DOI: http://dx.doi.org/10.21501/21454086.2403

Lámpsakos | No.19 | pp. 67-72 | enero-junio | 2018 | ISSN: 2145-4086 | Medellín–Colombia

Actualidad e importancia de la implementación de Big Data utilizando las herramientas Hadoop y Spark

Present and importance of the implementation of Big Data using the Hadoop and Spark tools

MsC. Lina Montoya Suarez
Universidad Católica Luis Amigó
lina.montoyasu@amigo.edu.co

Gustavo Andrés Gil Restrepo
Universidad Católica Luis Amigó
gustavo.gilre@amigo.edu.co

(Recibido el 06-22-2017, Aprobado el 01-18-2018, Publicado el 27-12-2018)

Estilo de Citación de Artículo:
L. M. Montoya, G.A. Gil, "Actualidad e importancia de la implementación de Big Data utilizando las herramientas Hadoop y Spark", Lámpsakos, no. 19, pp 67-72, 2018
DOI: http://dx.doi.org/10.21501/21454086.2403

Resumen: En el presente artículo se realizó una revisión sobre la actualidad e importancia del Big data a través de las herramientas Hadoop y Spark. En un principio se contextualiza el concepto Big Data desde diferentes autores y haciendo referencia a su importancia en las diferentes organizaciones, teniendo como premisa las tres V que deben estar presentes a la hora de implementar Big Data (Volumen, Variedad y Velocidad).

Luego se analizan las herramientas Hadoop y Spark, identificando su capacidad de hacer más eficiente el procesamiento de grandes volúmenes de datos, de diferentes tipos de datos y a gran velocidad dando solución a los problemas que se presentaban antes cuando se iba a trabajar sobre muchos datos.

Por último se hace una reflexión sobre la importancia del Big Data en la toma de decisiones de una organización, teniendo en cuenta que la toma de decisiones permite que una organización sea competitiva y pueda perdurar en el tiempo.

Palabras Claves: Big Data, Open Data, Volumen, Variedad, Velocidad, Hadoop, Spark, Map Reduce, HDFS, Minería de datos.

Abstract: In the present article, a review was made on the relevance and importance of Big Data through the Hadoop and Spark tools. First the Big Data concept is contextualized from different authors and making reference to its importance in the different organizations, having as a premise, the three V that must be present when implementing Big Data (Volume, Variety and Speed).

Then, the Hadoop and Spark tools are analyzed, identifying their capacity to make more efficient the processing of large volumes of data, of different types of data and at a high speed, solving the problems that arose before when many data were going to be worked on.

Finally, a reflection is made on the importance of Big Data in the decision-making of an organization, taking into account that decision-making allows an organization to be competitive and to last over time.

Key Words: Big Data, Open Data, Volume, Variety, Speed, Hadoop, Spark, Map Reduce, HDFS, Data Mining.

1. INTRODUCCIÓN

En la actualidad los datos se están generando exponencialmente, de una forma que no se imaginaba antes, hoy por hoy se dispone de dispositivos tecnológicos como celulares, portátiles, *Smart TV*, *Smart Watch*, *Tablet* entre otros; los cuales hacen parte de la vida cotidiana y gran parte de la población mundial [1]. Por lo que en la actualidad, hay una gran inquietud sobre el manejo y uso de grandes volúmenes de datos, es allí donde nacen diferentes disciplinas y tecnologías como la Big Data que requiere sacar un beneficio para las diferentes organizaciones y para toda la sociedad [2].

Durante el siglo XXI comenzó una convergencia de diferentes elementos como tecnologías, redes sociales, dispositivos móviles, banda ancha, internet de las cosas, que hicieron posible que en el año 2011 y 2012 se comenzara a usar la *Big Data* como

MsC. Lina Montoya Suarez
Gustavo Andrés Gil Restrepo
DOI: http://dx.doi.org/10.21501/21454086.2403

un elemento primordial para la competitividad de las organizaciones [3]. Debido a este surgimiento en la actualidad han llevado a que estas se preocupen por la gestión de la información y así replantear sus estrategias en cuanto el manejo, tratamiento y uso de la información, estas estrategias pueden influir en gran medida en la capacidad de mantenerse competitiva en el mercado local y mundial para perdurar en el tiempo [4].

Otro aspecto se puede observar en las entidades gubernamentales ya que son un actor muy importante en el tema de *Big Data*, pues cuentan con un gran volumen de información, generada por un municipio o nación, es allí donde aparece el término “open data” o “datos abiertos”, visto también como un movimiento que exige el acceso a los datos almacenados por las diferentes organizaciones del estado, para que haya transparencia en sus procesos y utilización de los datos en búsquedas de beneficios para todos [5].

El fin de este trabajo de investigación en primera instancia fue realizar una reflexión sobre *Big Data*, teniendo presente sus técnicas y herramienta por consiguiente se analizaron los resultados encontrados, y al final se dan la discusión y las conclusiones a las cuales se llegó.

2. ESTADO DEL ARTE

2.1 Big Data

El *Big Data* es una herramienta llamada a la creación de conocimiento, es por eso que la ciencia se puede apoyar mucho de esta herramienta desde una cultura de colaboración científica, de allí nace un movimiento llamado “*Open Science*” que va de la mano con el “*Open Data*” [6].

El *Big Data* hace referencia al gran volumen de datos generados por la humanidad a traves de diferentes dispositivos, el procesamiento y analisis de estos datos son de gran valor para la toma de decisiones de una organización.

Para que se pueda considerar que se esta implementado *Big Data*, los datos deben cumplir con tres caracteristicas: que sea un gran volumen de datos, que sea una gran variedad de datos y que sean procesados a gran velocidad, al cumplir con estas caracteristicas se puede garantizar unos resultados adecuados del *Big Data*.

2.2 Técnicas *Big Data*

El procesamiento de datos ha cambiado a través del tiempo, a medida que la capacidad del hardware y el software han evolucionado, pues el volumen de datos en la humanidad ha crecido exponencialmente, es por eso que a la par del hardware de alta capacidad también es necesario técnicas con algoritmos de alto rendimiento para el procesamiento de datos masivos [7].

Algunos de las técnicas más utilizadas en minería de datos son:

- Las Redes Neuronales: son un modelo compuesto por nodos organizados en capas e interconectados entre sí. Los nodos y sus valores se ordenan siempre buscando el funcionamiento más óptimo, buscando resolver problemas de predicción y clasificación.
- Los arboles de decisiones: son estructuras de nodos organizados jerárquicamente, siendo su principal aplicación la clasificación de los datos y la toma de decisiones de acuerdo a dicha clasificación.
- Los Algoritmos Genéticos: se basan en gran medida en la teoría Darwinista que plantea que los individuos más adaptados son los que permanecen. Así mismo los Algoritmos Genéticos buscan la mejor solución, estableciendo reglas que descartan diferentes soluciones hasta hallar la mejor por lo cual la búsqueda es optimizada al máximo posible.
- Los Vecinos más Cercanos: es una técnica que se encarga de agrupar los nodos de acuerdo a su grado de similitud para que el procesamiento de grandes volúmenes de datos sea más eficiente y en el menor tiempo posible.
- La Reglas de Inducción son una técnica que se encarga de determina reglas o patrones en un conjunto de datos a partir de condiciones “si - entonces” [8].

2.2.1 HADOOP

La herramienta Hadoop es un entorno de desarrollo que permite almacenar, procesar y analizar grandes cantidades de datos. Fue creada con el propósito de responder a las necesidades de implementación de *Big Data*. Unas de sus principales características es que es un software de código abierto. Es escalable, tolerable a los fallos y es distribuido [9].

Su característica de ser un software distribuido es debido a que en su ejecución, Hadoop trabaja con un conjunto de computadores interconectados entre sí a través de una red, siendo el propio software el que toma la decisión de la forma en que se distribuye la información entre ordenadores permitiendo el acceso y manipulación de dicha información desde cualquier computador.

DOI: http://dx.doi.org/10.21501/21454086.2403

La característica de ser un software escalable hace referencia a la posibilidad de que esta tecnología permite aumentar la capacidad sin límite del *cluster* de ordenadores, de acuerdo a la cantidad de computadores y de hardware que se añadan a la red. Otras tecnologías sí tienen límite de capacidad.

La característica de ser una herramienta tolerable a fallos es debido a que todo software distribuido tiende a fallar eventualmente, sin embargo, en estos casos Hadoop permite que cuando un nodo falle, el sistema siga funcionando sin problemas ya que el nodo maestro transfiere las funciones a otro nodo que esté funcionando sin problemas.

La característica de ser un software *open-source* o de código abierto, permite que sea de acceso público, que se pueda descargar de forma gratuita, que pueda ser modificado de acuerdo a las necesidades y que pueda ser distribuido si se desea.

En los sistemas informáticos tradicionales, una sola base de datos alimenta los diferentes ordenadores conectados a una red para sus propios. Esto hace que cada computador tenga que esperar a que otro computador finalice sus procesos, para poder comenzar con los propios, generando así un cuello de botella en la ejecución de procesos con grandes volúmenes de datos.

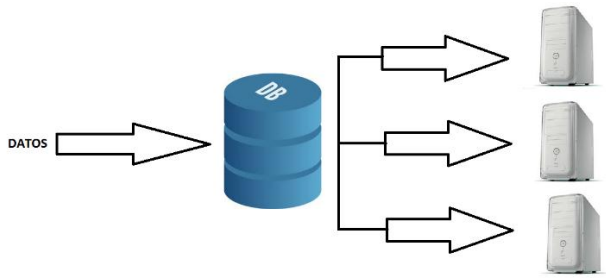

Fig. 1. Procesamiento de datos antes de HADOOP

Los sistemas distribuidos que implementan Hadoop en la gestión de sus datos, cuentan con una réplica de la base de datos para cada ordenador lo que hace posible que cada uno realice sus procesos en paralelo a los otros ordenadores de la red, haciendo posible el procesamiento de un mayor volumen de datos, con más eficiencia y en menor tiempo.

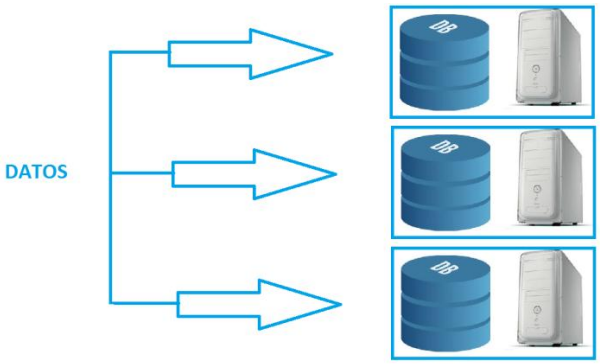

Fig. 2. Procesamiento de datos con HADOOP

Componentes

Hadoop Distributed File System, HDFS: es uno de los principales componentes de Hadoop ya que permite crear diferentes sistemas de ficheros lo que permite tener replicas, mayor capacidad y rendimiento. A su vez brinda la posibilidad de que los datos estén disponibles en cada ordenador para la ejecución rápida y paralela de procesos [10].

Map Reduce: El paradigma *Map Reduce* es un modelo de programación que permite el procesamiento de grandes volúmenes de datos en forma paralela, facilitando el manejo tolerable de errores en la manipulación de datos masivos lo que a su vez también permite de la forma más sencilla que diferentes procesos trabajen simultáneamente e interactúen entre sí [11].

Hadoop creado por *Apache* y desarrollado en el lenguaje Java, es un *Framework* que permite trabajar con miles de nodos y volúmenes de datos expresados en peta bytes. Uno de sus principales componente es el *HDFS* (*Hadoop Distributed File System*), que permite manipular grandes volúmenes de datos a través de un sistema distribuido de archivos. Su aplicación ha sido principalmente en el *Big Data*, llegando a ser la herramienta más usada en esta área.

El *HDFS* funciona a través de dos tipos de nodos, el nodo maestro o *Namenode* y los nodos esclavos o *Datanodes*. El nodo maestro se encarga de manejar los punteros, de ordenar los nodos y de almacenar su ubicación, mientras los nodos esclavos solo se encargan de almacenar archivos a través de bloques [12].

El componente *Map Reduce*, tuvo sus inicios en Google cuando los empleados Sanjay Ghemanwat y Jeffrey Dean hicieron un desarrollo que permitía realizar tareas de forma simultánea.

DOI: http://dx.doi.org/10.21501/21454086.2403

Actualmente *Map Reduce* en Hadoop se encarga de reducir los datos en fragmentos más pequeños con el fin de que su procesamiento sea mejor y en menos tiempo. La función *Map* se encarga de hacer un mapeo creando un paralelo con cada dato de entrada, creando así una lista de datos pares, mientras la función *Reduce* llama la lista de datos pares buscando el resultado deseado y permitiendo trabajar en paralelo [13].

2.2.2 Spark

En la Universidad de Berkeley California se desarrolló el *Framework Spark* con tres principales características: capacidad en analítica, fácil en su manipulación y que desarrolla procesos a altas velocidades.

Se puede decir que *Apache Spark* es una versión mejorada de *Map Meduce*, pues aprovecha el procesamiento simultaneo de grandes volúmenes de datos y aparte, proporciona el Grafo Acíclico Dirigido (DAG), el cual divide el proceso en diferentes tareas que se ejecutan en un *cluster* y que brindan más velocidad y la posibilidad de ejecutar volúmenes de datos más grandes [14].

El *Framework Spark* fue donado de Universidad de California a la fundación Apache como una iniciativa de código abierto y cuenta con una interfaz que permite programar diferentes *cluster* completos de forma paralela y evitando la complejidad ciclomática lo cual lo hace tolerante a fallos y de mayor rendimiento en procesamiento de grandes volúmenes de datos [15].

En el 2009 fue creado por Matei Zaharia en el AMP LAB en la Universidad de Berkeley. En el 2010 pasó a ser un software de código abierto, en 2013 lo recibió la Fundación Apache y lo tomó con un proyecto de alta importancia, de tal forma que para el 2015 ya contaba con 1000 contribuyentes [16].

2.3 Beneficios Big Data

Todas estas tecnologías aplicadas al *Big Data* buscan aportar competitividad a las diferentes organizaciones, sean públicas o privadas, pues es la competitividad la que permite que perduren en el tiempo y puedan alcanzar sus objetivos [17].

La información ha tomado gran relevancia al interior de las organizaciones, tanto así que la simbolizan como el petróleo de actualidad, pues su valor puede ser retribuido monetariamente si es bien utilizada [18].

Al interior de toda organización se debe tener muy claro dónde está el origen de la información que se requiere para dar respuesta a los problemas que acontecen día a día, también cómo se va a filtrar lo realmente relevante para la organización y por último cual es la gestión o cuales son las acciones que se van a emplear para que esta información realmente cobre valor [19]. Por lo tanto en la minería de datos se buscan los datos que realmente tienen valor para la organización, pues el volumen de datos con los que trabaja el *Big data* es demasiado grande, por lo tanto lo importante es encontrar patrones que ayuden a entender el comportamiento del sistema circundante [20].

El *Big Data* es un gran insumo para la generación de conocimiento en las organizaciones, pues brinda información oportuna para la toma de decisiones, que se traducen en acciones que al final se transforman en conocimiento, siendo la gestión de este conocimiento la que posibilita anticiparse a los cambios del entorno [21].

3 ANÁLISIS DE LOS RESULTADOS

Al revisar la literatura se encontró; en la actualidad el *Big Data* es una herramienta que está siendo utilizada cada vez más al interior de las organizaciones tanto a nivel nacional como internacional para dar respuesta a los diversos problemas é incógnitas que surgen en la operatividad del día a día. Es fundamental reconocer el crecimiento exponencial de la generación de datos, lo que ha hecho necesario el cumplimiento de requerimientos tanto en software y como de hardware los cuales son necesarios al momento de manipular grandes volúmenes de datos.

De esta forma se identificó que la Fundación Apache ha sido una organización que ha hecho grandes aportes en cuanto a desarrollos de software que ayudan a implementar el *Big Data* en las organizaciones, así pues, Apache participó en el desarrollo del *Framework* Hadoop, el cual en la actualidad es el más utilizado y el que ha permitido avanzar en el procesamiento de grandes volúmenes de datos, a través de sus componentes *HDFS* (que permite la réplica de los datos) y *Map Reduce* (qu permite el procesamiento en paralelo de los datos).

Otro punto que se identificó como un gran desarrollo para *Big data*, es la tecnología SPARK la cual permite trabajar en conjunto con Hadoop y que es mucho más rápida y eficiente pues trabaja utilizando la memoria y no el disco duro. En la Tabla 1 se observan la comparación de características entre Map Reduce y Spark.

DOI: http://dx.doi.org/10.21501/21454086.2403

Tabla 1. Características *Map Reduce* y Spark

Características	Map Reduce	Spark
Desarrollado por	Google	Universidad de Berkeley
Diseñado para	Procesar segmento de datos	Procesar en tiempo real
Lenguaje de desarrollo	Java	Scala
Soporte de proceso en memoria	Si	No
Resultados son almacenados en	Disco duro	Memoria
Tolerable a fallos por	Tener replica de datos	Registro de transformación
Cuello de botella	Acceso a disco frecuente	Gran consumo de memoria
Capacidad de datos	102.5 Tb	100 Tb
Tiempo de proceso	72 min	23 min
Cantidad de nodos	2100	206

4 CONCLUSIONES

Después de investigar sobre el ámbito que rodea al *Big Data* se concluye que es una herramienta que tiene la capacidad de influir positivamente en el futuro de una organización, pues da la posibilidad de tomar mejores decisiones y de identificar de una forma muy aproximada la realidad del entorno que nos rodea. Es por esto que cada vez más organizaciones a nivel mundial están adoptando el *Big Data* como una de sus principales herramientas para la toma de decisiones y así mismo están implementando diferentes tecnologías como Hadoop y Spark que les ayuda a obtener respuestas casi que en tiempo real, algo que era imposible en el pasado.

También se concluye que es momento de que las organizaciones en Colombia comiencen a implementar *Big Data,* para direccionar sus estrategias y de esta forma se logrará tener una economía más competitiva y actualizada con las Tecnologías de la información vigentes.

5 REFERENCIAS

[1] V. M. Schönberger and K. Cukier, *Big data: la revolución de los datos masivos*. Turner, 2013.

[2] J. Serrano-Cobos, “Big data y analítica web. Estudiar las corrientes y pescar en un océano de datos,” *El Prof. la Inf.*, vol. 23, no. 6, pp. 561–565, 2014.

[3] L. J. Aguilar, *Big Data, Análisis de grandes volúmenes de datos en organizaciones*. Alfaomega Grupo Editor, 2016.

[4] J. G. Cantero, “Nuevas estrategias de gestión de la información,” *Big Data*, vol. 95, p. 51, 2013.

[5] A. Ferrer-Sapena and E. Sánchez-Pérez, “Open data, big data:?` hacia dónde nos dirigimos?,” *Anu. ThinkEPI 2013*, vol. 7, pp. 150–156, 2013.

[6] A. López Borrull and A. Canals, “La colaboración científica en el marco de nuevas propuestas científicas: Open Science, e-Science y Big Data,” *La Colab. científica una aproximación Multidiscip. Val. Nau Llibres*, pp. 91–100, 2013.

[7] M. A. Murazzo, N. R. Rodriguez, M. J. Guevara, and F. G. Tinetti, “Identificación de algoritmos de cómputo Intensivo para big data y su implementación en clouds,” in *XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina)*, 2016.

[8] M. Coto-Jiménez, “Minería de datos: concepto y aplicaciones,” *Una ojeada a Clasif. del suelo Globos Cantolla II vuelo Vert. Arduino uno para prototipado rápido Gener. números aleatorios El bosón Higgs, la partícula divina*, p. 60, 2014.

[9] A. C. C. Herráez, “Big data con Hadoop-I,” 2015.

[10] B. Sarmiento, M. Hernández, and X. Gómez, “Herramientas y antecedentes Big Data,” *Rev. Investig. y Desarro. en TIC*, vol. 5, no. 2, 2017.

[11] A. Hernández Dominguez and A. Hernández Yeja, “Acerca de la aplicación de MapReduce+ Hadoop en el tratamiento de Big Data,” *Rev. Cuba. Ciencias Informáticas*, vol. 9, no. 3, pp. 49–62, 2015.

[12] L. F. Vásquez Rugel, L. A. Caviedes Ruiz, and others, “Sistema de archivos por capas en Hadoop HDFS,” Espol, 2017.

[13] J. L. Larroque, “Indexado de Wikipedia a través de una arquitectura Map Reduce,” Facultad de Informática, 2017.

[14] A. Fenna Víchez, “Captura y gestión de open data en entornos de smart city,” 2017.

[15] M. Niño and A. Illarramendi, “ENTENDIENDO EL BIG DATA: ANTECEDENTES, ORIGEN Y DESARROLLO POSTERIOR,” *DYNA New Technol.*, vol. 2, no. 1, pp. 1–8, 2015.

[16] S. A. Valenzuela, C. L. Vidal, J. D. Morales, and L. P. López, “Ejemplos de Aplicabilidad de Giraph y Hadoop para el Procesamiento de Grandes Grafos,” *Inf. tecnológica*, vol. 27, no. 5, pp. 171–180, 2016.

[17] K. Esser, W. Hillebrand, D. Messner, J. Meyer-Stamer, and others, “Competitividad sistémica: nuevo desafío para las empresas y la política,” *Rev. la CEPAL*, vol. 59, no. 8, pp. 39–52, 1996.

[18] D. Cohen Karen, E. Asin Lares, D. G. Lankenau

DOI: http://dx.doi.org/10.21501/21454086.2403

Caballero, and D. Alanis Davila, “Sistemas de informacion para los negocios: Un enfoque para la toma de decisiones.,” 2005.

[19] G. Ponjuán Dante, “Gestión de información en las organizaciones: principios, conceptos y aplicaciones,” 1998.

[20] H. Orallo, J. RAMIREZ, C. R. QUINTANA, M. Josej. H. Orallo, M. J. R. Quintana, and C. F. Ramírez, *Introducción a la Minería de Datos*. Pearson Prentice Hall, 2004.

[21] A. Blázquez Manzano, “La información y comunicación, claves para la gestión del conocimiento empresarial,” 2013.