. MapReduce está compuesto de un maestro, llamado JobTracker, y un esclavo, TaskTracker, por cada nodo. El primero se encarga de programar las

tareas, los componentes que manejan el esclavo, y éste ejecuta las tareas según las indicaciones del maestro.

MapReduce es usado en soluciones donde se pueda procesar de forma paralela y, además, con grandes cantidades de información, es decir, con volúmenes de petabytes, de lo contrario no sería una solución adecuada. Otra condición es que se puede usar MapReduce en procesos que se puedan disgregar en operaciones map () y reduce (), las cuales se definen en función de datos estructurados [19]. MapReduce se diseñó como un modelo de programación, para que se pudiera realizar procesamiento de datos de gran tamaño, y, de la misma forma, para que resolviera el problema existente de escalabilidad. MapReduce "es un modelo de programación para el procesamiento de datos". Puede ser ejecutado en varios lenguajes de programación, como Java, Ruby, Python, and C++ [22].

IV. Bases de datos NoSQL

En 1998 aparece el término NoSQL, que significa no solo SQL. El nombre fue creado por Carlo Strozzi, para denominar su base de datos que no ofrecía SQL. Las NoSQL no presentan el modelo de las bases de datos relacionales; estas no tienen esquemas, no usan SQL, tampoco permiten joins (unión), no almacenan datos en tablas de filas y columnas de manera uniforme, presentan escalabilidad de forma horizontal, para su labor usan la memoria principal del computador; su objetivo es gestionar grandes volúmenes de información. Las bases de datos NoSQL tienen como característica principal que su estructura es distribuida, es decir, los datos se hallan distribuidos en varias máquinas [21, 24, 25]. Las bases de datos NoSQL permiten obtener los datos con mayor velocidad que en otras con modelo relacional.

En la Tabla 1 se presenta un ejemplo de una clase de bases de datos NoSQL, con las características mencionadas anteriormente.

Tabla 1
Ejemplo de Base de Datos NoSQL Clave-Valor

Clave	Valor
1	Nombre: Julio; Apellidos: Ríos; Nacionalidad: española
2	Nombre: María; Apellidos: Gutiérrez Castro; Nacionalidad: colombiana; Edad: 30
3	Nombre: Petra; Nacionalidad: italiana

Existen varias clases de bases de datos NoSQL, dependiendo de su forma de almacenar los datos, tales como: almacenamiento Clave-Valor, orientadas a columnas y las orientadas a documentos. A continuación se presentan algunas Bases de Datos NoSQL:

A. DynamoDB

DynamoDB fue desarrollada y probada de manera interna en Amazon; guarda muy fácil y económicamente cualquier cantidad de información. Los datos son almacenados en unidades de estado sólido SSD (Solid State Drive), las cuales permiten mayor velocidad a la hora de encontrar la información, pues estas unidades funcionan de manera diferente a como lo hace el disco duro del computador [40]. Con el uso de SSD se tiene un excelente rendimiento, mayor fiabilidad y un alto grado de seguridad de los datos.

B. Cassandra

Proyecto iniciado por Facebook; es del tipo código abierto (Open Source). Se puede decir que después de la implementación de Cassandra, las redes sociales se dispararon en popularidad [41]. Es una base de datos distribuida, y almacena los datos en forma de clave-valor; fue desarrollada en java, además, hoy en día es usada en la red social Twitter.

Otras características importantes de Cassandra es que es *descentralizada*, lo que significa que cada nodo es idéntico, y, además, que no existe ningún punto único de fallo; que es escalable, es decir, que el software puede atender un número mayor de solicitudes de los usuarios sin que se note algún tipo de degradación en su rendimiento, y que es tolerante a fallos, es decir, que puede reemplazar nodos que fallen en el clúster sin perder tiempo.

C. Voldemort

Voldemort fue creada por LinkedIn, con el fin de solucionar los problemas de escalabilidad que tenían las bases de datos relacionales; los datos los almacena en forma de clave-valor; es de ambiente distribuido, los datos se replican automáticamente en los diferentes nodos o servidores, donde cada nodo es independiente de los demás; permite con cierta facilidad la expansión del clúster, sin necesidad de reequilibrar todos los datos. El código fuente está disponible bajo la licencia Apache 2.0 [42].

D. Google BigTable

BigTable fue creado por Google en el año 2004, con la idea inicial de que fuera distribuido para varias máquinas, por lo que necesitaban que fuese altamente eficiente. El sistema divide la información en columnas, y para almacenarla utiliza tablas multidimensionales compuestas por celdas [25]. El sistema de archivos usado por BigTable es GFS (Google File System) es de tipo distribuido, del mismo propietario Google, y se desarrolló con el objetivo de almacenar información en sistemas de archivos distribuidos con cierta velocidad. Puede almacenar hasta tres copias de la información. Maneja dos servidores diferentes: uno llamado Master, que se encarga de guardar la dirección donde se alojan los archivos, y otro llamado Chunk Server, que es donde almacena los datos. Para terminar, GFS no depende de un sistema operativo específico, es decir, funciona en cualquier plataforma.

E. HBase

HBase es una base de datos de tipo código abierto (Open Source); almacena los datos de forma clave-valor; también almacena y recupera los datos de forma aleatoria, es decir, que al momento de escribir los datos lo hace a su manera, y al leerlos funciona de igual forma. Trabaja con los tres tipos de datos: no estructurados, semiestructurados y estructurados, siempre y cuando no sean tan grandes. HBase no permite consultas SQL y, además, está diseñada para ejecutarse en un clúster de equipos, lo que indica que no puede trabajar en un solo servidor. En la medida que se aumenten más servidores, HBase no presenta inconvenientes en ese sentido, y, también, cuando uno de ellos presenta algún tipo de inconveniente se puede sustituir por otro sin mayor problema [26, 27].

F. Riak

Riak es una base de datos que almacena la información en forma de clave-valor y es de ambiente distribuido, presenta la característica de que es tolerante a fallos, lo que indica que puede eliminar errores y sus efectos antes de que ocurra una falla, buscando de esta manera maximizar la fiabilidad del sistema. Utiliza JSON (JavaScript Object Notation - Notación de Objetos de JavaScript), que es un formato para el intercambio de datos. Además, Riak tiene mayor ventaja a la hora de trabajar en la Web, en la familia de bases de datos de su especie, medida en las peticiones de muchos usuarios simultáneamente [28, 29].

G. CouchDB

CouchDB es el acrónimo en inglés de Clusterof UnreliableCommodity Hardware; fue creado en el año 2005, por Damien Katz. En el 2011 se hace el lanzamiento al público de la versión 1.1.1. Se considera que CouchDB es un servidor de base de datos documental, lo cual indica que los datos no los almacena en tablas, sino que la base de datos está compuesta por documentos, que a su vez trabajan como objetos. Hace uso de JSON, que es un formato para el intercambio de datos, usado cuando los datos son de gran volumen; por eso, para las consultas hace uso de JavaScript; debido a lo anterior, es muy usado por empresas como Yahoo y Google [28, 30].

CouchDB presenta una característica importante: se puede instalar desde un datacenter hasta un Smartphone, y se puede ejecutar en un celular Android, en un MacBook o en un datacenter, lo que quiere decir que se pueden almacenar datos pequeños en un celular, como también grandes volúmenes de datos en

un servidor. También es muy flexible para estructurar y distribuir datos. Otra característica importante es la facilidad con la que permite hacer replicaciones.

Una desventaja consiste en que no permite consultas dinámicas, pues las realiza de manera estática; por ejemplo, para buscar un libro por el nombre de autor, primero crea un índice con todos los nombres de autores para todos los documentos.

Una ventaja desde el punto de vista de seguridad que maneja CouchDB es que cada vez que un documento se almacena nunca se sobrescribe el original, se crea uno nuevo con las modificaciones sucedidas de los datos; lo anterior indica que CouchDB guarda una copia de seguridad de los documentos viejos [28, 31].

H. MongoDB

MongoDB es una base de datos con el perfil NoSQL orientada a documentos, bajo la filosofía de código abierto. La importancia de MongoDBradica en su versatilidad, su potencia y su facilidad de uso, al igual que en su capacidad para manejar tanto grandes como pequeños volúmenes de datos. Es una base de datos que no tiene concepto de tablas, esquemas, SQL, columnas o filas. No cumple con las características ACID, que es el acrónimo de Atomicity, Consistency, Isolation and Durability (Atomicidad, Consistencia, Aislamiento y Durabilidad, en español).

MongoDB permite las operaciones CRUD, que es el acrónimo de Create, Read, Update and Delete (Crear, Obtener, Actualizar y Borrar); para almacenar y recuperar los datos hace uso de JSON, pero utiliza BSON, que es una forma binaria de JSON, el cual ocupa menos espacio al almacenar los datos. Además, BSON es más eficiente y rápida para convertir a un formato de datos de un lenguaje de programación. Otra característica de MongoDB es que realiza consultas dinámicas, es decir, puede realizar consultas sin demasiada planificación. MongoDB se desarrolló en C++ [28, 32, 33, 34].

I. BaseX

Es una base de datos de tipo documental, la cual permite almacenar, recuperar y gestionar datos de documentos; es de la clase de bases de datos NoSQL; tiene como característica importante que permite escalar y, además, que es de alto rendimiento. Su arquitectura es cliente/servidor, permitiendo realizar lecturas y escrituras de datos de manera simultánea. Cumple con el estándar, ACID (acrónimo de Atomicity, Consistency, Isolation and Durability-Atomicidad, Consistencia, Aislamiento y Durabilidad). Soporta grandes documentos en XML, JSON y formatos binarios. BaseX está desarrollado bajo Java y XQuery [35].

V. Modelo de datos

Los datos se clasifican en estructurados, no estructurados y semiestructurados.

A. Datos estructurados

Este tipo de datos se dividen en estáticos (array, cadena de caracteres y registros) y dinámicos (listas, pilas, colas, árboles, archivos). Se puede definir que los datos estructurados son aquellos de mayor facilidad para acceder, pues tienen una estructura bien especificada [31, 36]. Un array es una colección finita de elementos en formatos definidos del mismo tipo, es decir, son homogéneos, y ordenados por un índice; con estos formatos se facilita la administración de los datos; ejemplo de ellos, un campo que contiene una fecha DD, MM, AA, que contiene seis caracteres, o un formato con la dirección de la persona, que puede ser alfanumérico, con tamaño de 40 caracteres.

B. Datos semiestructurados

Estos datos no tienen un formato definido, lo que tienen son etiquetas que facilitan separar un dato de otro. Un dato de estos se lee con un conjunto de reglas de cierto nivel de complejidad [36].

Los datos semiestructurados presentan las siguientes características [23]:

- Son datos irregulares, que pueden no tener un esquema en particular, es el caso del ejemplo que se presenta en las Tablas 2, 3, y 4.

Tabla 2
Datos semiestructurados

Nombre	Teléfono	Sexo	Correo
Pedro Pérez	2127409	M	pedroperez@gmail.com

Tabla 3
Datos semiestructurados

Nombre	Apellido	Teléfono	Correo
Mario	Rodríguez	0987526221	mario@gmail

Tabla 4
Datos semiestructurados

Primer apellido	Segundo apellido	Nombres	Correo	Teléfono
Martínez	Arévalo	Julio	aremar@gmail.com	24356712

- En este tipo de datos semiestructurados se pueden presentar datos incompletos, es el caso del ejemplo que se observa en las Tablas 5 y 6.

Tabla 5
Datos semiestructurados

Nombre	Teléfono	Sexo	Correo
Martínez	2127409		pedroperez@gmail.com

Tabla 6
Datos semiestructurados

Nombre	Apellido	Teléfono	Correo
Mario	Rodríguez	0987526221	mario@gmail

- Los componentes de este tipo de datos, pueden cambiar de tipo (ver Tabla 7).

Tabla 7
Datos semiestructurados

Primer apellido	Segundo apellido	Nombres	Correo	Teléfono
Martínez	Arévalo	Julio	aremar@gmail.com	24356712

- Otra característica de los datos semiestructurados es que pueden aparecer datos nuevos cuya estructura nada tiene que ver con la ya existente, es decir, para seguir el ejemplo, se puede observar las Tablas 7 y 8: la primera presenta cinco campos, y la segunda, seis, y los dos registros dentro del mismo archivo de datos.

Tabla 8
Datos semiestructurados

Primer apellido	Segundo apellido	Primer nombre	Segundo nombre	Teléfono	Correo
Vargas	Castro	Néstor	Julio	25678349	neva@gmail.com

Algunas de las anteriores características se presentan debido a que cada quien publica sus datos a su manera, y esto se presenta en internet; al observar cualquier página web se puede visualizar tal situación, es decir, no existe un formato o estructura definida para presentar los datos.

Es de aclarar que los ejemplos anteriores, expuestos en las Tablas 2 a 8, son parte de un archivo con datos semiestructurados, donde sería difícil realizar cualquier gestión o procesamiento con este tipo de datos, pues el primer motivo es la diferencia de tamaño en los campos de cada registro.

C. Datos no estructurados

Son aquellos que no pueden ser normalizados, no tienen tipos definidos ni están organizados bajo algún patrón; tampoco son almacenados de manera relacional, o con base jerárquica de datos, debido a que no son un tipo de dato predefinido; es decir, no tienen un formato normalizado determinado. Sin embargo, los datos deben poder ser organizados, clasificados, almacenados, eliminados, buscados de alguna forma. Estos datos se pueden observar a diario en correos electrónicos, archivos de texto, un documento de algún procesador de palabra, hojas electrónicas, una imagen, un objeto, archivos de audio, blogs, mensajes de correo de voz, mensajes instantáneos, contenidos Web y archivos de video, entre otros [31, 37].

En este caso de datos no estructurados, no tienen un identificador definido, no se puede reconocer su estado físico ni lógico; tampoco se puede identificar su tipo o clase; su tamaño no se puede encajar en una tabla predefinida, es el caso de los datos contenidos en una página web. Se puede tener el siguiente ejemplo: "Pedro nació el día 24 de noviembre de 1978, y el 20 de septiembre se graduó Julio en la universidad". Como se puede observar, no es tan fácil la administración de este tipo de información, no estructurada.

Lo que sí se puede respecto a los datos no estructurados es hacer uso de los metadatos, es decir, usar datos que puedan describir otros datos. Por ejemplo, en una biblioteca se tiene en fichas o en un sistema de información datos de los libros como: autor, título, editorial, ISBN y tema, entre otros. Lo anterior con el fin de hallar con facilidad un determinado libro; esta es la forma como los metadatos ayudan a buscar datos.

VI. Tecnologías de análisis de datos

A. BigQuery

"Google BigQuery es un servicio web que permite hacer un análisis interactivo de enormes conjuntos de datos hasta miles de millones de filas. Escalable y fácil de usar, permite a los desarrolladores BigQuery y las empresas aprovechar los análisis de datos de gran alcance en la demanda" [38]. BigQuery es un servicio que presta Google con el fin de almacenar y consultar grandes datos no estructurados.

B. ThinkUp

Según [39], "ThinkUp es un potente motor de análisis de datos que permite extraer información de Twitter, Facebook y Google+". Para la instalación es necesario un servidor con PHP y una base de datos en MySQL. ThinkUp se desarrolló bajo licencia GPL y su gran potencial es la extracción de datos; es una aplicación web gratuita, de código abierto, puede almacenar actividades sociales en una base de datos con el control de cada persona [43].

C. Infosphere Streams

Es una plataforma desarrollada por IBM, que permite el análisis de datos en milisegundos [44]. Streams analiza y transforma datos en memoria y en tiempo real, no como sucede con otras aplicaciones, que

primero gestionan, almacenan y por último analizan los datos. Con Streams, los datos se analizan directamente, es decir, en tiempo real, lo cual permite obtener resultados más rápidamente [45, 53]. Un Stream es una secuencia continua de elementos, que para este caso son datos; permite manejar altas tasas de transferencia de datos hasta millones de eventos o mensajes por segundo.

D. Biginsights Infosphere

Es una plataforma desarrollada por IBM para Hadoop, buscando suplir las necesidades de las empresas [45, 53], lo cual se puede lograr facilitando el trabajo de los analistas de sistemas, sin volverlos programadores en una herramienta de difícil manejo. Otra forma es facilitar la consulta de los datos almacenados.

E. System PureData

El sistema PureData es una herramienta de IBM; permite realizar análisis de Big Data en menos tiempo que otras herramientas de análisis; la velocidad de lectura de datos promedia los 128 gigabytes por segundo; fue diseñado para manejar más de 1000 consultas simultáneamente; se puede decir que las consultas son tres veces más rápidas que la versión anterior de InfoSphereWarehouse software; permite el análisis de datos tanto estructurados como no estructurados. SystemPureData permite cargar cinco terabytes en una hora [47].

F. Infosphere Information Server

Es una plataforma de integración de datos, producto desarrollado por IBM; permite limpiar y transformar datos, para luego entregar información confiable a la empresa o negocio. Esta herramienta permite trabajar inteligencia de negocios, facilitando la mejor toma de decisiones; ayuda en el almacenamiento de los datos; reduce costos de operación, al permitir fácilmente la relación entre los sistemas, de manera que proporciona información a otras aplicaciones y a procesos de negocios, lo cual trae consigo mayor agilidad en el negocio de la empresa, es decir, lo que sucede es una transformación del negocio en la empresa [46, 47, 45].

G. Sap Hana

Sap Hana (System Applications Products High-Performance Analytic Appliance) es una herramienta para el análisis de Big Data, la cual se compone de hardware y software, con gran velocidad de procesamiento de datos y en los tiempos de respuesta cuando de consultas se trata; lo anterior debido a que para el procesamiento de datos usa tecnología in-memory [6]. Esta tecnología permite realizar procesamiento de grandes cantidades de datos en la memoria principal del servidor, lo cual trae consigo ofrecer resultados con mayor prontitud, comparados con datos almacenados en el disco del servidor. La tecnología in-memory promete un desempeño entre diez y veinte veces más veloz que los modelos tradicionales basados en disco [48, 49].

H. Oracle Big Data Appliance

Es un software desarrollado por la empresa Oracle, que combina hardware con software optimizado, ofreciendo una solución completa y fácil de implementar para la organización de Big Data. En la parte de hardware, está compuesto por un rack de 18 servidores; cada servidor tiene 64 GB de memoria, es decir, el rack tiene 1,152 GB de capacidad total de memoria. Además, cada servidor tiene dos CPU, y cada uno con ocho núcleos, es decir, que en su totalidad posee 288 núcleos el rack [50].

I. HDinsight

Es un producto Microsoft, basado en Hadoop, permite gestionar datos estructurados y no estructurados de cualquier tamaño, que se pueden llegar a combinar perfectamente con herramientas de Inteligencia de Negocios de Microsoft, fortaleciendo de esta forma los servicios a usuarios y público en general con ayuda de software como Office y SharePoint [24].

J. Textalytics

Textalytics es un software desarrollado por Daedalus (Data Decisions and Language S. A.), dedicado al análisis de texto; extrae con facilidad significado de lo escrito en medios sociales y todo tipo de documentos. Dichos datos se transforman en modelos estructurados para poder ser procesados y gestionados con facilidad. Textalytics, permite realizar tareas

tales como extracción de conceptos, relación entre conceptos, corrección ortográfica, corrección gramatical, corrección de estilo, entre otras funciones, es Multiidioma, pues acepta contenidos en español, inglés, francés y otros idiomas [51].

VII. Beneficios del Big Data

Las empresas que saben sacar provecho del Big Data pueden mejorar su estrategia y así permanecer en el mercado posicionadas, pues hará uso de nuevos conocimientos, con el gran volumen de datos o información que maneja a diario, que inicialmente no se les dio la suficiente importancia, por no tener una herramienta tecnológica que permitiera procesarla. Con la tecnología de Big Data, las empresas pueden ofrecer mejores productos, desarrollar excelentes relaciones con sus clientes, además, se transforman en más ágiles y competitivas [17].

Es importante tener en cuenta algunos pasos para la implementación de Big Data, como se menciona en [52].

- Entender el negocio y los datos. Este primer paso pide un análisis detallado con las personas que hoy laboran y entienden los procesos y los datos que la empresa maneja.
- El segundo paso consiste en determinar los problemas y cómo los datos pueden ayudar. Al momento de conocer los procesos es muy posible que se encuentren los problemas de la empresa o del negocio.
- Establecer expectativas razonables, es decir, definir metas alcanzables; esto se puede lograr si al implementar la solución de un problema éste no presenta alguna mejora, y se debe buscar otra solución.
- Existe una recomendación especial, y es que cuando se inicia un proyecto de Big Data es necesario trabajar en paralelo con el sistema que hoy está funcionando.
- Al tratar de implementar un proyecto de Big Data se debe ser flexible con la metodología y las herramientas; esto se debe a que las dos anteriores son recientes y pueden llegar a presentar problemas al implementarlas. Esto se puede solucionar realizando investigación e inversión en este tipo de tecnología.
- Es importante mantener el objetivo de Big Data en mente; esto porque el proceso es pesado y porque no es tedioso, máxime cuando los métodos y herramientas que usan Big Data para el análisis de datos aún pueden presentar problemas, y la idea es que se mantenga en mente la meta final del proyecto sin desanimarse pronto.

VIII. Conclusiones

Dentro del estado del arte se encuentran desde diversas definiciones del término Big Data por parte de varios investigadores hasta las tecnologías existentes para iniciar un proyecto en una institución de cualquier ramo productivo, comercial o educativo.

Se estudiaron y analizaron las herramientas tecnológicas que se pueden usar a la hora de desarrollar un proyecto de Big Data. Es así como se pudieron observar empresas desarrolladoras de software que presentan herramientas para enfrentar proyectos de Big Data, con sus características.

Se pudieron identificar las características más importantes en la gestión de Big Data, desde los diferentes formatos de datos que hoy existen o se manejan por los usuarios, hasta conocer las tecnologías necesarias para convertir datos no estructurados en información y conocimiento que beneficie tanto a personas como a empresas en la toma de decisiones. Dicha herramienta para tal labor es Hadoop, que, como se mencionó anteriormente, permite convertir datos poco útiles en información estructurada, ayudando de esta forma a los tomadores de decisiones.

Parte de la investigación arrojó que hoy existe un sinnúmero de herramientas tecnológicas para realizar análisis de datos, la gran mayoría basadas en Hadoop, algunas en ambiente web y otras para escritorio, y algunas en ambiente de la nube. Se nota el esfuerzo que han realizado varias empresas desarrolladoras de software, al servicio de los usuarios.

También se pudo conocer una metodología para implementar un proyecto de Big Data, de forma que pueda servir de guía a quienes deseen sacarle un mayor

usufructo a los datos y convertirlos en conocimiento, que les sea útil a las empresas u organizaciones, buscando mayor beneficio en estrategias empresariales.

Referencias

[1] ZDNet.com, CBS Interactive,*What is "Big Data?"*. Disponible en: http://www.zdnet.com/topic-big-data/, 2013.

[2] thinkupapp.com,(2012). Disponible en:http://thinkupapp.com/, 2012.

[3] E. Dans. Disponible en:http://www.enriquedans.com/2011/10/big-data-una-pequena-introduccion.html, 2011.

[4] E. Plugge, P. Membrey & T. Hawkins, *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing*, Published Apress Media LLC, New York, 2010.

[5] B. Hopkins, *Beyond the Hype of Big Data*. Disponible en: http://www.cio.com/article/692724/Beyond_the_Hype_of_Big_Data, 2011.

[6] Business Software, Disponible en: http://www.businessoftware.net/que-es-big-data/, 2013.

[7] Zdnet.com, Big Data. Disponible en: http://www.zdnet.com/search?q=big+data, 2010.

[8] M. Salgado, Oracle apuesta por Big Data con tecnología y proyectos. Disponible en: http://www.computerworld.es/big-data/oracle-apuesta-por-big-data-con-tecnologia-y-proyectos, 2014.

[9] P. Russom, Big Data Analytics, TDWI (The Data Warehousing Institute), 2012.

[10] S. Montoro, Server and Cloud Platform. Disponible en: http://lapastillaroja.net/2012/02/nosql-for-non-programmers/, 2012.

[11] searchstorage.techtarget.com, *Examining HDFS and NameNode in Hadoop architecture*. Disponible en: http://searchstorage.techtarget.com/video/Examining-HDFS-and-NameNode-in-Hadoop-architecture, 2012.

[12] computerworld.es, Disponible en: http://www.computerworld.es/sociedad-de-la-informacion/el-mercado-del-big-data-crecera-hasta-los-32400-millones-de-dolares-en-2017, 2013.

[13] -01.ibm.com, *IBM Big Data and analytics platform.*Disponible en: http://www-01.ibm.com/software/data/bigdata, 2012.

[14] ibm.com, ¿Qué es Big Data? Disponible en: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html, 2012.

[15] es.wikipedia.org, *RFID*. Disponible en: http://es.wikipedia.org/wiki/RFID, 2010.

[16] E. Redmond, & J. Wilson, *Seven Databases in Seven Weeks*, USA: O'Reilly Media, Inc., Pragmatic Programmers, LLC.2012.

[17] Emc.com, *Big Data transforms Business*. Disponible en: http://www.emc.com/microsites/ebook/index.htm#/slide-intro, 2012.

[18] T. Olavsrud, *Big Data Causes Concern and Big Confusion*.Disponible en:http://www.cio.com/article/700804/Big_Data_Causes_Concern_and_Big_Confusion?page=2&taxonomyId=3002, 2012.

[19] hadoop.apache.org, Disponible en: http://hadoop.apache.org/, 2013.

[20] Chuck Lam, Hadoop in Action, Publisher: Manning Publications Co., Stamford, 2011.

[21] Cloudera.com, Cloudera, Inc. Disponible en: http://www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html, 2013.

[22] P. Zikopoulos, C. Eaton, D. DeRoos, T. Deutsch, &G. Lapis, *Understanding Big Data*, USA: McGraw-Hill Books, 2012.

[23] Universidad Simón Bolívar, *Laboratorio Docente de Computación*. Disponible en: http://ldc.usb.ve/~ruckhaus/materias/ci7453/clase3.pdf.

[24] Microsoft, *SharePoint*. Disponible en: http://office.microsoft.com/es-es/sharepoint/informacion-general-de-sharepoint-2013-caracteristicas-del-software-de-colaboracion-FX103789323.asp, 2014.

[25] S. Montoro, Disponible en: http://www.versioncero.com/articulo/596/almacenamiento-distribuido-no-relacional, 2009.

[26] N. Dimiduk, & A. Khurana, *HBase in Action*, USA: Manning Publications Co, 2013.

[27] textalytics.com, *El motor de análisis de texto más fácil de usar*. Disponible en: https://textalytics.com/inicio, 2013.

[28] C. Preimesberger, *eWeek.*Disponiblen: http://search.proquest.com/view/885430073/1366B171EE72EDB474F/1?accountid=43790, 2011.

[29] Basho Technologies, Inc., Disponible en: http://docs.basho.com/riak/latest/, 2011-2014.

[30] T. Juravich, *CouchDB and PHP Web Development Beginner's Guide*, Birmingham – Mumbai: Packt Publishing Ltd., 2012.

[31] L. Joyanes, *Big Data: Análisis de grandes volúmenes de datos en organizaciones*, Editorial Alfaomega, 2013.

[32] networkworld.com, *9 Open Source Big Data Technologies to Watch*. Disponible en: http://www.networkworld.com/slideshow/51090/, 2012.

[33] K. Chodorow, *MongoDB: The Definitive Guide*, Second Edition, USA: O'Reilly Media, Inc., 2013.

[34] S. Francia, *MongoDB and PHP*, USA:O'Reilly Media, Inc.,2012.

[35] BaseXTeam, Disponible en: http://basex.org/products/download/all-downloads/, 2013.

[36] P. Karl, *Moving Media Storage Technologies: Applications &Workflows for Video and Media Server Platforms*, USA: Elsevier, Inc, 2011.

[37] Adelman Sid, Moss Larissa T., & Abai Majid, *Data Strategy,* USA: Prentice Hall, 2005.

[38] Developers.google.com, *Google BigQuery*. Disponible en: https://developers.google.com/bigquery/, 2012.

[39] effectandaffect.es,*ThinkUp, un motor de análisis de datos*.Disponible en:http://www.effectandaffect.es/blog/thinkup-motor-analisis-datos/, 2012.

[40] T. White, *Hadoop: The Definitive Guide*, USA: O'Reilly, Media, Inc, 2009.

[41] T. Rodríguez, Amazon lanza DynamoDB, una base de datos NoSQL desarrollada internamente. Disponible en: http://www.genbetadev.com/programacion-en-la-nube/amazon-lanza-dynamodb-una-base-de-datos-nosql-desarrollada-integramente-por-ellos, 2012.

[42] The Apache Software Foundation, Welcome to Apache Cassandra. Disponible en: http://cassandra.apache.org/, 2009.

[43] The Apache Software Foundation, *ApacheHBase*. Disponible en: http://hbase.apache.org/, 2014.

[44] -03.ibm.com, InfoSphere Streams. Disponible en: http://www-03.ibm.com/software/products/en/infosphere-streams, 2013.

[45] project-voldemort, Voldemort is a distributed key-value storage system. Disponible en: http://www.project-voldemort.com/voldemort/, 2014.

[46] IBM International Business Machines Corporation, IBM InfoSphere Information Server. Disponible en: http://www-01.ibm.com/software/data/integration/info_server/, 2012.

[47] IBM Corporation Software Group Route 100 Somers, *IBM PureData System for Operational Analytics*. NY 10589. Disponible en: http://public.dhe.ibm.com/common/ssi/ecm/en/wad12351usen/WAD12351USEN.PDF, 2012.

[48] Mariño E., *Business Software, In-Memory: edificación de una empresa que opera en tiempo real*. Disponible en: http://www.americaeconomia.com/analisis-opinion/memory-edificacion-de-una-empresa-que-opera-en-tiempo-real, 2011.

[49] itelligence AG, *SAP In-Memory Computing*. Disponible en: http://www.itelligence.es/14878.php, 2013.

[50] J. P. Dijcks, *Oracle: Big Data for the Enterprise*. Disponible en: http://www.oracle.com/technetwork/database/bigdata-appliance/overview/wp-bigdatawithoracle-1453236.pdf?ssSourceSiteId=ocomes, 2013.

[51] StackpoleBeth, Disponible en: http://www.cio.com.mx/Articulo.aspx?id=13527, 2011.

[52] F. Carrasco, *Los 6 pasos que su organización debe seguir para confiar en Big Data*. América Latina. Disponible en: http://www.cioal.com/2013/07/31/los-6-pasos-que-su-organizacion-debe-seguir-para-confiar-en-big-data/, 2013.

[53] P. Zikopoulos, D. deRoos, K. Parasuraman, T. Deutsch, D. Corrigan, &J. Giles, *Harness the Power of Big Data*, McGraw-Hill Companies, 2013.