Información Importante

Diseño e implementación de un clúster de PCs bajo la tecnología Hadoop para la analítica de datos del consultorio TIC

Comité de proyectos de grado

Meleny Luna Ortiz

Director:

Juan Carlos García Ojeda

Docente Facultad Ingeniería de Telecomunicaciones

Codirector:

Rodolfo Sánchez García

Docente Facultad Ingeniería de Telecomunicaciones

Facultad de Ingeniería de Telecomunicaciones

Universidad Santo Tomás

Bucaramanga, 2017

ELABORADO POR:	**REVISADO POR:**	**APROBADO POR:**
_____ AUTOR *Estudiante de Pregrado*	_____ Director *Director del Trabajo de Investigación*	Comité de trabajos de grado

Dedicatoria

El presente trabajo de grado lo dedico primeramente a Dios, quien me ha acompañado y me ha dado la inteligencia a lo largo del proceso, a mi familia – especialmente a mis padres – que siempre me brindó su apoyo incondicional, y a todas las personas que se sumaron para cumplir esta meta.

Agradecimientos

Quiero agradecer a los Ingenieros Juan Carlos García Ojeda y Rodolfo Sánchez García por su acompañamiento en esta etapa final, por aportar sus conocimientos en el desarrollo del trabajo de grado, y brindarme consejos de vida.

A la Universidad Santo Tomas por formarme como Profesional e inculcarme valores humanísticos y muy especialmente a la ingeniera Natalia Flórez por guiar mi camino en la consecución de mi gran meta de ser ingeniera de Telecomunicaciones.

Agradezco a todos mis compañeros con los que compartí los momentos más gratos a lo largo de este valioso camino, y se convirtieron en amigos para el resto de la vida.

Resumen

Este proyecto de grado describe el diseño de un clúster de computadores para la gestión de grandes volúmenes de datos utilizando la tecnología Hadoop, con la finalidad de ofrecer servicios de analítica de datos a los diferentes sectores económicos que hacen vida en la región a través del consultorio TIC de la universidad Santo Tomas seccional Bucaramanga.

Para ello, en el cuerpo del trabajo se refiere el diseño, implementación y pruebas de carga de la infraestructura tecnológica propuesta.

Cabe resaltar que el montaje del clúster fue realizado con equipos que fueron dado de baja en la institución y bajo rendimiento, aspecto que se evidencia en las pruebas de cargas realizadas. Donde fue recurrente la aparición de errores en disco, memoria y comunicaciones.

Palabras Claves: Clúster de computadores, Big Data, Hadoop, Computo distribuido, Consultorio TIC, USTA.

Tabla de contenido

Figuras

Tablas

IMPLEMENTACIÓN DE UN CLÚSTER

Introducción

En este documento se describe el diseño e implementación de un clúster de computadores bajo la tecnología Hadoop como herramienta para la gestión, distribuida, de grandes volúmenes de información. Para comprender los beneficios de la implementación del clúster debemos empezar por definir ¿Qué es Hadoop? Hadoop es un tipo de tecnología que permite almacenar, procesar y analizar grandes volúmenes de datos por medio de la computación distribuida utilizando modelos sencillos de programación.

Las principales características de dicha tecnología se basan en la rentabilidad del sistema, el bajo costo en los terabytes de almacenamiento, flexibilidad recibiendo cualquier tipo de dato (estructurado – no estructurado), no basando su funcionamiento en esquemas, permitiendo agregar un sinnúmero de nodos y cuenta con respaldo cuando se presentan fallas en los nodos.

El interés que se tiene para el desarrollo del proyecto es adquirir nuevos conocimientos con el software Hadoop y el manejo de grandes volúmenes de información, adicional a esto, proveer al consultorio TIC de la Universidad de un nuevo servicio que puede ser ofrecido a diferentes empresas del sector e inclusive en la misma institución académica.

Inicialmente se realizó el diseño y adquisición de todos los equipos y herramientas que se necesitaban para realizar la implementación del clúster, se continuó con la configuración del clúster: instalación del sistema operativo en los equipos utilizando la versión UBUNTU 17.04, configuración de la red local, utilizando un switch (Cisco Systems, Catalyst 2950) para realizar la interconexión de cada uno de los equipos de cómputo que hacen parte de la infraestructura, por último, realizar la instalación y configuración local y distribuida de Hadoop.

Para evaluar el rendimiento de la infraestructura tecnológica se llevaron a cabo (9) experimentos que se basaron en contar palabras (anagramas de uno y dos palabras, tomadas de

GoogleBooks[1]) en volúmenes de datos de distintos tamaños (100 MB, 200 MB, 500 MB, 600 MB, 800 MB, 900 MB, 1 GB, 6 GB, 12 GB). Los cuales se ejecutaron 10 veces cada uno, con diferentes arquitecturas: Clúster y Equipo Unitario; para luego obtener el promedio y evaluar el beneficio de la infraestructura propuesta.

[1] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

1.1. Formulación del problema

La Universidad Santo Tomás de Bucaramanga, cuenta desde el año 2014 con el Consultorio TIC (Tecnología de la Información y las Comunicaciones), el cual es el encargado de promover la incorporación y apropiación de las TIC, buscando mitigar la brecha digital en la sociedad por medio del desarrollo comunitario, consultoría y capacitaciones.

Anteriormente, la metodología de recolección de datos era un proceso tedioso y se requería de varias herramientas para la consolidación y tiempo para realizarlo; en la actualidad, la adquisición de datos se realiza en muchos escenarios y se generan de una forma más sencilla.

A partir de la recolección y análisis de los datos, la toma de decisiones en las organizaciones se hace más sencilla; esto implica la realización de procesos de análisis, permiten involucrar a todas las variables que sean consideradas por la organización como pertinentes para su manejo, gestión y control, y encontrar información implícita dentro de ellos: ocurrencias que sean repetitivas, un patrón de eventos que tenga relación con un suceso, entre otros. [1]

Por otro lado, la cultura de los datos no está muy desarrollada en el país, lo cual impide una utilización óptima de un activo sustancial como son los datos. Por ejemplo, se estima que, en Colombia, sólo 65% de los datos son recolectados efectivamente y de estos solo el 58,87% de esos datos son relevantes. [2]

¿Cuenta el consultorio TIC con la infraestructura tecnológica necesaria para ofrecer servicios de analítica de datos a grandes volúmenes de información generados por empresas del sector productivo de la región?

1.2. Justificación

Debido al avance tecnológico las instituciones, empresas y organizaciones generan gran cantidad de datos producto de sensores, redes sociales, videos, fotografías, música, transacciones financieras, entre otros, que no pueden ser procesados por los sistemas de bases de datos convencionales. Lo anterior implica que las organizaciones están perdiendo oportunidades valiosas de obtener nuevo conocimiento de todo ese mar de información que se está generando.

En Colombia, se estima que, sólo el 65% de los datos son recolectados efectivamente y de estos sólo el 58.87% es relevante [2]. Por lo cual se infiera, que la cultura de datos no está muy desarrollada en el país, lo cual impide una utilización óptima de un activo sustancia como son los datos.

Razón por la cual se abre una oportunidad importante para que el CTIC explote servicios de analítica de datos en función de los datos recogidos diariamente por los diferentes sectores económicos del Área Metropolitana de Bucaramanga. Sin embargo, el CTIC no cuenta con la infraestructura tecnológica mínima necesaria para ofrecer tales servicios de analítica de datos. Motivo por el cual es necesario el diseño, montaje, y validación de una solución tecnológica que inicialmente permita la gestión de grandes volúmenes de información.

1.3. Objetivos

1.3.1. Objetivo General

Diseñar e implementar una infraestructura tecnológica para cómputo distribuido empleando la tecnología Hadoop como herramienta para ofrecer nuevos servicios en el consultorio TIC.

1.3.2. Objetivos específicos

- Identificar las características de un clúster de computadores, por medio de estudio de referentes tecnológicos de gestión de computación distribuida, para definir los componentes de la solución.

- Describir el funcionamiento del software Hadoop, para identificar sus componentes principales, capacidades, limitaciones y ventajas aplicadas para la gestión de computación distribuida, por medio del análisis de experimentos.

- Validar conceptual y tecnológicamente la infraestructura tecnológica para computo distribuido definida.

2. Marco de referencia

2.1. Acuerdo CAOBA

Colombia cuenta con un primer acuerdo público - privado que promueve el Big data y la analítica en Colombia, denominado CAOBA. Fue creado el 3 de marzo del 2016 por invitación directa del Ministerio de Tecnologías de la Información y las Comunicaciones (TIC) y el Departamento Administrativo de Ciencias, Tecnología e Innovación (COLCIENCIAS), integrado por distintas empresas del sector público y privado con el propósito de fortalecer la generación de soluciones de análisis de información, las universidades aportan la parte investigativa y las empresas ofertas de productos en tecnología con el objeto de generar servicios y soluciones alrededor de las tecnologías del BD&DA. [3]

2.2. Eventos TIC en Colombia, relacionados con Big Data.

En el primer Foro de Ingeniería de Información realizado en la Universidad de los Andes en el año 2016, el Ingeniero de Sistemas Latam de la compañía Cloudera, Joao Salcedo, impartió la conferencia titulada "Cómo sacarle proyecto a la plataforma Hadoop", en la que hace la presentación de la compañía localizada en California, Estados Unidos, y que fue creada por miembros de Oracle, Yahoo, Google y Facebook; mencionó como aspectos relevantes que trabajan en una plataforma de análisis y gestión de datos desarrollada sobre Hadoop con presencia en 20 países y brindando servicio a alrededor de 800 empresas. Además, Cloudera cuenta con un equipo especializado en conocimiento en estadística, programación y del negocio, donde se fusionan las tres ramas y generan datos de valor a cada una de las organizaciones de manera individual.

Otro de los invitados al Foro fue el Ingeniero Doug Cutting, quien actualmente ocupa el cargo de Arquitecto jefe de Cloudera y fue el desarrollador del Software Hadoop, impartió una conferencia titulada "The Future of Big Data and Why I Created Hadoop" donde principalmente

se refiere que se dedicó a crear motores de búsqueda, sistemas para recuperar información, guardar y procesar datos y en base a esto creó Hadoop. Además, explica que se trata de una multiplataforma de computación distribuida programada en java, de código abierto (Open source), más barata, flexible y fácil de escalar; en ella se hay muchos computadores organizados en clústeres, participantes de una infraestructura conectada por una red de telecomunicaciones. [4]

2.3. Experiencia BIOS

En el Ecoparque de selva Húmeda Tropical, Los Yarumos, de la ciudad de Manizales se encuentra operando el Centro de Bioinformática y Biología Computacional (BIOS) desde el 2013, gracias a un acuerdo de cooperación dada entre Colciencias Y MinTic y la ayuda de la Alcaldía de Manizales.

BIOS cuenta con computación de alto desempeño más rápida de Latinoamérica, que permite resolver problemas complejos a la gran cantidad de información generada de las organizaciones e instituciones, igualmente cuenta con talento humano altamente especializado que permite realizar procesamiento de datos, convirtiendo el Big data en un gran activo de la nación y aliado para la investigación. [5]

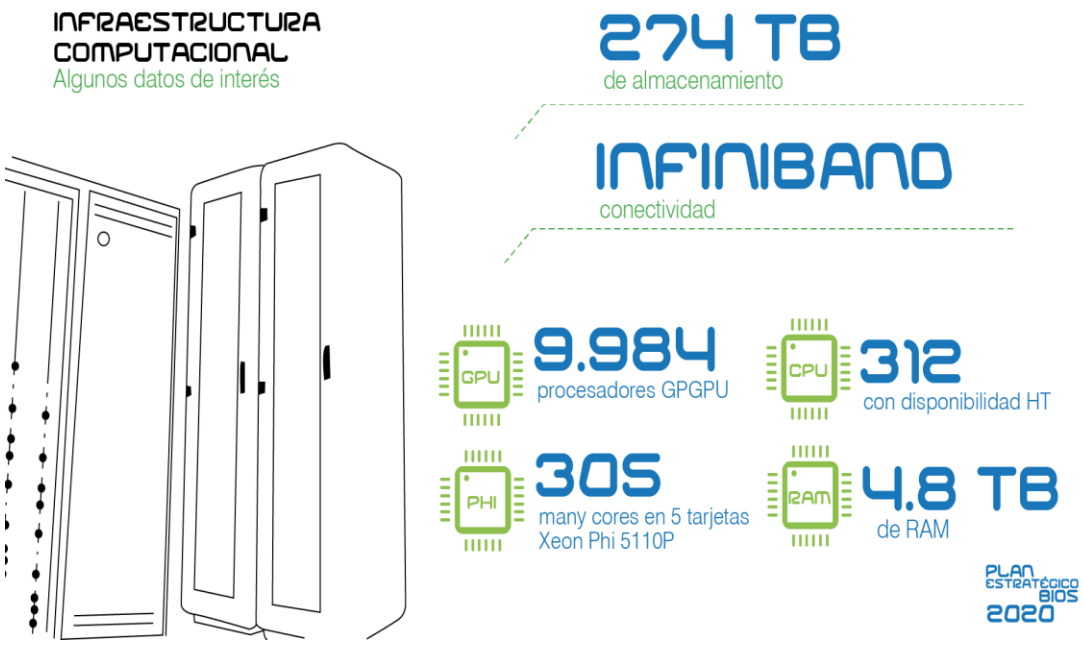


Figura 1. Infraestructura computacional
Fuente [5]

2.4. Computación distribuida y paralela

Computación distribuida, se define como al conjunto de computadores independientes que son capaces de distribuir y llevar a cabo una tarea y que, ante el usuario, se comporta como un único computador y se caracteriza por su heterogeneidad en la ejecución de procesos. Su forma de operar se realiza dividiendo el trabajo en pequeñas tareas individuales, que reciben los datos necesarios para llevar a cabo la tarea, la hacen y devuelven los datos para unirlos en un resultado único y final. [6].

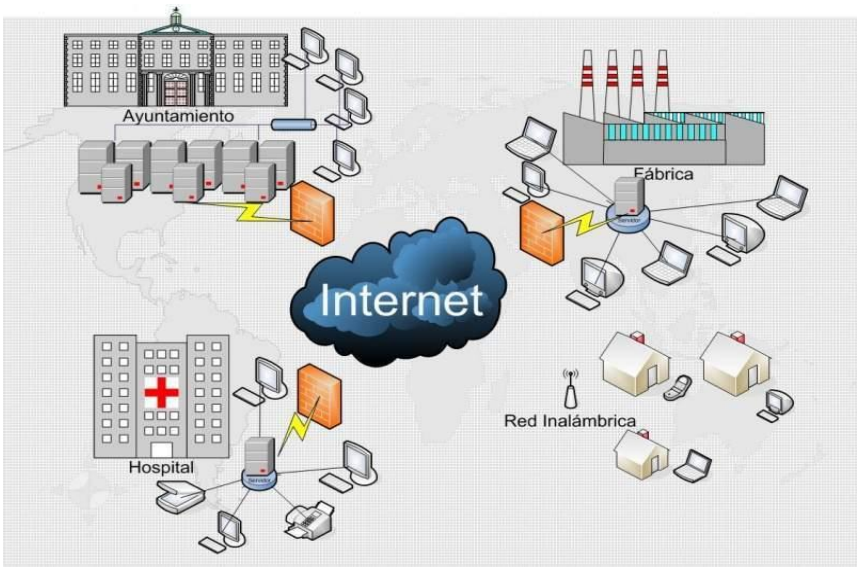


Figura 2. Computación distribuida**.**
FUENTE [6]

Computación en paralelo, se define como una forma de computo donde varias tareas se ejecutan simultáneamente, se basa en el principio de dividir un problema grande en varios pequeños para ser resueltos de forma simultánea. [7]

TABLA 1
Comparación entre computación paralela y distribuida.

COMPUTACION	
PARALELA	**DISTRIBUIDA**
Sistemas simples con varios procesadores trabajando sobre el mismo problema	Varios sistemas acoplados por un secuenciador de trabajo sobre problemas relacionados

2.5. Clúster

En informática, se define a la combinación dos o más computadores independientes, a través de software de código abierto (por ejemplo, plataforma LINUX) y redes para dar solución a problemas específicos. Este tiene como característica principal proveer de alta disponibilidad y alto rendimiento, a las tareas computacionales que se asignen a desarrollar en él. El clúster se

implementa para proporcionar una mayor potencia computacional de lo que una sola computadora puede hacer. [8].

2.6. Big data

Enormes cantidades de datos estructurados, no estructurados y semiestructurados que no pueden ser analizados utilizando procesos o herramientas tradicionales, aunque cabe resaltar que va de la mano con las bases de datos relacionales.

Big data normalmente maneja volumen de datos en términos de petabytes y exabytes y dichos datos son de gran variedad de fuentes que lo generan de diferentes escenarios como son los equipos móviles, gps, audios, videos, sensores, entre otros que necesitan ser procesados de manera rápida para que sea de ayuda para la toma de decisiones en el tiempo indicado. [9]

2.6.1. Beneficios de Big Data

El Big data trae beneficio en las organizaciones significativamente, como es el aprovechamiento de toda la información obtenida de diferentes medios, que ayudan a fidelizar y tener un punto diferenciador comparada con otras organizaciones, uno del ejemplo de los beneficios de Big Data se ven en la Tabla 2.

Tabla 2
Transformación de los negocios con Big Data

Toma de decisiones actual	**Toma de decisiones mediante Big Data**
"vista" a posteriori	Recomendaciones "mirando hacia adelante"
menos de un 10 por cien de datos disponibles x	Aprovechar los datos de fuentes diversas
agrupados, incompletos, inconexos	En tiempo real, relacionados y controlados
monitorización empresarial	Optimización empresarial

Nota:" Big Data El poder de los datos " Bill Schmarzo; 2014

2.7. Analítica de datos

Se le llama análisis de datos a todas las tareas y actividades orientadas a la exploración de datos que ayudan de manera fundamental a la toma de decisiones en tiempo real y revelan patrones para la optimización de los procesos de las empresas. [10]

2.8. Hadoop

Es un software de código abierto, que permite la computación distribuida de grandes cantidades de datos en clústeres; está diseñado para extender una arquitectura de computo monolítico, a una arquitectura multimaquina con alto grado de tolerancia a fallas [11]

2.8.1. Características de Hadoop.

Redimensionable: Permite agregar nuevos nodos sin necesidad de variar el formato de los datos, carga de datos y escritura de procesos.

Rentable: Bajo la computación paralela en los servidores el precio por terabyte de almacenamiento es mínima.

Flexible: Hadoop permite cualquier tipo de datos, estructurados o no, y su funcionamiento es sin esquemas, agrupa los datos de forma arbitraria permitiendo el análisis de datos de alta calidad.

Tolerante a fallas: si un nodo deja de funcionar, Hadoop permite redirigir el trabajo sin generar retrasos en el análisis.

2.8.2. Funcionamiento de Hadoop

Hadoop se apoya principalmente en dos conceptos:

2.8.2.1. HDFS (Hadoop Data File System).

HDFS es un sistema de archivos distribuidos, escalables y portátiles escrito en JAVA; su arquitectura se basa en dos pilares básicos que son descritos a continuación:

- **Namenode**: se ocupa del control de acceso y tiene la información sobre la distribución de datos en el resto de los nodos. Cabe destacar que namenode cumple la función de un equipo maestro en una arquitectura cliente / servidor tradicional.
- **Datanodes**: son los encargados de ejecutar el cómputo, es decir, las funciones de mapeo y reducción, sobre los datos almacenados de manera local en cada uno de ellos. Los datanodes actúan como esclavos en una arquitectura cliente / servidor tradicional.

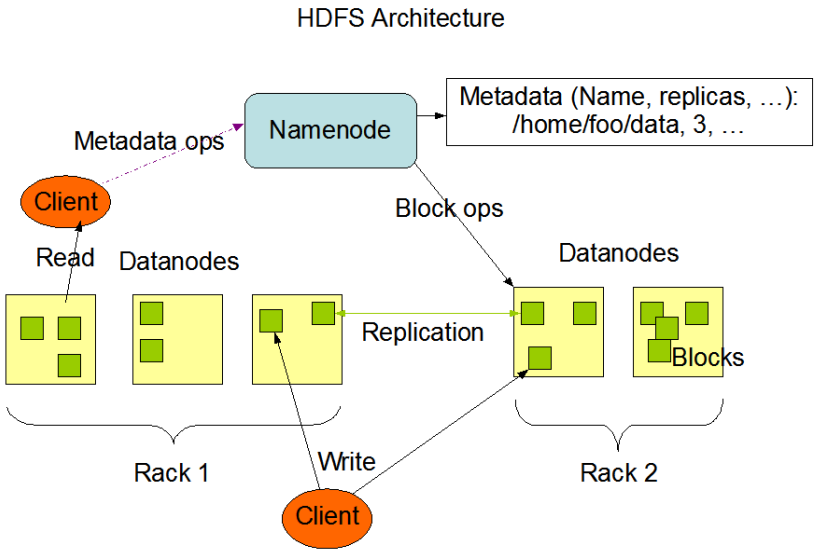


FIGURA 3. ARQUITECTURA HDFS
Fuente [13]

2.8.2.1.1. Características de HDFS.

- **Tolerancia a fallos:** Si se llegara a generar la caída de uno de los datanodes, el clúster sigue funcionando sin generar pérdida de información o retrasos
- **Acceso a datos en streaming:** no se necesitan descargar los datos.
- **Facilidad para el trabajo con grandes volúmenes de datos:** los clusteres almacenan gran cantidad de ficheros de todo tipo.
- **Modelo sencillo de coherencia:** No implementa la regla POSIX en un 100% para poder aumentar la tasa de transferencia de datos.

- **Portabilidad de convivencia entre hardware heterogéneo:** Se ejecuta en máquinas de diferentes fabricantes. [13]

2.8.2.2. MapReduce

Modelo de programación asociado con el procesamiento y generación de conjuntos de grandes volúmenes de información con un algoritmo que implementa estrategias de computo paralelo y distribuido sobre un clúster, a continuación, se define las tres fases. [12]

- **Map:** Se aplica en paralelo, se le hace mapeo a cada ítem de dato de entrada asignándole una lista de pares key/value. [13]**.**
- **Shuffle and sort**: Realiza dos actividades, la primera ordena por claves los resultados del mapeo, y la segunda, se encarga de recoger los valores intermedios de una clave para agruparlos. [13]**.**
- **Reduce:** se aplica en paralelo para cada grupo asociado a una clave. El resultado es la producción de una colección de valores para cada dominio, aunque también puede darse el caso de generar una llamada vacía. El resultado es la obtención de una lista de valores. [13]**.**

En la figura 4, se muestra el funcionamiento del procedimiento MapReduce. MapReduce tiene como objetivo principal particionar el archivo de entrada y distribuirlo en los diferentes Datanodes de la arquitectura. Primero, el Namenode particiona el archivo de entrada en archivos más pequeños, los cuales son distribuidos en cada uno de los Datanodes. Segundo, los Datanodes, respectivamente, realizan el mapeo del contenido de cada uno de los archivos. Es decir, cada palabra (Token) se le asigna un identificador, de manera tal que pueda ser identificado de forma única en fases posteriores del proceso (e.g., <id,”token”>). Tercero, en cada uno de los Datanodes

se procede a reducir los mapeos de manera tal que se trabajen con tokens simplificados. De esta forma se elimina la duplicidad de tokens en cada reducción. Finalmente, el Namenode recoge cada una de las salidas del proceso reduce y procede a contabilizar los resultados [14].

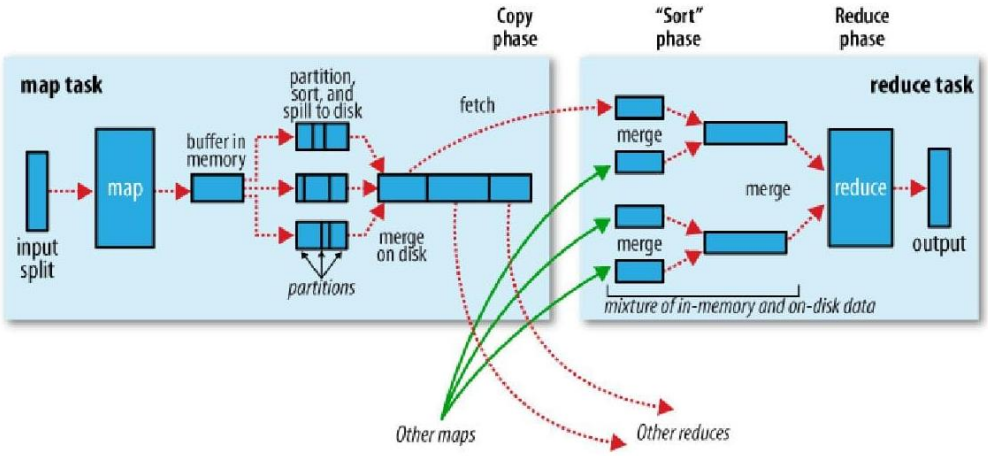


FIGURA 4. MAPREDUCE
FUENTE [13]

3. Desarrollo del proyecto

Se realizo el diseño del clúster en base a los equipos e implementos que se tenían disponibles en la institución, igualmente, se tuvo en cuentas los parámetros y normas básicas que se necesitan para la implementación del mismo.

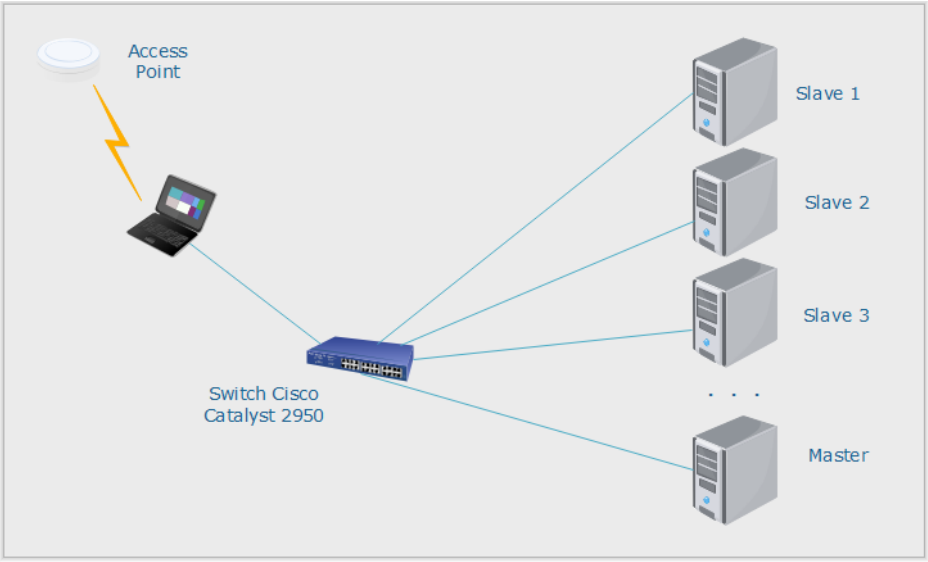


Figura 5. Diseño de arquitectura del clúster

Se implementó un modelo funcional del clúster a partir de cuatro equipos de cómputo dados de baja por la institución; además, se incorporaron para la interconexión un switch y un Access Point para gestionar el acceso a internet de los equipos. Para la configuración del clúster se implementó una arquitectura maestro-esclavo, etiquetando los computadores como "Master" y "Slave 1", "Slave 2" y "Slave 3". El Access Point recibe servicio de internet de la red interna de la institución y provee el servicio a los computadores para permitir la instalación, configuración y descarga de librerías para la puesta a punto del clúster. A cada uno de los equipos de cómputo se le asigna una IP estática clase C, se listan en la tabla 5.

Tabla 3
Característica de Equipos de Computo

EQUIPO	**HP INTEL CORE 2 DUO**
Memoria	1,9 Gib
Procesador	Intel Core 2 Duo CPU E4600
Gráficos	Intel Q33
Tipo de sistema operativo	64 bits
Disco	500 GB
Sistema Operativo	Ubuntu 17.04

Nota: se evidencia las características básicas de los equipos de cómputo como lo es memoria, procesador, disco y sistema operativo.

Tabla 4
Característica del Switch

EQUIPO	**CISCO SYSTEMS**	
Modelo	Catalyst 2950	
Subtipo	Fast Ethernet	
Puertos	Tipo: 10/100	Cantidad: 24
Norma	IEEE 802.1D, IEEE 802.1Q, IEEE 802.1P, IEEE 802.3, IEEE 802.3u	
Modo de comunicación	Full-duplex, Half-duplex	
Voltaje requerido	AC 110/220 V	
Frecuencia	50/60 Hz	

Nota: se especifican las características del switch, como marca del equipo, modelo, numero de puertos, modo de comunicación, voltaje requerido y frecuencia de transmisión.

Tabla 5
Direcciones IP

EQUIPO	**IP**	**USUARIO**
Master	192.168.137.100	Master
Slave 1	192.168.137.101	Slave1
Slave 2	192.168.137.102	Slave2
Slave 3	192.168.137.103	Slave3

Nota: Se listan las direcciones IP y usuario de cada uno de los equipos del clúster.

3.1. Experimentos

3.1.1. Descripción de los archivos

Los archivos utilizados para realizar la carga del sistema están disponibles en el sitio Web de GoogleBook[2] . En dicha página el lector tiene acceso a un conjunto variado de grandes archivos que contienen datos Ngram viewer.
Para la experimentación del clúster Hadoop se seleccionaron los siguientes volúmenes de datos como se muestra en la tabla 6.

Tabla 6.

Enlace de los volúmenes de datos de los experimentos.

Experimentos	**Tamaño del archivo**	**Enlace**
1	129.16 MB	http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-z.gz
2	220.5 MB	http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-6.gz

[2] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

3	577 MB	http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-2.gz
4	601.22 MB	http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-gb-all-1gram-20120701-1.gz
5	877.57 MB	http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-us-all-1gram-20120701-r.gz
6	923.35 MB	http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20090715-0.csv.zip
7	1.38 GB	http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-1.gz
8	6GB	Combinación de los archivos anteriores
9	12GB	Combinación y replicación de los archivos anteriores

La figura 6 muestra los volúmenes de datos de entrada y salida del clúster. Como se aprecia, el volumen de datos de salida de cada uno de los experimentos representa un porcentaje ínfimo del volumen de datos de entrada. Por ejemplo para el experimento #1 el volumen de salida representa el 1.48 % del volumen de entrada; para el experimento #2 el volumen de salida representa el 1,32 % del volumen de entrada; para el experimento #3 el volumen de salida representa el 1,36% del volumen de entrada; para el experimento #4 el volumen de salida representa el 1,20 % del volumen de entrada: para el experimento #5 el volumen de salida representa el 1,24 % del volumen de entrada; para el experimento #6 el volumen de salida representa el 1,02% del volumen de entrada; para el experimento #7 el volumen de salida representa el 1,4 % del volumen de entrada; para el experimento #8 el volumen de salida representa el 0,88 % del volumen de entrada; para el experimento #9 el volumen de salida representa el 0,45 % del volumen de entrada.

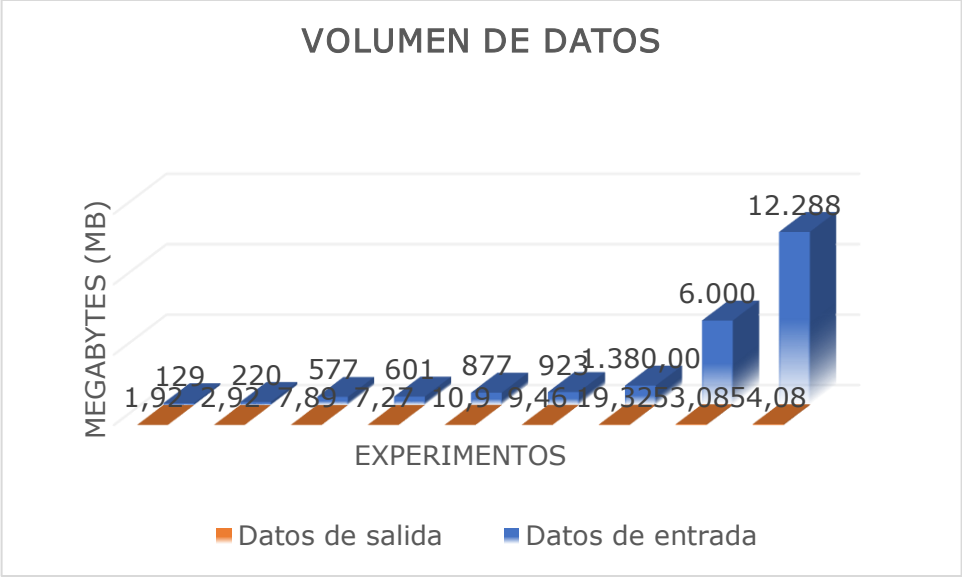


Figura 6. Tamaño de los volúmenes de datos de entrada y salida del clúster

3.1.2. Descripción de los experimentos.

Se realizaron nueve (9) experimentos con el objetivo de analizar el rendimiento del clúster comparado con el equipo unitario. Los experimentos se basaron en contar palabras (anagramas de uno y dos palabras) en volúmenes de datos de distintos tamaños (100 MB, 200 MB, 500 MB, 600 MB, 800 MB, 900 MB, 1 GB, 6 GB, 12 GB).

Se ejecutaron 10 veces cada uno de los experimentos tanto para el clúster como para el equipo unitario, donde se obtuvo el promedio de cada uno de estos, igualmente, se calculó la desviación estándar que nos permite saber que tan dispersos están los datos en relación con la media.

3.1.3. Resultados equipo unitario.

En esta sesión del documento se muestran los resultados de los experimentos realizados al equipo unitario.

3.1.3.1. Volumen de datos > 100 MB.

La tabla 7 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

Tabla 7

Tiempo de ejecución equipo unitario 100 MB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDO (S)
1	0m52,960s	52,960
2	0m52,777s	52,777
3	0m52,669s	52,669
4	0m51,637s	51,637
5	0m52,840s	52,840
6	0m53,775s	53,775
7	0m52,631s	52,631
8	0m51,694s	51,694
9	0m52,634s	52,634
10	0m55,706s	55,706
	Promedio	**52,932**
	Desviación Estándar	**1,1497**

La Figura 7 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del equipo unitario para un volumen de datos mayor a 100 MB. Como se puede apreciar, el tiempo promedio de ejecución es 52,9323 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 55,9323 segundos y de ejecución mínimo de 51,637 segundos con una desviación estándar respecto a la media de ejecución de 1,1497 segundos.

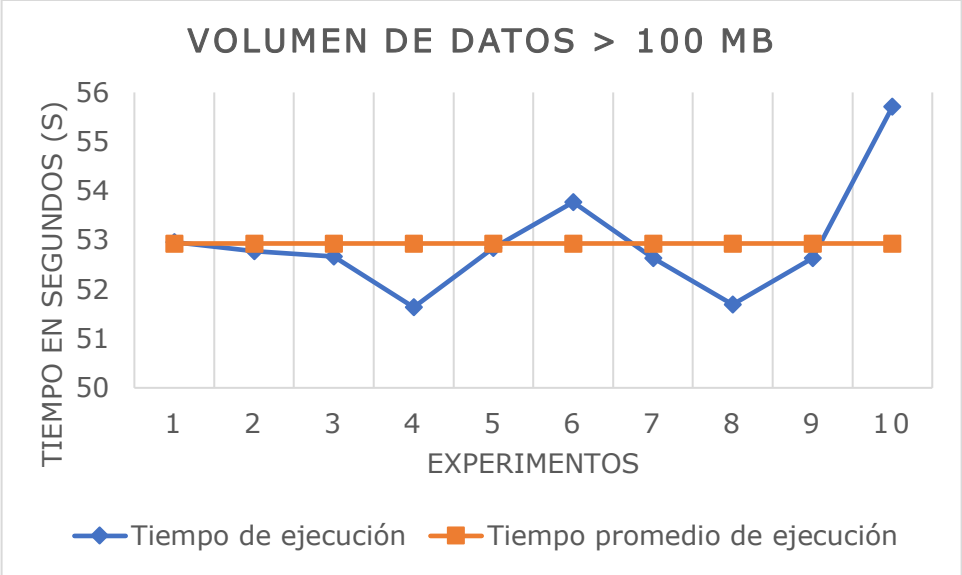


Figura 7. Tiempo de ejecución del equipo unitario para un volumen de datos > 100 MB

3.1.3.2. Volumen de datos > 200 MB.

La tabla 8 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

TABLA 8

Tiempo de ejecución equipo unitario 200 MB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDO (S)
1	1m31,870s	91,870
2	1m32,197s	92,197
3	1m33,944s	93,944
4	1m29,244s	89,244
5	1m31,775s	91,775
6	1m31,814s	91,870
7	1m30,185s	90,185
8	1m30,690s	90,690
9	1m32,776s	92,776
10	1m30,684s	90,684
	Promedio	**91,465**
	Desviación Estándar	**1,3090**

La Figura 8 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del equipo unitario para un volumen de datos mayor a 200 MB. Como se puede apreciar, el tiempo promedio de ejecución es 91,4656 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 92,776 segundos y de ejecución mínimo de 89,244 segundos con una desviación estándar respecto a la media de ejecución de 1,3090 segundos.

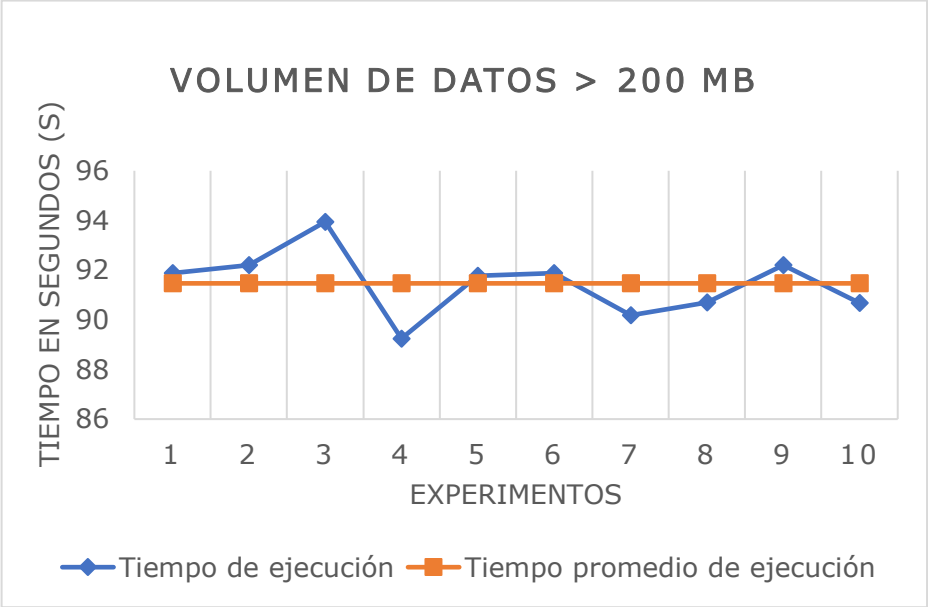


Figura 8. Tiempo de ejecución del equipo unitario para un volumen de datos > 200 MB

3.1.3.3. Volumen de datos > 500 MB.

La tabla 9 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

TABLA 9

Tiempo de ejecución equipo unitario 500 MB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDOS (S)
1	3m61,651s	241,651
2	3m47,326s	227,326
3	3m48,225s	228,225
4	3m46,863s	228,863

5	3m49,021s	229,021
6	3m50,180s	230,180
7	3m46,343s	226,343
8	3m42,874s	222,874
9	3m47,171s	227,171
10	3m54,540s	234,540
	Promedio	**229,619**
	Desviación Estándar	**5,16021**

La Figura 9 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del equipo unitario para un volumen de datos mayor a 500 MB. Como se puede apreciar, el tiempo promedio de ejecución es 229,619 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 229,021 segundos y de ejecución mínimo de 222,874 segundos con una desviación estándar respecto a la media de ejecución de 5,16021 segundos.

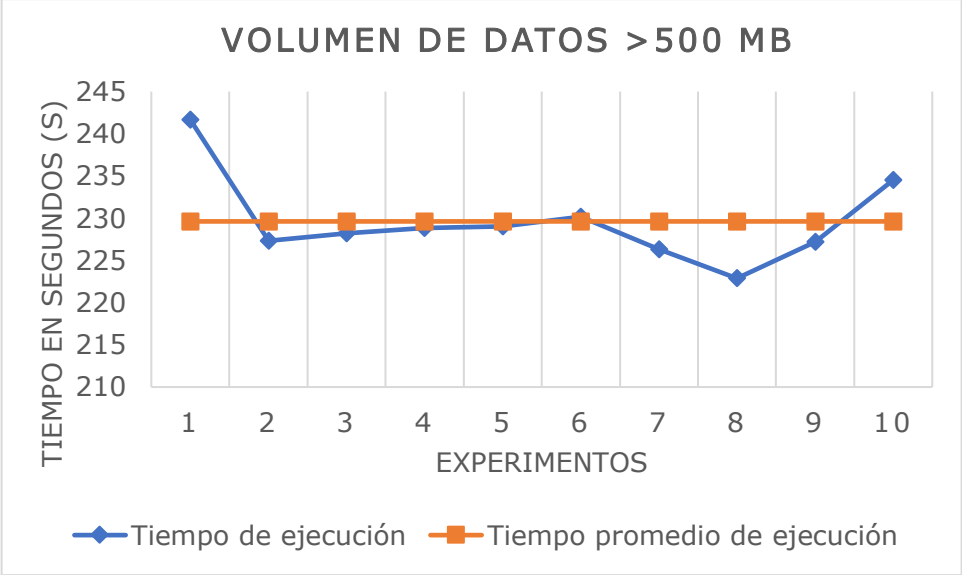


Figura 9.Tiempo de ejecución del equipo unitario para un volumen de datos > 500 MB

3.1.3.4. Volumen de datos > 600 MB.

La tabla 10 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

Tabla 10

Tiempo de ejecución equipo unitario 600 MB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDOS (S)
1	3m54,078s	234,078
2	3m58,733s	238,733
3	3m52,327s	232,327
4	3m54,495s	234,495
5	3m51,893s	231,893
6	4m0,251s	240,251
7	4m8,814s	248,814
8	3m51,984s	231,984
9	3m55,188s	235,188
10	3m54,132s	234,132
	promedio	**236,189**
	Desviación Estándar	**5,22785**

La Figura 10 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del equipo unitario para un volumen de datos mayor a 600 MB. Como se puede apreciar, el tiempo promedio de ejecución es 236,189 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 248,814 segundos y de ejecución mínimo de 231,132 segundos con una desviación estándar respecto a la media de ejecución de 5,22785 segundos.

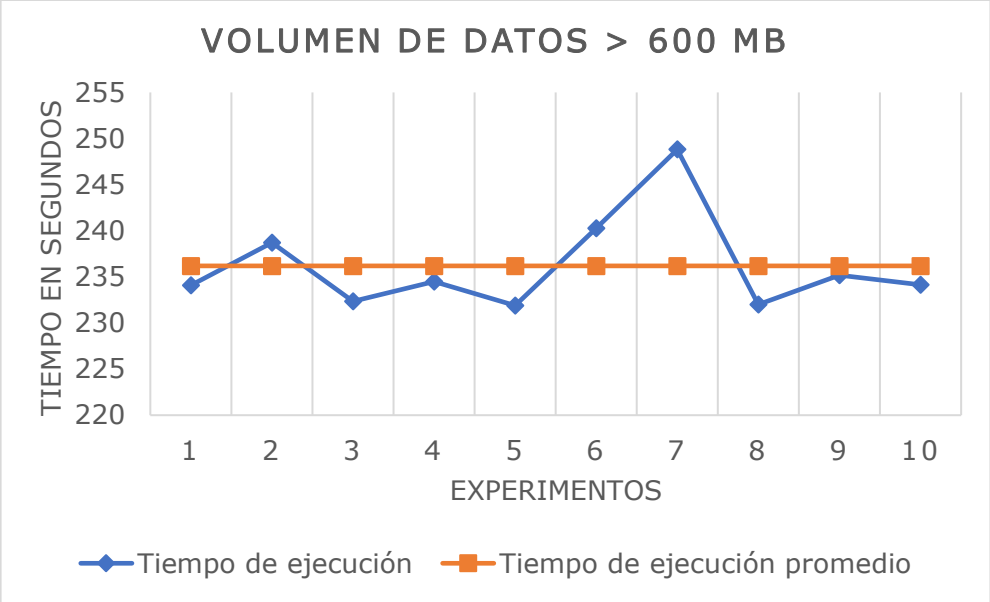


Figura 10. Tiempo de ejecución del equipo unitario para un volumen de datos > 600 MB

3.1.3.5. Volumen de datos > 800 MB.

La tabla 11 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

Tabla 11

Tiempo de ejecución equipo unitario 800 MB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDO
1	5m10,302s	310,302
2	5m17,647s	317,647
3	5m20,032s	320,032
4	5m11,635s	311,635
5	5m21,251s	321,251
6	5m21,717s	321,717
7	5m11,960s	311,960
8	5m28,613s	328,613
9	5m25,252s	325,252
10	5m22,759s	322,759
	Promedio	**319,116**
	Desviación Estándar	**6,14903**

La Figura 11 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del equipo unitario para un volumen de datos mayor a 800 MB. Como se puede apreciar, el tiempo promedio de ejecución es 319,116 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 328,613 segundos y de ejecución mínimo de 310,302 segundos con una desviación estándar respecto a la media de ejecución de 6,14903 segundos.

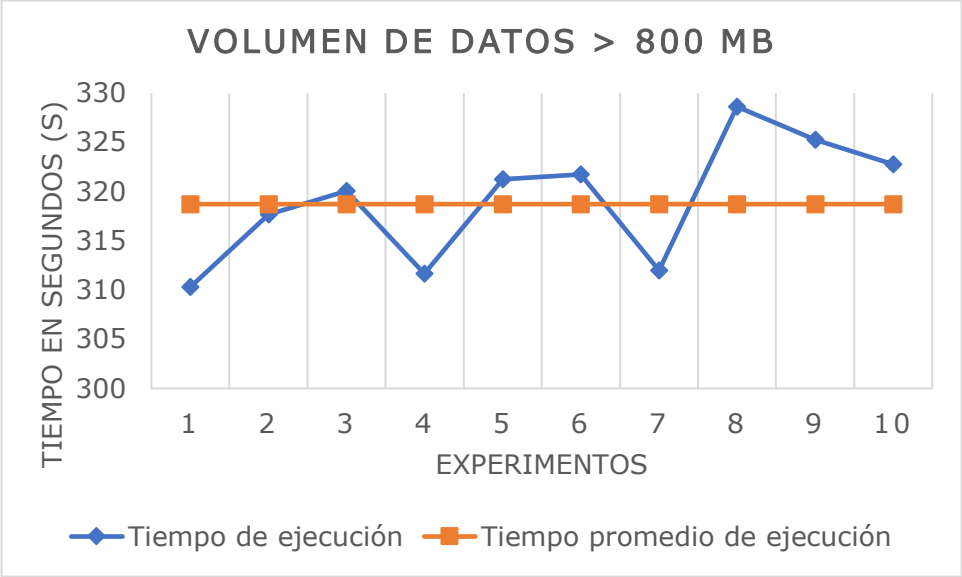


Figura 11. Tiempo de ejecución del equipo unitario para un volumen de datos > 800 MB

3.1.3.6. Volumen de datos > 900 MB.

La tabla 12 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

TABLA 12

Tiempo de ejecución equipo unitario 900 MB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDOS (S)
1	6m20,938s	380,938
2	6m27,927s	387,927
3	6m19,449s	379,449
4	6m15,900s	375,900
5	6m22,766s	382,766

6	6m27,930s	387,930
7	6m31,531s	391,531
8	6m18,704s	378,704
9	6m53,529s	413,529
10	6m21,407s	381,407
	Promedio	**386,008**
	Desviación Estándar	**10,8074**

La Figura 12 muestra el contraste del tiempo de ejecución vs. el tiempo de ejecución promedio del equipo unitario para un volumen de datos mayor a 900 MB. Como se puede apreciar, el tiempo promedio de ejecución es 3896,008 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 413,529 segundos y de ejecución mínimo de 375,900 segundos con una desviación estándar respecto a la media de ejecución de 10,8074 segundos.

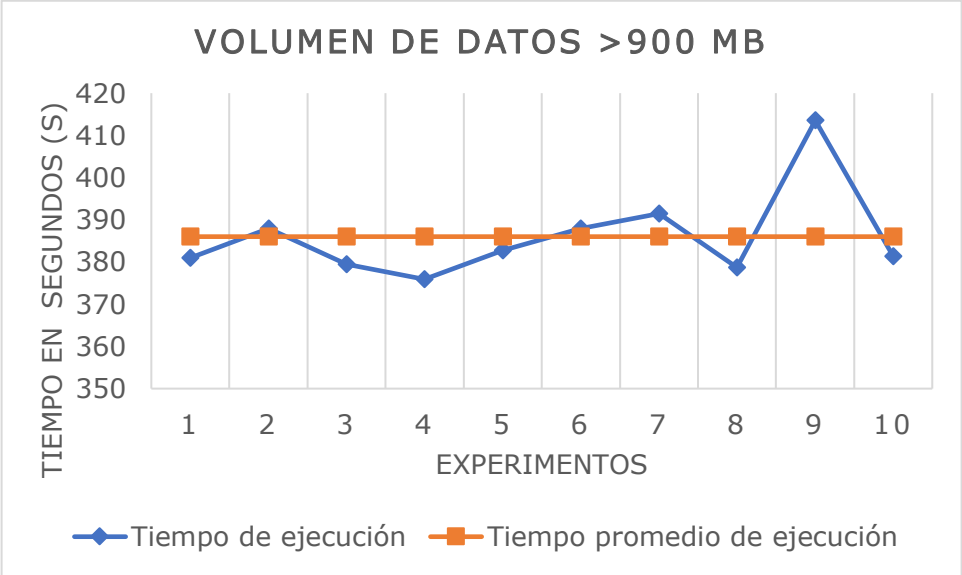


Figura 12. Tiempo de ejecución del equipo unitario para un volumen de datos > 900 MB

3.1.3.7. Volumen de datos > 1 GB.

La tabla 13 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

Tabla 13

Tiempo de ejecución equipo unitario 1 GB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDOS (S)
1	8m55,465s	535,465
2	8m55,898s	535,898
3	8m50,038s	530,038
4	8m58,435s	538,435
5	9m3,244s	543,244
6	8m59,136s	539,136
7	9m0,267s	540,267
8	9m2,266s	542,266
9	8m58,231s	538,231
10	9m6,081s	546,081
	Promedio	**538,906**
	Desviación Estándar	**4,51787**

La Figura 13 muestra el contraste del tiempo de ejecución vs. el tiempo de ejecución promedio del equipo unitario para un volumen de datos mayor a 100 MB. Como se puede apreciar, el tiempo promedio de ejecución es 52,9323 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 55,9323 segundos y

de ejecución mínimo de 530,038 segundos con una desviación estándar respecto a la media de ejecución de 4,51787 segundos.

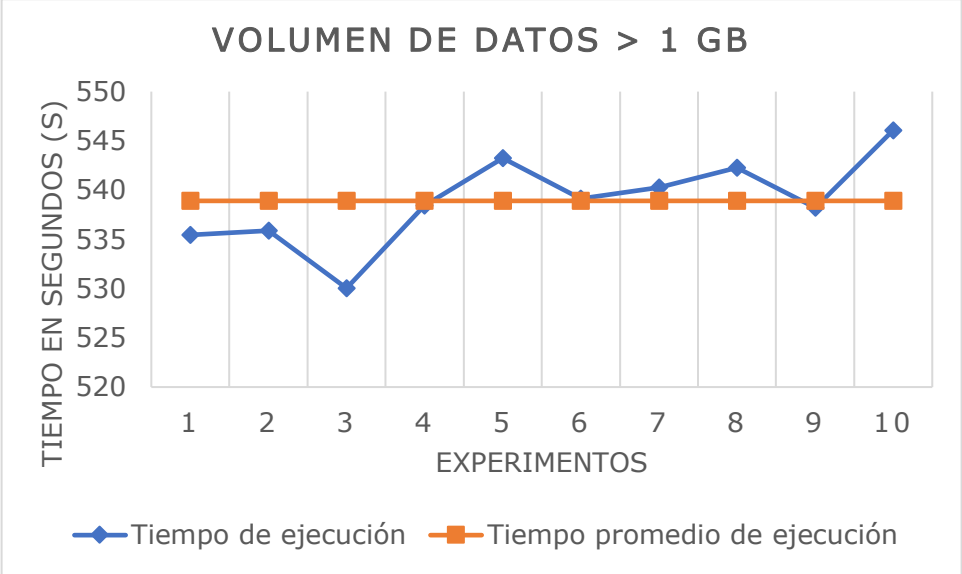


Figura 13. Tiempo de ejecución del equipo unitario para un volumen de datos > 1 GB

3.1.3.8. Volumen de datos > 6 GB.

La tabla 14 muestra el tiempo de computo de 10 ejecuciones realizadas en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar del mismo.

Tabla 14
Tiempo de ejecución equipo unitario 6 GB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDO
1	88m59,036s	5339,036
2	88m0,507s	5280,507
3	87m18,752s	5238,752
4	85m20,257s	5120,257
5	84m56,184s	5096,184
6	85m42,115s	5142,115
7	86m13,518s	5173,518
8	85m17,866s	5117,866
9	85m49,357s	5149,357
10	87m21,956s	5241,956
	Promedio	**5189,954**
	Desviación Estándar	**80,66717**

La Figura 14 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del equipo unitario para un volumen de datos mayor a 6 GB. Como se puede apreciar, el tiempo promedio de ejecución es 5189,954 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 5339,036 segundos y de ejecución mínimo de 5096,184 segundos con una desviación estándar respecto a la media de ejecución de 80,66717 segundos.

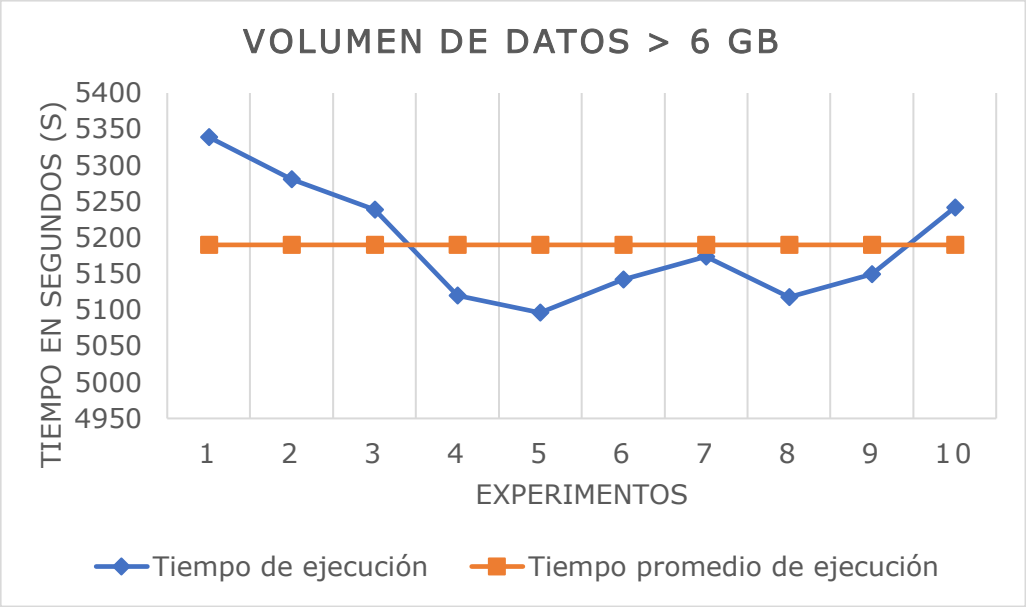


Figura 14. Tiempo de ejecución del equipo unitario para un volumen de datos > 6 GB

3.1.3.9. Volumen de datos > 12 GB.

La tabla 15 muestra el tiempo de computo de 10 experimentos ejecutados en el equipo unitario; de igual forma se presenta el promedio y la desviación estándar de los mismos

Tabla 15
Tiempo de ejecución equipo unitario 12 GB.

EXPERIMENTOS	EQUIPO UNITARIO	SEGUNDO
1	113m57,582s	6.837,58
2	113m28,777s	6.808,78
3	115m47,131s	6.947,13
4	113m7,771s	6.787,77

5	113m40,503s	6.820,50
6	114m31,748s	6.871,74
7	116m47,435s	7.007,43
8	115m6,559s	6.906,55
9	114m58,552s	6.898,55
10	117m54,067s	7.074,06
	Promedio	**6.896,01**
	Desviación Estándar	**91,8210**

La Figura 15 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del equipo unitario para un volumen de datos mayor a 12 GB. Como se puede apreciar, el tiempo promedio de ejecución es 6.896,01 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 7.074,06 segundos y de ejecución mínimo de 6.787,77 segundos con una desviación estándar respecto a la media de ejecución de 91,8210 segundos.

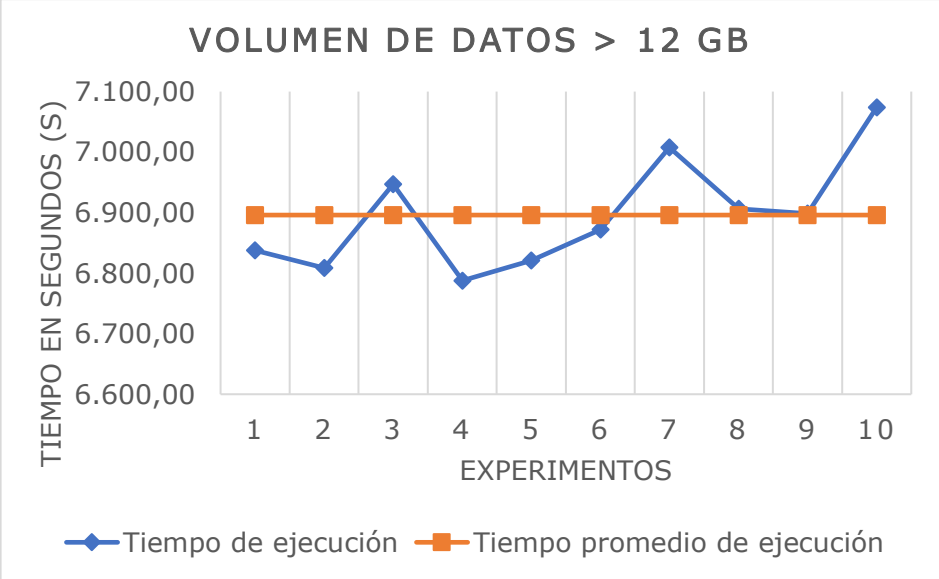


Figura 15. Tiempo de ejecución del equipo unitario para un volumen de datos > 12 GB

3.1.4. Resultados clúster

A continuación, se presentan los resultados obtenidos de la carga efectuada al clúster con cada uno de los volúmenes de datos.

3.1.4.1. Volumen de datos > 100 MB

La tabla 16 muestra el estado de almacenamiento para un volumen de datos mayor a 100 MB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 7 bloques del sistema de archivo de datos de Hadoop (HDFS) de 292.78 MB. Asimismo, el Slave1 emplea 5 bloques HDFS de 132.17 MB. Del mismo modo, el Slave2 usa 4 bloques HDFS de 290.82 MB. Finalmente, el Slave3 dispone de 5 bloques de 162.62 MB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 16
Utilización (Almacenamiento) del clúster Volumen de datos > 100 MB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MÁSTER	45.58 GB	292.78 MB	15.13 GB	7
SLAVE1	45.58 GB	132.17 MB	5.2 GB	5
SLAVE2	45.58 GB	290.82 MB	4.8 GB	4
SLAVE3	45.58 GB	162.62 MB	4.8 GB	5

La tabla 17 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 17
Tiempo de ejecución del clúster 100 MB

EXPERIMENTOS	CLÚSTER	SEGUNDO
1	1m20,456s	80,456
2	2m34,532s	154,532
3	1m30,044s	90,044
4	1m21,483s	81,483
5	1m20,371s	80,371
6	1m22,504	82,504
7	1m20,040s	80,040
8	1m20,364s	80,364
9	1m20,276s	80,273
10	2m24,491s	144,491
	Promedio	**95,4558**
	Desviación Estándar	**28,7425**

La Figura 16 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 100 MB. Como se puede apreciar, el tiempo promedio de ejecución es 95,4558 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 154,532 segundos y de ejecución mínimo de 80,040 segundos con una desviación estándar respecto a la media de ejecución de 28,7425 segundos.

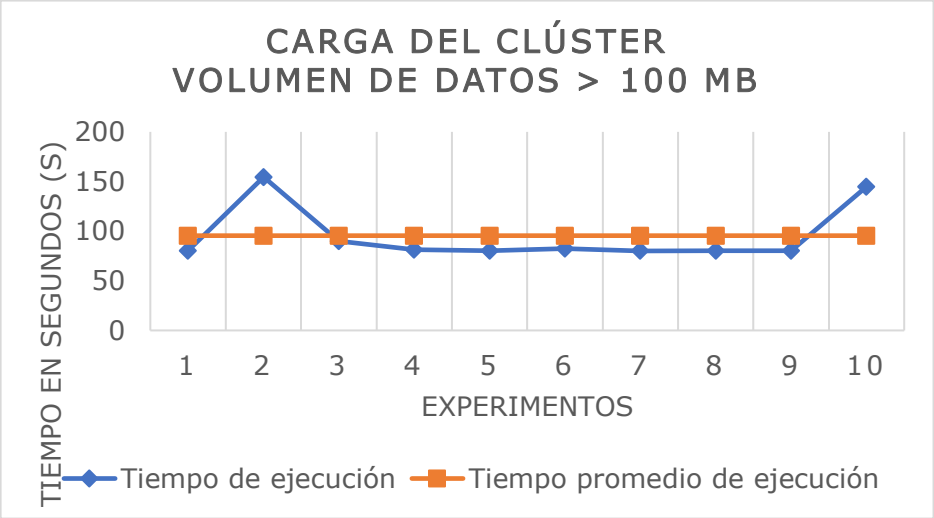


Figura 16. Tiempo de ejecución del Clúster para un volumen de datos > 100 MB

3.1.4.2. Volumen de datos > 200 MB

La tabla 18 muestra el estado de almacenamiento para un volumen de datos mayor a 200 MB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 7 bloques del sistema de archivo de datos de Hadoop (HDFS) de 385.86 MB. Asimismo, el Slave1 emplea 4 bloques HDFS de 96.24 MB. Del mismo modo, el Slave2 usa 4 bloques HDFS de 382.89 MB. Finalmente, el Slave3 dispone de 6 bloques de 292.63 MB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 18
Utilización (Almacenamiento) del clúster Volumen de datos > 200 MB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MÁSTER	45.58 GB	385.86 MB	15.13 GB	7
SLAVE1	45.58 GB	96.24 MB	5.2 GB	4
SLAVE2	45.58 GB	382.89 MB	4.8 GB	4
SLAVE3	45.58 GB	292.63 MB	4.81 GB	6

La tabla 19 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 19
Tiempo de ejecución 200 MB

EXPERIMENTO	CLUSTER	SEGUNDOS
1	1m27,638s	87,638
2	1m24,101s	84,101
3	1m19,222s	79,222
4	1m25,552s	85,552
5	1m23,406s	83,406
6	1m16,817s	76,817
7	1m18,244s	78,244
8	1m15,541s	75,541
9	1m22,423s	82,423
10	1m19,878s	79,878
	Promedio	**81,282**
	Desviación Estándar	**3,9543**

La Figura 17 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 200 MB. Como se puede apreciar, el tiempo promedio de ejecución es 81,2822 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 87,638 segundos y

de ejecución mínimo de 75,541 segundos con una desviación estándar respecto a la media de ejecución de 3,9543 segundos.

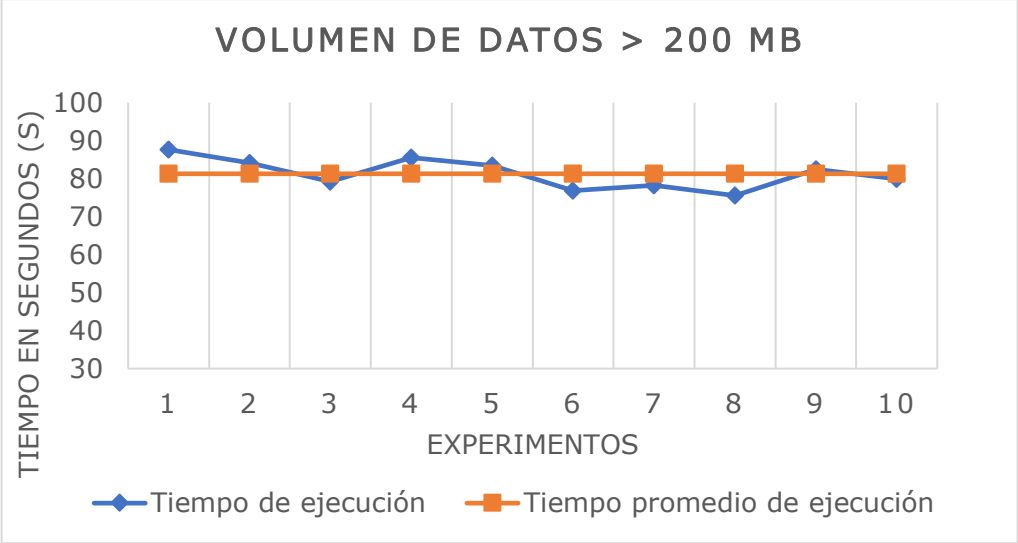


Figura 17. Tiempo de ejecución del Clúster para un volumen de datos > 200 MB

3.1.4.3. Volumen de datos > 500 MB

La tabla 20 muestra el estado de almacenamiento para un volumen de datos mayor a 500 MB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 10 bloques del sistema de archivo de datos de Hadoop (HDFS) de 750.2 MB. Asimismo, el Slave1 emplea 5 bloques HDFS de 323.58 MB. Del mismo modo, el Slave2 usa 8 bloques HDFS de 750.13 MB. Finalmente, el Slave3 dispone de 7 bloques de 426.63 MB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66 %, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 20
Utilización (almacenamiento) del clúster Volumen de datos > 500 MB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MASTER	45.58 GB	750.2 MB	15.13 GB	10
SLAVE1	45.58 GB	323.58 MB	5.2 GB	5
SLAVE2	45.58 GB	750.13 MB	4.8 GB	8

SLAVE3	45.58 GB	426.63 MB	4.81 GB	7

La tabla 21 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 21
Tiempo de ejecución 500 MB

EXPERIMENTO	CLÚSTER	SEGUNDOS
1	2m57,491s	152,539
2	2m32,529s	152,529
3	3m19,326s	199,326
4	2m55,927s	175,927
5	2m21,001s	141,001
6	3m1,576s	181,576
7	3m13,907s	193,907
8	2m19,758s	139,758
9	2m20,073s	140,073
10	2m19,537s	139,537
	Promedio	**161,617**
	Desviación estándar	**23,7750**

La Figura 17 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 500 MB. Como se puede apreciar, el tiempo promedio de ejecución es 161,6173 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 199,326 segundos y de ejecución mínimo de 139,537 segundos con una desviación estándar respecto a la media de ejecución de 23,7750 segundos.

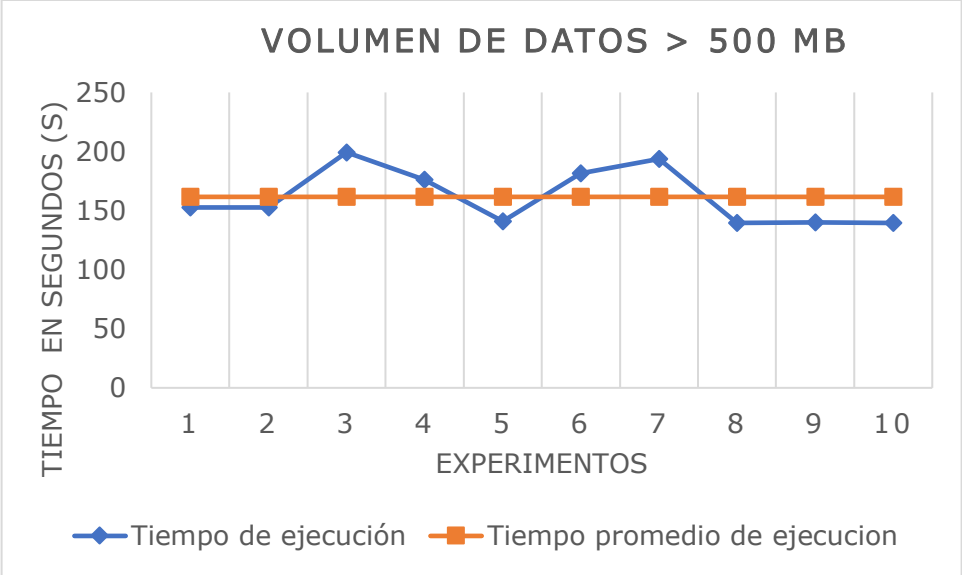


Figura 18. Tiempo de ejecución del Clúster para un volumen de datos > 500 MB

3.1.4.4. Volumen de datos > 600 MB

La tabla 22 muestra el estado de almacenamiento para un volumen de datos mayor a 600 MB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 10 bloques del sistema de archivo de datos de Hadoop (HDFS) de 773.94 MB. Asimismo, el Slave1 emplea 8 bloques HDFS de 613.32 MB. Del mismo modo, el Slave2 usa 4 bloques HDFS de 296.99 MB. Finalmente, el Slave3 dispone de 8 bloques de 637.61 MB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 22
Utilización (almacenamiento) del clúster Volumen de datos > 600 MB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MASTER	45.58 GB	773.94 MB	15.13 MB	10
SLAVE1	45.58 GB	613.32 MB	5.2 GB	8
SLAVE2	45.58 GB	296.99 MB	4.8 GB	4
SLAVE3	45.58 GB	637.61 MB	4.81 GB	8

La tabla 23 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 23
Tiempo de ejecución del clúster 600 MB

EXPERIMENTO	CLUSTER	SEGUNDOS
1	2m24,592s	144,592
2	2m22,043s	142,043
3	3m4,635s	184,635
4	4m0,153s	240,153
5	2m19,607s	139,607
6	3m21,498s	201,498
7	3m0,407s	180,407
8	3m13,724s	193,724
9	3m16,829s	196,829
10	3m52,544s	232,544
	Promedio	**185,603**
	Desviación estándar	**35,5070**

La Figura 19 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 600 MB. Como se puede apreciar, el tiempo promedio de ejecución es 185,603 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 240,153segundos y

de ejecución mínimo de 139,607 segundos con una desviación estándar respecto a la media de ejecución de 35,5070 segundos.

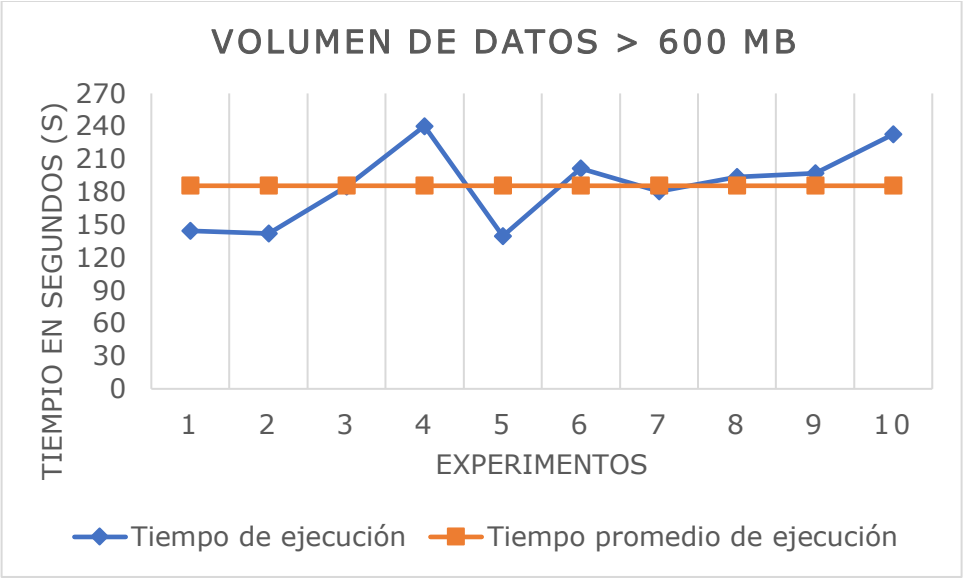


Figura 19. Tiempo de ejecución del Clúster para un volumen de datos > 600 MB

3.1.4.5. Volumen de datos > 800 MB

La tabla 24 muestra el estado de almacenamiento para un volumen de datos mayor a 800 MB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 19 bloques del sistema de archivo de datos de Hadoop (HDFS) de 1.9 GB. Asimismo, el Slave1 emplea 9 bloques HDFS de 747.94 MB. Del mismo modo, el Slave2 usa 15 bloques HDFS de 1.77 GB. Finalmente, el Slave3 dispone de 14 bloques de 1.3 GB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 24
Utilización (almacenamiento) del clúster Volumen de datos > 800 MB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MASTER	45.58 GB	1.9 GB	15.05 GB	19

SLAVE1	45.58 GB	747.94 MB	5.21 GB	9
SLAVE2	45.58 GB	1.77 GB	4.81 GB	15
SLAVE3	45.58 GB	1.3 GB	4.82 GB	14

La tabla 25 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 25
Tiempo de ejecución del clúster 800 MB

EXPERIMENTO	CLÚSTER	SEGUNDOS
1	3m49,588s	229,588
2	3m2,948s	182,948
3	3m3n200s	183,200
4	3m59,740s	239,740
5	3m6,791s	186,791
6	5m58,454s	358,454
7	3m52,616s	232,616
8	4m40,847s	280,847
9	4m41,126s	281,126
10	3m2,923s	182,923
	Promedio	**235,823**
	Desviación estándar	**57,6867**

La Figura 19 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 800 MB. Como se puede apreciar, el tiempo promedio de ejecución es 23,5823 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 358,454 segundos y de ejecución mínimo de 182,923 segundos con una desviación estándar respecto a la media de ejecución de 57,6867 segundos.

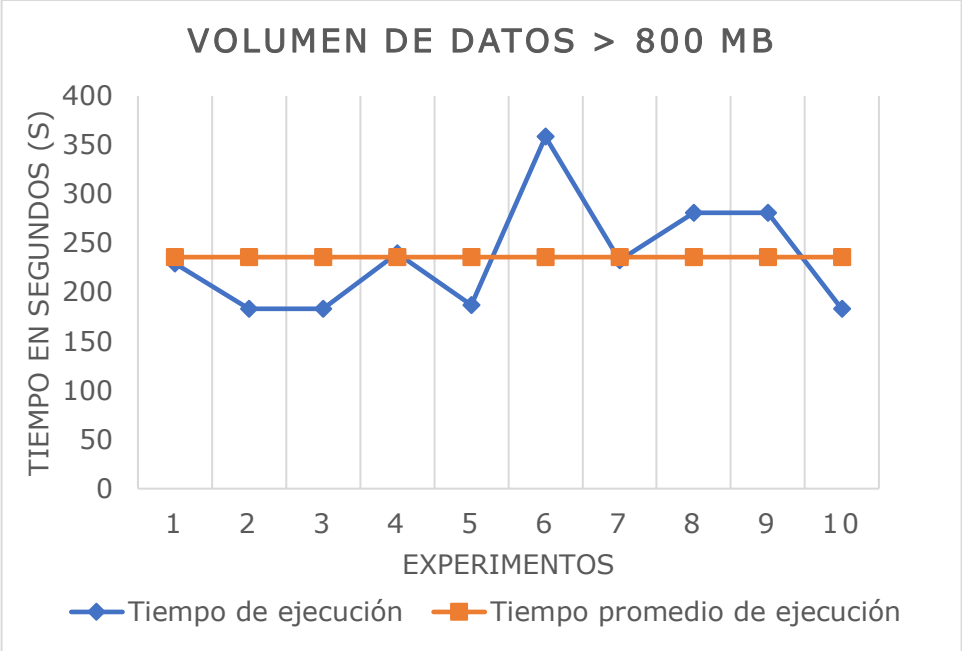


Figura 20. Tiempo de ejecución del Clúster para un volumen de datos > 800 MB

3.1.4.6. Volumen de datos > 900 MB

La tabla 26 muestra el estado de almacenamiento para un volumen de datos mayor a 900 MB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 13 bloques del sistema de archivo de datos de Hadoop (HDFS) de 1.08 GB. Asimismo, el Slave1 emplea 8 bloques HDFS de 654.61 MB. Del mismo modo, el Slave2 usa 6 bloques HDFS de 575.23 MB. Finalmente, el Slave3 dispone de 12 bloques de 971.81 MB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 26
Utilización (almacenamiento) del clúster Volumen de datos > 900 MB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MASTER	45.58 GB	1.08 GB	15.13 GB	13
SLAVE1	45.58 GB	654.61 MB	5.2 GB	8
SLAVE2	45.58 GB	575.23 MB	4.8 GB	6
SLAVE3	45.58 GB	971.81 MB	4.81 GB	12

La tabla 27 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 27
Tiempo de ejecución del clúster > 900 MB

EXPERIMENTOS	CLÚSTER	SEGUNDOS
1	4m32,307s	272,307
2	3m55,268s	235,268
3	3m43,012s	223,012
4	3m42,834s	222,834
5	3m23,219s	203,219
6	3m24,592s	204,592
7	3m31,718s	211,718
8	3m21,745s	201,745
9	3m35,977s	215,977
10	3m27,071s	207,071
	Promedio	**219,774**
	Desviación estándar	**21,3174**

La Figura 21 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 900 MB. Como se puede apreciar, el tiempo promedio de ejecución es 219,774 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 272,307 segundos y de ejecución mínimo de 201,745 segundos con una desviación estándar respecto a la media de ejecución de 21,3174 segundos.

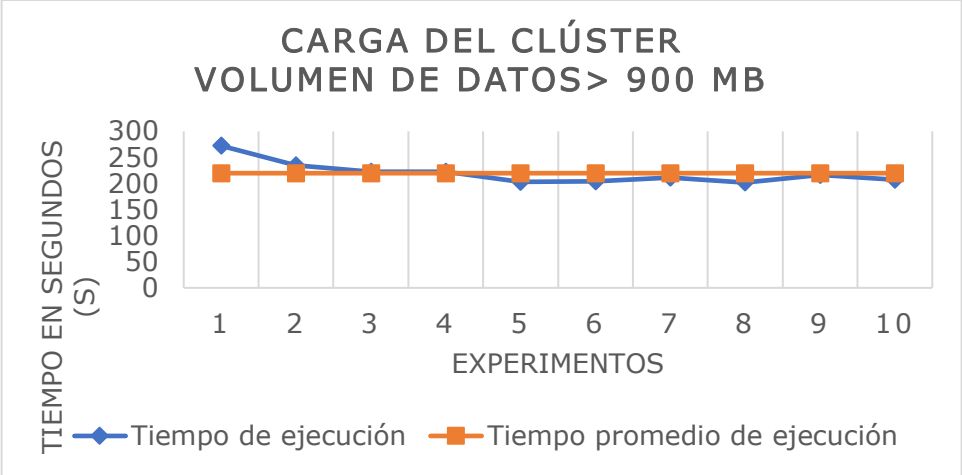


Figura 21. Tiempo de ejecución del Clúster para un volumen de datos > 900 MB

3.1.4.7. Volumen de datos > 1 GB

La tabla 28 muestra el estado de almacenamiento para un volumen de datos mayor a 1 GB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 17 bloques del sistema de archivo de datos de Hadoop (HDFS) de 1.56 GB. Asimismo, el Slave1 emplea 9 bloques HDFS de 777.25 MB. Del mismo modo, el Slave2 usa 10 bloques HDFS de 1.05 GB. Finalmente, el Slave3 dispone de 15 bloques de 1.31 GB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 28
Utilización (almacenamiento) del clúster Volumen de datos > 1 GB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MASTER	45.59 GB	1.56 GB	15.13 GB	17
SLAVE1	45.59 GB	777.25 MB	5.2 GB	9
SLAVE2	45.59 GB	1.06 GB	4.8 GB	10
SLAVE3	45.59 GB	1.31 GB	4.81 GB	15

La tabla 29 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 29
Tiempo de ejecución del clúster > 1 GB

EXPERIMENTO	CLÚSTER	SEGUNDOS
1	5m59,533s	359,533
2	5m9,000s	309,000
3	6m35,711s	395,711
4	5m59,177s	359,177
5	5m53,289s	353,289
6	6m53,895s	413,895
7	7m50,485s	470,485
8	6m54,336s	414,336
9	4m11,475s	251,475
10	6m48,060s	408,060
	Promedio	**373,496**
	Desviación estándar	**61,6911**

La Figura 21 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 1 GB. Como se puede apreciar, el tiempo promedio de ejecución es 373,496 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 470,485 segundos y de ejecución mínimo de 251,475 segundos con una desviación estándar respecto a la media de ejecución de 61,6911 segundos.

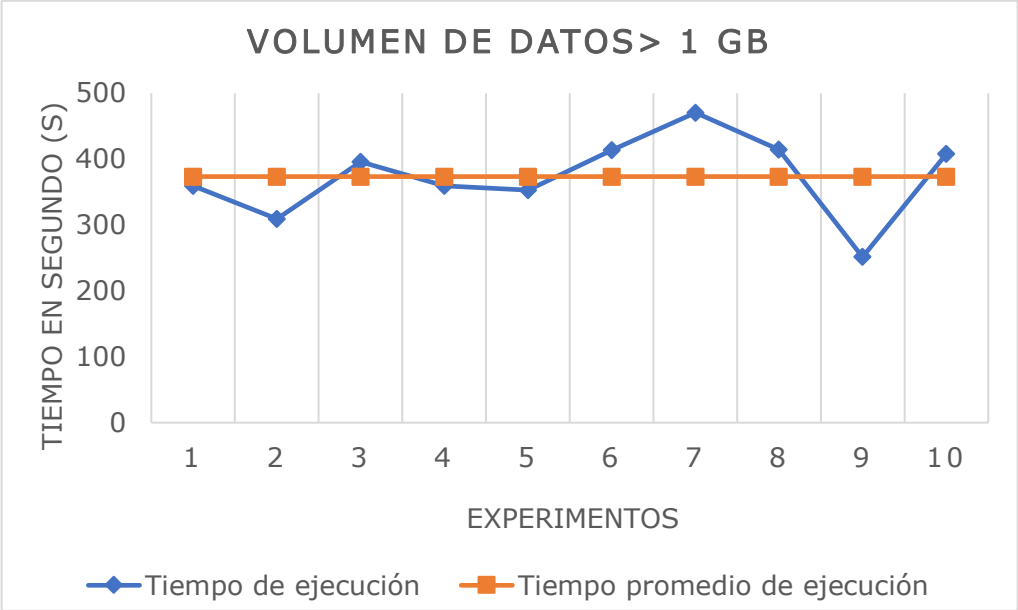


Figura 22. Tiempo de ejecución del Clúster para un volumen de datos > 1 GB

3.1.4.8. Volumen de datos > 6 GB

La tabla 30 muestra el estado de almacenamiento para un volumen de datos mayor a 6 GB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 284 bloques del sistema de archivo de datos de Hadoop (HDFS) de 6.07 GB. Asimismo, el Slave1 emplea 268 bloques HDFS de 3.99 GB. Del mismo modo, el Slave2 usa 248 bloques HDFS de 4.13 GB. Finalmente, el Slave3 dispone de 255 bloques de 4.2 GB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 48%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 87%, respectivamente.

Tabla 30
Utilización (almacenamiento) del clúster Volumen de datos > 6 GB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MASTER	45.58 GB	6.07 GB	23.34 GB	284
SLAVE1	45.58 GB	3.99 GB	5.25 GB	268
SLAVE2	45.58 GB	4.13 GB	5.48 GB	248
SLAVE3	45.58 GB	4.2 GB	5.75 GB	255

La tabla 31 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 31
Tiempo de ejecución del clúster > 6 GB

EXPERIMENTOS	CLÚSTER	SEGUNDOS
1	79m22,510s	4.762,51
2	76m18,160s	4.578,16
3	78m39,817s	4.719,82
4	77m12,858s	4.632,86
5	77m20,969s	4.640,97
6	78m54,915s	4.734,92
7	76m19,478s	4.579,48
8	95m48,899s	5.748,90
9	77m15,418s	4.635,42
10	74m57,832s	4.497,83
	Promedio	**4.753,09**
	Desviación estándar	**359,0007**

La Figura 23 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 6 GB. Como se puede apreciar, el tiempo promedio de ejecución es 4.753,09 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 5.748,90 segundos y de ejecución mínimo de 4.497,83 segundos con una desviación estándar respecto a la media de ejecución de 359,0007 segundos.

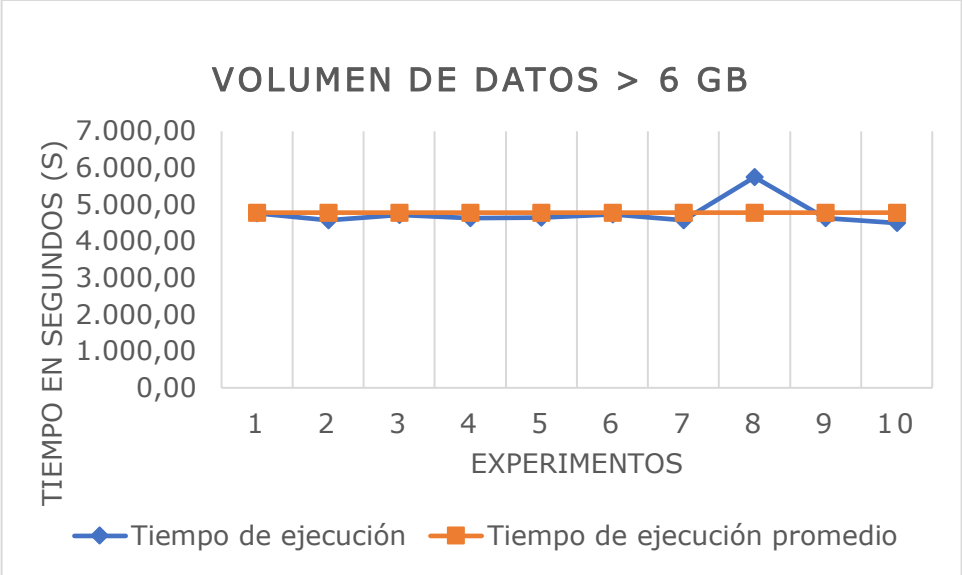


Figura 23. Tiempo de ejecución del Clúster para un volumen de datos > 6 GB

3.1.4.9. Volumen de datos > 12 GB

La tabla 32 muestra el estado de almacenamiento para un volumen de datos mayor a 12 GB. Cabe resaltar que cada nodo tiene una capacidad de almacenamiento de 45.58 GB; para un total de 182,32 GB de almacenamiento global del clúster. De igual forma se identifica, en el clúster Hadoop, que el nodo Master utiliza 329 bloques del sistema de archivo de datos de Hadoop (HDFS) de 10.79 GB. Asimismo, el Slave1 emplea 295 bloques HDFS de 6.86 GB. Del mismo modo, el Slave2 usa 281 bloques HDFS de 7.61B. Finalmente, el Slave3 dispone de 285 bloques de 7.14 GB. Es importante resaltar que el espacio disponible de HDFS sin utilizar en el nodo Master es del 66%, del nodo Slave1 es 88%, y de los nodos Salve2 y Slave3 es 89%, respectivamente.

Tabla 32
Utilización (almacenamiento) del clúster Volumen de datos > 12 GB

NODOS	CAPACIDAD	HDFS	LINUX FILE SYSTEM	BLOQUES
MASTER	45.58 GB	10.79 GB	23.39 GB	329
SLAVE1	45.58 GB	6.86 GB	5,25 GB	295
SLAVE2	45.58 GB	7.61 GB	5.48 GB	281
SLAVE3	45.58 GB	7.14 GB	5.81 GB	285

La tabla 33 muestra el tiempo de computo de 10 experimentos ejecutados en el clúster; de igual forma se presenta el promedio y la desviación estándar de los mismos.

Tabla 33
Tiempo de ejecución del clúster > 12 GB

EXPERIMENTOS	CLUSTER	SEGUNDOS
1	59m53,510s	3.593,51
2	77m53,794s	4.673,79
3	58m59,770s	3.539,77
4	60m5,269s	3.605,27
5	59m26,296s	3.566,30
6	59m37,751	3.577,75
7	59m28,940	3.568,94
8	96m27,353s	5.787,35
9	98m23,127s	5.903,13
10	99m32,436s	5.972,44
	Promedio	**4.378,82**
	Desviación estándar	**1039,760**

La Figura 23 muestra el contraste del tiempo de ejecución vs. el tiempo promedio de ejecución del clúster para un volumen de datos mayor a 6 GB. Como se puede apreciar, el tiempo promedio de ejecución es 4.753,09 segundos; de igual forma se muestran los diferentes tiempos de ejecución (10 en total), donde se observa un tiempo de ejecución máximo de 5.748,90 segundos y de ejecución mínimo de 4.497,83 segundos con una desviación estándar respecto a la media de ejecución de 359,0007 segundos.

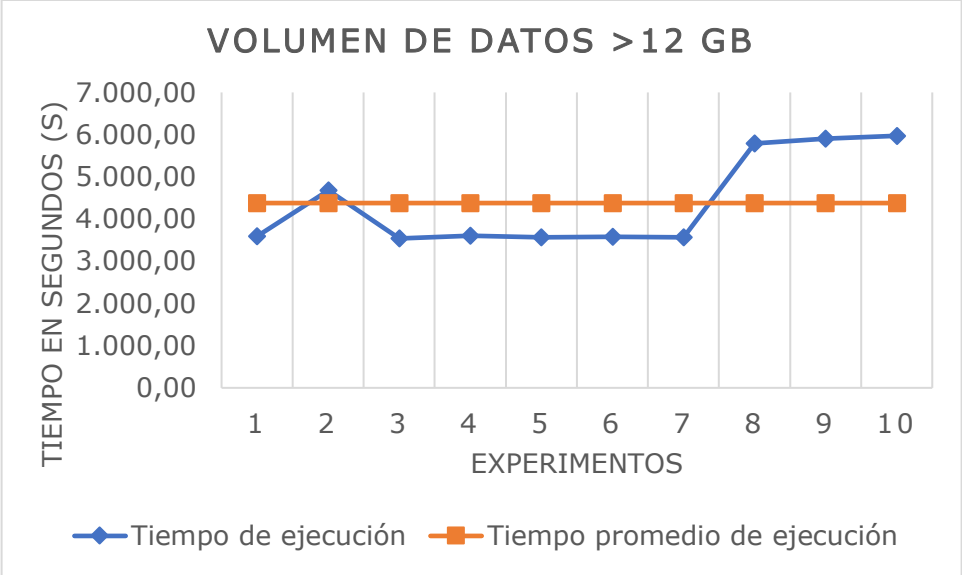


Figura 24. Tiempo de ejecución del Clúster para un volumen de datos > 12 GB

3.1.5. Comparación resultados (eficiencia)

Para comparar el rendimiento tanto del clúster como del equipo unitario se propone la ecuación Eq. 1

$$Gain(\overline{T}_s, \overline{T}_c) = \begin{cases} \frac{|\overline{T}_s - \overline{T}_c|}{\overline{T}_c}, si\ \overline{T}_c > \overline{T}_s \\ \frac{|\overline{T}_s - \overline{T}_c|}{\overline{T}_s}, si\ \overline{T}_s > \overline{T}_c \\ 0, si\ \overline{T}_s = \overline{T}_c \end{cases} \quad (1)$$

Donde $\overline{T}_c$ y $\overline{T}_s$ representan el tiempo promedio de ejecución del clúster y el equipo unitario, respectivamente. Cuando $\overline{T}_c > \overline{T}_s$,entonces el equipo unitario es de mejor rendimiento, cuando $\overline{T}_s > \overline{T}_c$, entonces el clúster es de mejor rendimiento y cuando $\overline{T}_s = \overline{T}_c$ no hay diferencia de rendimiento.

A continuación, se presenta la comparación del rendimiento del clúster y el equipo unitario en la gestión de grandes volúmenes de datos.

3.1.5.1. Volumen de datos > 100 MB

La Figura 25 muestra la comparación de rendimiento entre el clúster y el equipo unitario, para volúmenes de datos mayores a 100 MB. Luego de realizados los 10 experimentos, el equipo unitario exhibe un rendimiento del 44,54 % sobre el clúster, es decir, aplicando la Eq. 1 se observa que $\frac{|95{,}74-52{,}93|}{95{,}74} = 0{,}4454$[3].

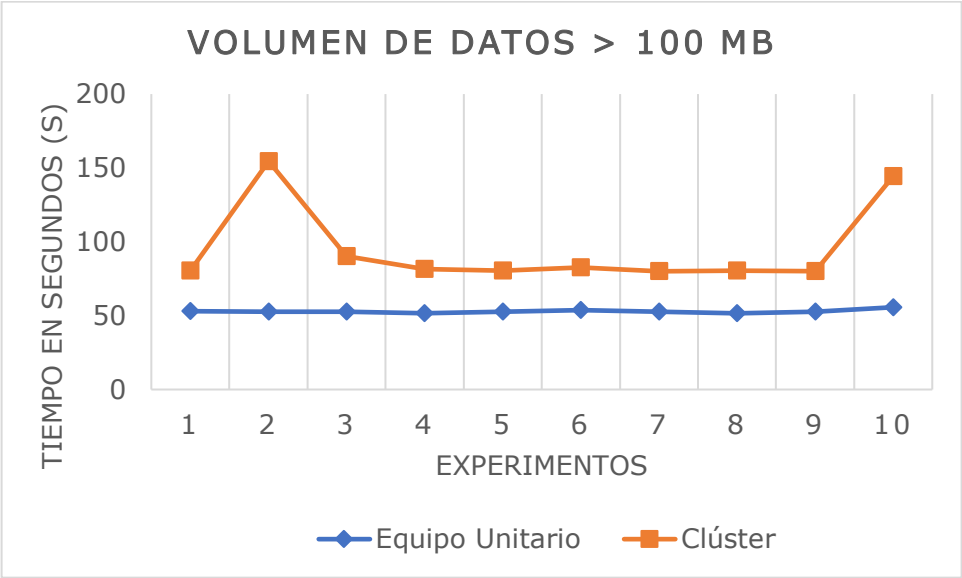


Figura 25. Comparación de eficiencia para un volumen de datos > 100 MB

3.1.5.2. Volumen de datos > 200 MB

La Figura 26 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 11,13% sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|81{,}28-91{,}46|}{91{,}46} = 0{,}1113$[4].

[3] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 17 y 7, respectivamente.
[4] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 19 y 8, respectivamente.

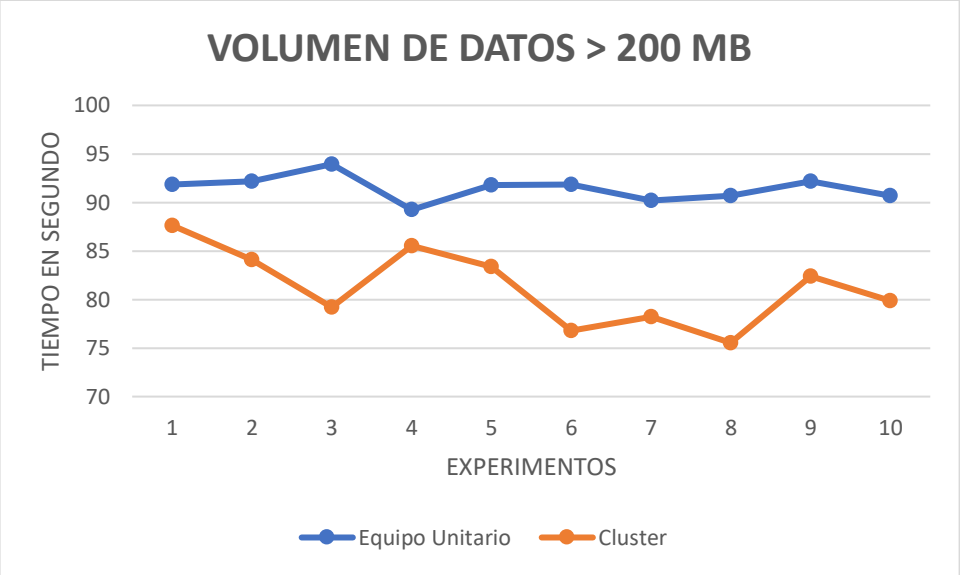


Figura 26. Comparación de eficiencia para un volumen de datos > 200 MB

3.1.5.3. Volumen de datos > 500 MB

La Figura 27 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 29,61 % sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|161,61-229,61|}{229,61} = 0,2961$[5].

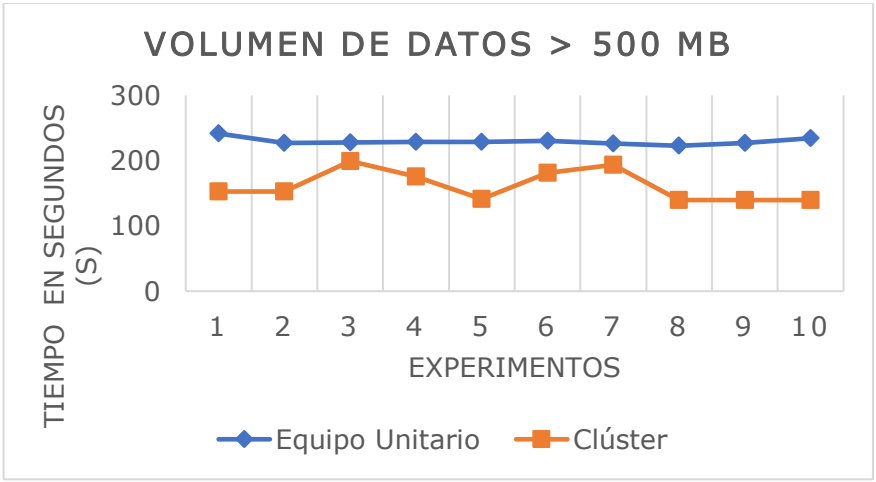


Figura 27. Comparación de eficiencia para un volumen de datos > 500 MB

[5] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 21 y 9, respectivamente.

3.1.5.4. Volumen de datos > 600 MB

La Figura 28 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 21,41 % sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|185{,}60-236{,}18|}{236{,}18} = 0{,}2141$[6].

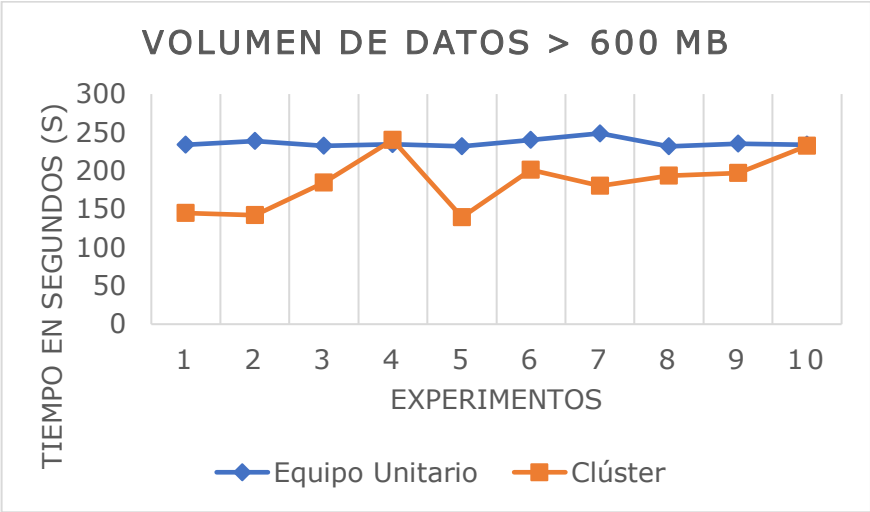


Figura 28. Comparación de eficiencia para un volumen de datos > 600 MB

3.1.5.5. Volumen de datos > 800 MB

La Figura 29 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 26,10 % sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|235{,}82-319{,}11|}{319{,}11} = 0{,}2610$[7].

[6] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 23 y 10, respectivamente.
[7] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 25 y 11, respectivamente.

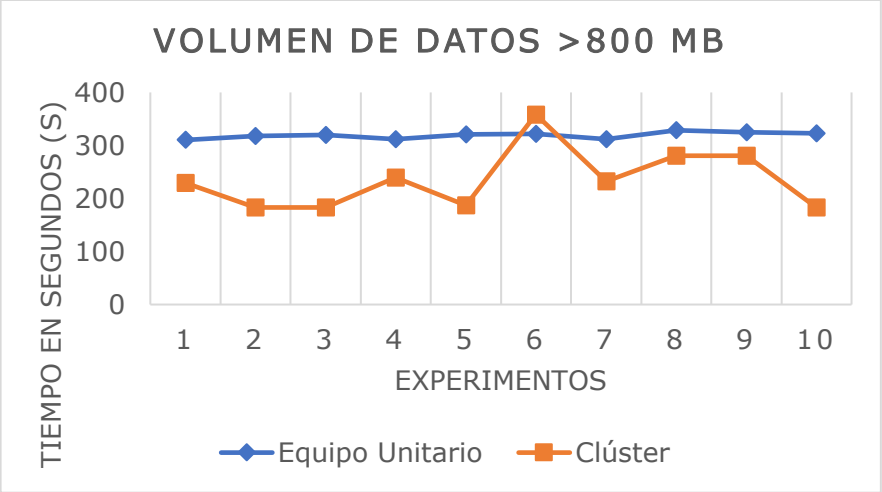


Figura 29. Comparación de la eficiencia para un volumen de datos > 800 MB

3.1.5.6. Volumen de datos > 900 MB

La Figura 22 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 43,06 % sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|219{,}77-386{,}00|}{386{,}00} = 0{,}4306$[8].

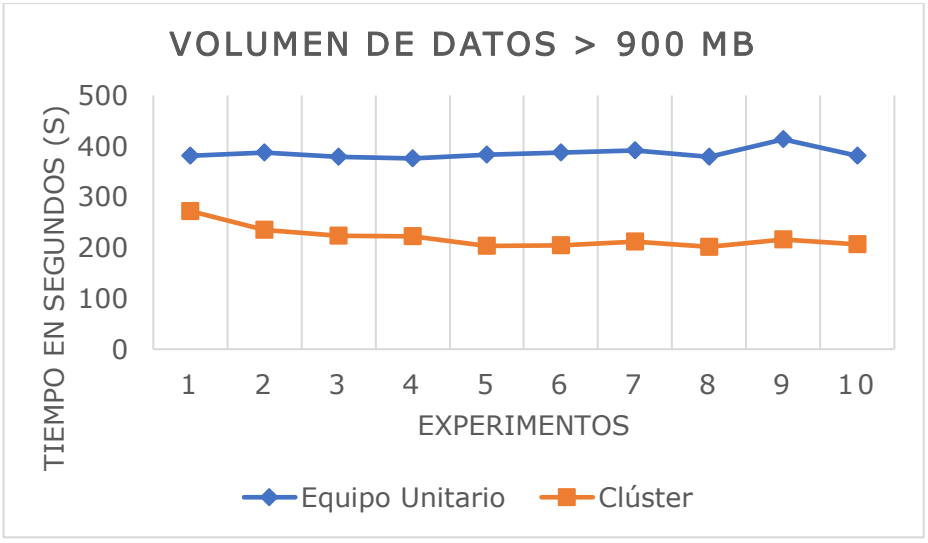


Figura 30. Comparación de eficiencia para un volumen de datos > 900 MB

3.1.5.7. Volumen de datos > 1 GB

[8] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 27 y 12, respectivamente.

La Figura 31 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 30,69 % sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|373{,}49-538{,}90|}{538{,}90} = 0{,}3069$ [9].

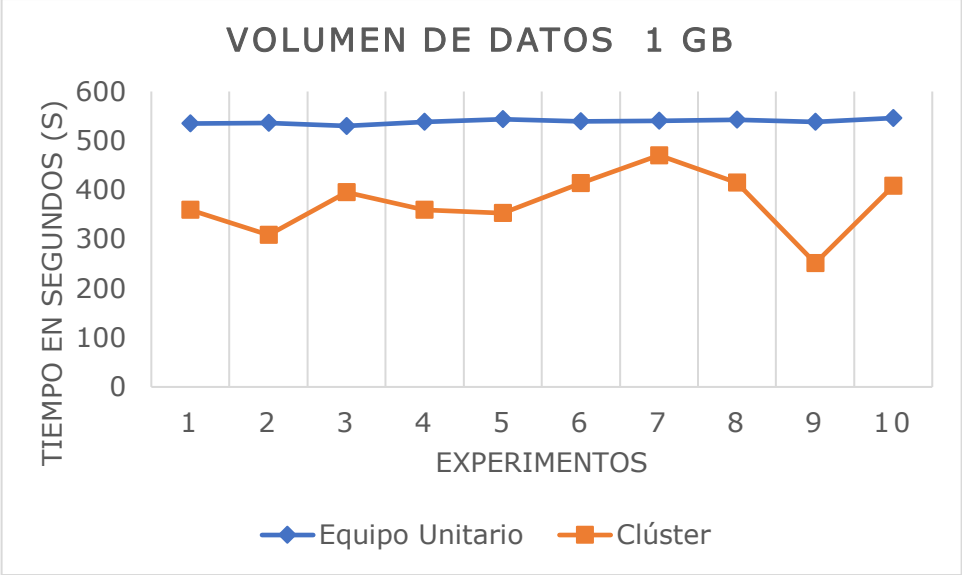


Figura 31. Comparación de eficiencia para un volumen de datos > 1 GB

3.1.5.8. Volumen de datos > 6 GB

La Figura 32 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 8,4175 % sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|5189{,}95-4.753{,}09|}{5189{,}95} = 0{,}08417$ [10].

[9] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 29 y 13, respectivamente.
[10] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 31 y 14, respectivamente.

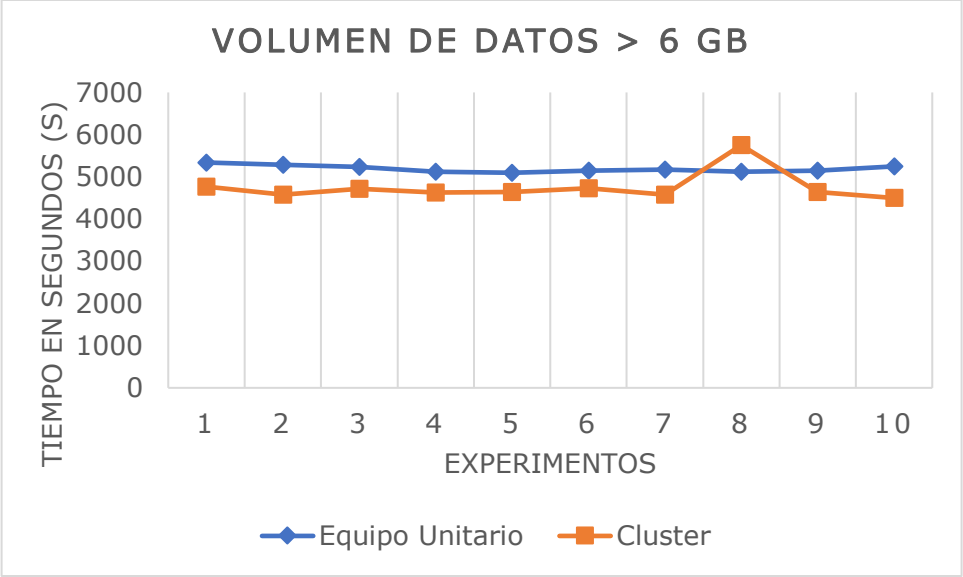


Figura 32. Comparación de eficiencia para un volumen de datos > 6 GB

3.1.5.9. Volumen de datos > 12 GB

La Figura 33 muestra la comparación de rendimiento entre el clúster y el equipo unitario. Luego de realizados los 10 experimentos, el clúster exhibe un rendimiento del 36,50 % sobre el equipo unitario, es decir, aplicando la Eq. 1 se observa que $\frac{|6.896,01-4.378,82|}{6.896.01} = 0,3650$ [11].

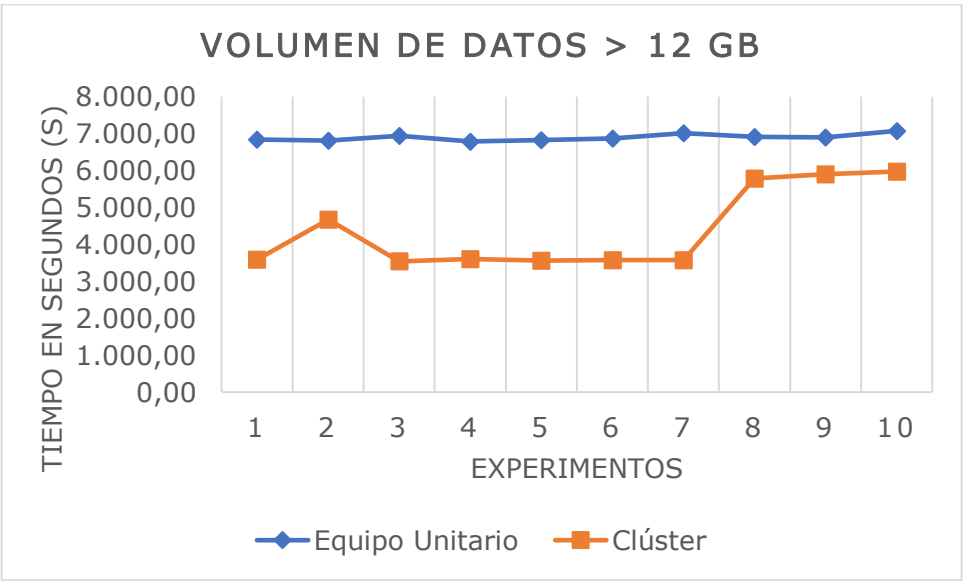


Figura 33. Comparación de eficiencia para un volumen de datos > 12 G

[11] Note que Tc y Ts son calculados a partir de los resultados mostrados en las tablas 33 y 14, respectivamente.

3.2. Evidencia de errores

Algunos de los errores presentados por el clúster están asociados con la comunicación de los datos, el almacenamiento en disco, y gestión de la memoria virtual del clúster. A continuación, se listan alguno de estos errores.

3.2.1. Directorio de salida existe

Este error ocurre cuando el nombre del directorio ya existe en HDFS (org.apache.hadoop.mapred.FileAlreadyExistsException or 'Output directory output already exists'.)

3.2.2. Tipo de dato equivocado

Este error se presenta cuando se utilizan tipos de datos Hadoop erróneos. Por ejemplo, IntWritable en lugar de Text, en los trabajos (i.e., jobs) a ejecutar en el MapReduce.

3.2.3. Error de Conexión al Sistema HDFS

Al ejecutar el comando -put al sistema de archivos HDFS, se genera un error de conexión.

3.2.4. Error configuración de java

El error se presenta cuando no está configurado de manera correcta el JAVA_HOME en el archivo de configuraciones de Hadoop.

3.2.5. Error de protocolo de seguridad SSH

Este error se presenta cuando no se han configura las llaves publicas y privadas de la arquitectura Hadoop.

4. Conclusiones

- Este trabajo describe las características principales de un clúster de computadores. Primero hay que tener presente que, en un ambiente de computo distribuido un clúster depende en gran medida de equipos de cómputo, infraestructura de red, y software para la gestión de computo distribuido; mientras que, en un ambiente de computo paralelo un clúster depende en gran medida en lo equipos de cómputo y el software explote todo el poder de cómputo de dichos equipos.
- Se describe el funcionamiento del software Hadoop en términos de: almacenamiento distribuido (HDFS Hadoop Distributed File System), computo distribuido (MapReduce), y administración de clúster (Yarn).
- Para validar conceptual y tecnológicamente la infraestructura del clúster se realizaron nueve (9) experimentos al sistema, con volumen de datos de diferentes tamaños (refiérase a la sección 1.1 del documento) y se ejecutaron 10 veces para promediar el tiempo de ejecución de cada experimento y evaluar el rendimiento del mismo. Como resultados relevantes se pudo observar que, en el caso de 100 Mb, el rendimiento del equipo unitarios fue superior en un 45% al clúster. Sin embargo, para volúmenes superiores a los 200 Mb e inferiores a 12Gb, el rendimiento del cluster fue superior al equipo unitario en un porcentaje promedio de 25%.

5. Trabajo Futuro

Como trabajo futuro se propone incrementar nuevos equipos de cómputo al clúster, sabiendo que entre más nodos gestione el clúster la eficiencia va a hacer superior y proporciona mayor respaldo a fallos en el sistema.

Se sugiere también implementar nuevos entornos de trabajo (framework) tales como PIG o JAQL, para el análisis de grandes conjuntos de datos en redes sociales (Twitter, LinkedIn, estaciones meteorológicas) los cuales siguen manejando secuencias MapReduce, lo que permite dedicar menos tiempo al desarrollo de aplicaciones y fijar todo el esfuerzo al análisis de datos.

Igualmente se propone crear una infraestructura tecnológica que pueda ser gestionada a distancia, sin necesidad de tener el ingeniero en sitio para poder subir los archivos y realizar las distintas tareas de análisis; además, podría implementarse como fase final tener acceso por medio de una página web donde el usuario se pueda autenticar, subir los archivos de Big Data y lanzar los distintos servicios proporcionados por el clúster.

6. Referencias

[1] BSA, «The Software Alliance,» Octubre 2015. [En línea]. Available: http://data.bsa.org/wp-content/uploads/2015/10/BSADataStudy_es.pdf.

[2] M. d. T. d. l. I. y. l. Comunicaciones, 01 Mayo 2017. [En línea]. Available: http://www.mintic.gov.co/portal/604/w3-article-14678.html.

[3] Pontificia Universidad Javeriana , «Alianza Caoba,» [En línea]. Available: http://alianzacaoba.co/que-es-caoba/alianzas-2/.

[4] Andes, «Cómo sacarle porvecho a la platarorma Hadoop,» *Primer Foro de Ingenieria de Información*, pp. 24-26, 2016.

[5] Centro de Bioinformatico y Biologia Computacional, «BIOS,» 2017. [En línea]. Available: http://bios.org.co/es-es/.

[6] EcuRed, «EcuRed,» 23 Agosto 2017. [En línea]. Available: https://www.ecured.cu/Computaci%C3%B3n_distribuida.

[7] Almasi y Gottlieb, Highly parallel computing, Redwood City: ACM, 1989.

[8] G. Thiruvathukal, «IEEE,» 7 Marzo 2005. [En línea]. Available: http://ieeexplore.ieee.org.bdatos.usantotomas.edu.co:2048/document/1401797/.

[9] R. Barranco, «IBM,» 18 Junio 2012. [En línea]. Available: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/.

[10] V. Barrera, «i2ds,» 14 Mayo 2016. [En línea]. Available: http://www.i2ds.org/algunas-consideraciones-sobre-la-analitica-de-datos-o-data-analytics/.

[11] IBM, «IBM,» 2017. [En línea]. Available: https://www-01.ibm.com/software/cl/data/infosphere/hadoop/que-es.html.

[12] PowerData, «PowerData,» 2017. [En línea]. Available: http://cdn2.hubspot.net/hub/239039/file-884064500-pdf/docs/PWD_-_BIG_DATA_-_hadoop_-_que_significa_haddop_en_el_mundo_del_big_data.pdf.

[13] D. Jeffrey y G. Sanjay, «MapReduce: simplified date processing on large clusters.,» New York, 2008.

[14] J. Ullman, «MapReduce Algorithms,» de *In Proceedings of the 2nd IKDD Conference on Data Sciences*, New York, NY, 2015.

Apéndices

En esta sección se listan las instrucciones para el montaje del Clúster.

Sección 1. Instalación del sistema operativo en los equipos

Paso 1. Crear dispositivo de arranque

1- Empleando la aplicación YUMI, crear el arranque *booteable* del sistema operativo Ubuntu versión 17.04.

Paso 2. Preparar el hardware de los equipos

1- Conectar a cada equipo, los periféricos necesarios para su operación (mouse, teclado, pantalla).
2- Conectar tanto la pantalla como el equipo a la corriente eléctrica.

Paso 3. Instalación del sistema operativo

1- Encender el equipo y entrar al modo BIOS para configurar el arranque a modo USB (USB Device)
2- En la interfaz del YUMI, seleccionar el sistema operativo a instalar (Ubuntu 17.04).
3- En el escritorio, se crea un icono para comenzar la instalación. Dar doble clic al icono.
4- En la configuración de las particiones se sugiere dejar 1000 MB para el área de intercambio (swap), 50000 MB para la ext4/ (root) y el resto del espacio del disco para ext4/home.
5- En la siguiente ventana, seleccionar la zona horaria (Bogotá GMT -5) y país (Colombia).

Sección 2. Configuración de la red local

Paso 1. Configurando la red local.

1- Etiquetar en forma externa los computadores como Maestro y Esclavo. Dejar un solo maestro y los esclavos, numerarlos del 1 al 4 (*slave* 1, *slave* 2…)

2- Usando cable UTP Cat 5E (disponible en el almacén de laboratorio), cablear desde cada tarjeta de red Ethernet de los equipos hacia un puerto del switch configurado.

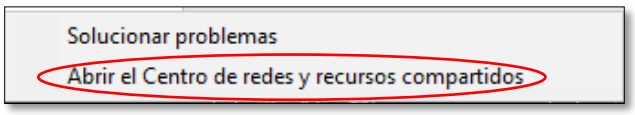
3- Establecer un orden “uno a uno” para cada uno de los equipos esclavos (puerto 1 – slave 1, puerto 2 – slave 2…).

4- Conectar el equipo Master al puerto 10 del Switch configurado.

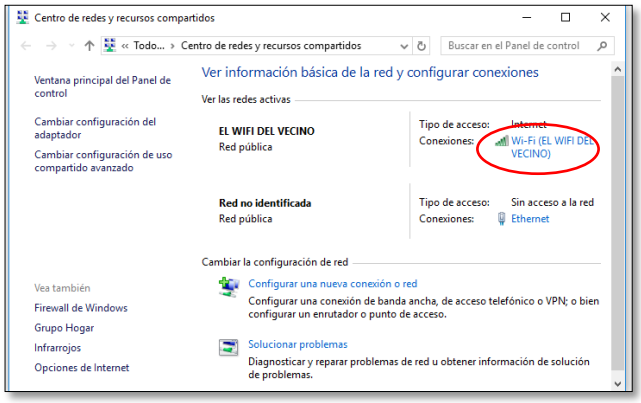
Paso 2. Configurando conectividad a Internet en la red

1- Dar clic derecho al icono de “Configuración de red e Internet”, ubicado en la parte inferior izquierda de la barra de tareas.

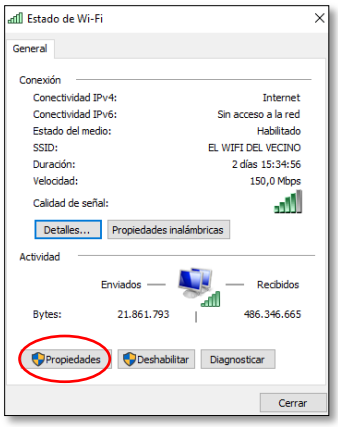
2- Dar clic a la opción “Abrir el centro de redes y recursos compartidos”.



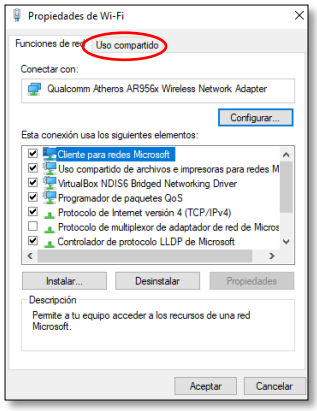
3- Dar clic sobre la red que vamos a utilizar para compartir.



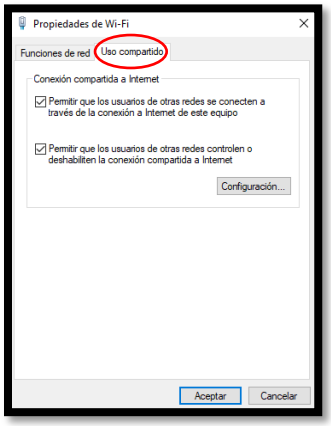
4- En la caja de dialogo Estado de la red, dar clic en el botón Propiedades.



5- En la caja de dialogo Propiedades del Wi-Fi, dar clic en la pestaña Uso compartido.



6- Habilitar la casilla Permitir que los usuarios de otras redes se conecten a través de la conexión a Internet de este equipo.



7- Clic en Aceptar.

IMPLEMENTACIÓN DE UN CLÚSTER

Paso 3. Configurando el acceso de los equipos maestro y esclavos.

Configurar en cada uno de los equipos la IP asignada en la siguiente tabla.

Equipo	IP	Mascara
Master	192.168.137.100	255.255.255.0 /24
Slave 1	192.168.137.101	255.255.255.0 /24
Slave 2	192.168.137.102	255.255.255.0 /24
Slave 3	192.168.137.103	255.255.255.0 /24
Slave 5	192.168.137.105	255.255.255.0 /24

Instalación y configuración de Hadoop

Se debe tener instalado LINUX en versiones superiores de 1.5

Hadoop clúster utilizando cinco sistemas (un Maestro, cuatro esclavos), a continuación, se muestra la tabla con la dirección IP y el usuario de cada sistema.

Equipo	IP	Usuario.
Master	192.168.137.100	Master
Slave 1	192.168.137.101	Slave1
Slave 2	192.168.137.102	Slave2
Slave 3	192.168.137.103	Slave3
Slave 5	192.168.137.105	Slave5

Para realizar la instalación de Hadoop es necesario la instalación previa de JAVA como se muestra a continuación en cada uno de los equipos del sistema.

Instalar Java JDK

IMPLEMENTACIÓN DE UN CLÚSTER

La Instalación del JDK de java se lleva a cabo ejecutando los siguientes comandos en el terminal.

```
$ Sudo apt-get update
$ Sudo apt-get install default-jdk
```

Se instala la última versión disponible en el repositorio de aplicaciones de Ubuntu.

Para comprobar la versión instalada y su correcta instalación de ejecuta el siguiente comando

```
$ Java -version
```

En nuestro caso la ubicación donde queda instalado el java es

```
/usr/lib/jvm/ java-1.8.0-openjdk-amd64
```

para la configuración de la ruta de acceso y JAVA_HOME se agregan los siguientes comandos

```
$ export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
$ export PATH=$PATH:$JAVA_HOME/bin
$    echo    "$    export    JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-
amd64">>~/.bashrc
$ source~/.bashrc
$ echo "export PATH=$PATH:$JAVA_HOME/bin">>~/.bashrc
$ source~/.bashrc
```

Asignación de nodos

Se edita el archivo hosts en /etc/ carpeta en todos equipos, se especifica la dirección IP de cada sistema seguido del nombre del host.

```
$ nano /etc/hosts
192.168.137.100      master
192.168.137.101      slave1
192.168.137.102      slave2
192.168.137.103      slave3
192.168.137.105      slave5
```

Instalar SSH

```
$ sudo apt-get install ssh
```

se debe configurar el acceso al usuario ssh, se realiza con los siguientes comandos

```
$ su – seed
```

ahora, se genera una clave con contraseña para que los nodos se puedan comunicar entre si sin pedir contraseña

```
$ ssh-keygen -t rsa
$ ssh-copy-id -i ~/.ssh/id_rsa.pub seed@master
$ ssh-copy-id -i ~/.ssh/id_rsa.pub seed@slave1
$ ssh-copy-id -i ~/.ssh/id_rsa.pub seed@slave2
$ ssh-copy-id -i ~/.ssh/id_rsa.pub seed@slave3
$ ssh-copy-id -i ~/.ssh/id_rsa.pub seed@slave5
$ chmod 0600 ~/.ssh/authorized_keys
```

Conectar Master con master

$ssh Master
$ssh Slave

Instalar Hadoop

En el servidor Master, descargar e instalar Hadoop siguiendo los siguientes comandos

Recordar que Hadoop siempre está instalado en la dirección /usr/local/Hadoop-2.7.4

```
$ cd /usr/local
$ wget http://mirrors.sonic.net/apache/hadoop/common/hadoop-2.7.4/hadoop-2.7.4.tar.gz
$ tar xvzf Hadoop-2.7.4.tar.gz
```

El siguiente comando se debe ejecutar en todos los slave del sistema

```
$ sudo chown -R seed /usr/local
```

Configurar Hadoop

Hadoop se configura en el Master, haciendo los siguientes cambios

Core-site.xml

Abrir el **Core-site.xml** archivo y editar, como se muestra a continuación

```
$ cd /usr/local/Hadoop-2.7.4/etc/Hadoop
```

```
$ sudo nano Core-site.xml
```

```
<configuration>
<property>
      <name>fs.default.name</name>
      <value>hdfs://192.168.137.100:9000/</value>
</property>
<property>
      <name>dfs.permissions</name>
      <value>false</value>
</property>
<property>
      <name>hadoop.tmp.dir</name>
      <value>/tmp/hadoop-${user.name}</value>
</property>
</configuration>
```

Hdfs-site.xml

Abrir el **Hdfs-site.xml** archivo y editar, como se muestra a continuación:

```
$ cd /usr/local/Hadoop-2.7.4/etc/Hadoop
$ sudo nano Hdfs-site.xml
```

```
<configuration>
<property>
      <name>dfs.data.dir</name>
      <value>file:///usr/local/hdfs/namenode</value>
</property>
<property>
      <name>dfs.name.dir</name>
      <value>file:///usr/local/hdfs/datanode</value>
</property>
<property>
      <name>dfs.replication</name>
      <value>3</value>
</property>
</configuration>
```

mapred-site.xml

Abrir el **mapred-site.xml** archivo y editar, como se muestra a continuación

```
$ cd /usr/local/Hadoop-2.7.4/etc/Hadoop
```

$ sudo nano **mapred-site.xml**

```
<configuration>
<property>
      <name>mapred.job.tracker</name>
      <value>192.168.137.100:9001</value>
</property>
</configuration>
```

Yarn-site.xml

Abrir el **yarn-site.xml** archivo y editar, como se muestra a continuación:

$ cd /usr/local/Hadoop-2.7.4/etc/Hadoop

$ sudo nano **yarn-site.xml**

```
<configuration>
<property>
      <name>yarn.resurcemanager.hostname</name>
      <value>192.168.137.100</value>
</property>
<property>
      <name>yarn.resourcemanager.address</name>
      <value>${yarn.resourcemanager.hostname}:8032</value>
</property>
<property>
      <name>yarn.resourcemanager.webapp.address</name>
      <value>${yarn.resourcemanager.hostname}:8088</value>
</property>
<property>
      <name>yarn.resourcemanager.schedule.address</name>
      <value>${yarn.resourcemanager.hostname}:8030</value>
</property>
<property>
      <name>yarn.resourcemanager.resource-tracker.address</name>
      <value>${yarn.resourcemanager.hostname}:8031</value>
</property>
<property>
      <name>yarn.resourcemanager.admin.address</name>
      <value>${yarn.resourcemanager.hostname}:8033</value>
</property>
<property>
      <name>yarn.nodemanager.aux-services</name>
      <value>mapreduce_shuffle</value>
</property>
```

```
<property>
      <name>yarn.scheduler.minimum-allocation-vcores</name>
      <value>1</value>
</property>
<property>
      <name>yarn.scheduler.maximum-allocation-vcores</name>
      <value>2</value>
</property>
<property>
      <name>yarn.scheduler.minimum-allocation-mb</name>
      <value>128</value>
</property>
<property>
      <name>yarn.scheduler.maximum-allocation-mb</name>
      <value>2048</value>
</property>
<property>
      <name>yarn.nodemanager.resource.memory-mb</name>
      <value>4096</value>
</property>
<property>
      <name>yarn.nodemanager.resource.cpu-vcores</name>
      <value>4</value>
</property>
</configuration>
```

hadoop-env.sh

Abrir el **hadoop-env.sh** archivo y editar JAVA_HOME, HADOOP_CONF_DIR Y HADOOP_OPTS, como se muestra a continuación.

```
$ cd /usr/local/Hadoop-2.7.4/etc/Hadoop

$ sudo nano hadoop-env.sh
```

```
export JAVA_HOME=/usr/
export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
export HADOOP_CONF_DIR=/usr/local/hadoop-2.7.4/etc/hadoop/conf
```

Instalación de hadoop en servidores slave

IMPLEMENTACIÓN DE UN CLÚSTER

Se instala Hadoop en todos los servidores Slave desde el Master siguiendo los siguientes comandos.

```
$ cd /usr/local
$ scp -r hadoop-2.7.4 seed@slave1:/usr/local/hadoop-2.7.4
$ scp -r hadoop-2.7.4 seed@slave2:/usr/local/hadoop-2.7.4
$ scp -r hadoop-2.7.4 seed@slave3:/usr/local/hadoop-2.7.4
$ scp -r hadoop-2.7.4 seed@slave5:/usr/local/hadoop-2.7.4
```

Configuración de Hadoop em el servidor Master

En el servidor Master configurar lo siguiente

```
$ cd /usr/local/hadoop-2.7.4
```

Configuración de nodos secundarios

```
$ sudo nano /etc/hadoop/slaves
```

Master
slave1
slave2
slave3
slave5
Nombre de formato nodo maestro sobre Hadoop

```
$ cd /usr/local/hadoop-2.7.4/etc/hadoop

$ hadoop namenode -format
```

Ejecución del clúster

```
$ cd /usr/local/hadoop-2.7.4/sbin

$ ./start-all.sh
```

Después de lanzar los servicios con el comando anterior, se ejecuta JPS

```
$jps
Resultado
14799 NameNode

15314 Jps

14880 DataNode
14977 SecondaryNameNode
```