Revista de Investigación
Hatun Yachay Wasi 2(1), 98 - 108 (2023)
ISSN: 2955 - 8255

ARTICULO ORIGINAL

Recibido: febrero 2023 Aceptado: abril 2023

# ESTIMACIÓN DEL RENDIMIENTO DE ARQUITECTURA HOMOGÉNEA Y/O HETEROGÉNEA PARA BIG DATA

## HOMOGENEOUS AND/OR HETEROGENEOUS ARCHITECTURE PERFORMANCE ESTIMATION FOR BIG DATA

Claudio Isaias Huancahuire Bravo[1] Abbon Alex Vasquez Ramirez[1]
Javier Rozas Huacho[2]

[1] Universidad Tecnológica de los Andes-Abancay-Perú
[2] Universidad Nacional de San Antonio Abad del Cusco-Perú

**Correspondencia:**
Mag. Claudio Huancahuire Bravo
**chuancahuireb@utea.edu.pe**

**Como citar este articulo:** Huancahuire, C., Vasquez, A., & Rozas, J. (2023). Estimación del rendimiento de arquitectura homogénea y/o heterogénea para Big data. *Hatun Yachay Wasi 2*(1), 98 - 108. https://doi.org/10.57107/hyw.v2i1.39

**RESUMEN**

La cuarta revolución industrial interactúa con otras vertientes como Cloud Computing, Internet de las Cosas, Ciencia de Datos, Ingeniero de datos, Inteligencia Artificial con Machine Learning. Porque cada vez es más inevitable, no transformar los datos del mundo real en datos digitales como: Textos, audio, imágenes, videos, etc, para su tratamiento y una óptima toma de decisión, en el contexto que se requiera. En consecuencia, de las tecnologías mencionadas deviene el término de Big Data, que subyace con términos estructurados, semi estructurados y no estructurados y todo ello tiene que ser procesado, administrado y gestionado mediante técnicas de ETL, Power BI Desktop y Power BI de servicio cloud, Looker Studio, Arquitectura de Hadoop para Big Data, ASF-Apache Software Foundation, brinda un respaldo al ecosistema de Hadoop, para crear, diseñar y aplicar como investigación, aplicación y distribución en Universidades, PYMES y Empresas e industrias respectivamente, además las multinacionales empresas como Oracle cloud, IBM, Amazon, Azure y Google, se basan con esta tecnología de código abierto – open source de Hadoop.

**Palabras clave:** Hadoop, HDFS, MapReduce, predicción, YARN, clúster, datos estructurados, datos no estructurados

**ABSTRACT**

The fourth industrial revolution interacts with other aspects such as Cloud Computing, Internet of Things, Data Science, Data Engineering, Artificial Intelligence with Machine Learning. Because it is increasingly inevitable, not to transform real world data into digital data such as: Texts, audio, images, videos, etc., for its treatment and optimal decision making, in the context that is required. Consequently, from the aforementioned technologies comes the term Big Data, which underlies structured, semi-structured and unstructured terms and

all of this has to be processed, administered and managed using ETL, Power BI Desktop and Power BI cloud service techniques. , Looker Studio, Hadoop Architecture for Big Data, ASF-Apache Software Foundation, provides support to the Hadoop ecosystem, to create, design and apply as research, application and distribution in Universities, SMEs and Companies and industries respectively, as well as multinationals Companies such as Oracle cloud, IBM, Amazon, Azure and Google, are based on this open source technology – Hadoop open source.

**Keywords:** Hadoop, HDFS, MapReduce, prediction, YARN, cluster, structured data, unstructured data

## INTRODUCCIÓN

En la actualidad los datos digitales se están generando exponencialmente, de una forma que no se imaginaba antes; hoy por hoy se disponen de dispositivos tecnológicos como celulares, portátiles, Smart TV, Smart Watch, Tablet entre otros, los cuales hacen parte de la vida cotidiana a nivel mundial (Li et al., 2015)

Por lo que, en la actualidad, hay una gran inquietud sobre el manejo y uso de grandes volúmenes de datos, es de allí donde se originan diferentes disciplinas y tecnologías como la Big Data que requiere obtener un beneficio para las diferentes organizaciones y sociedades (Serrano, 2014).

El crecimiento en el volumen de datos generados por diferentes sistemas y actividades cotidianas en la sociedad ha forjado la necesidad de modificar, optimizar y generar métodos y modelos de almacenamiento y tratamiento de estos que suplan las falencias que presentan las bases de datos y los sistemas de gestión de datos tradicionales. En este sentido, Big Data, es un término que incluye diferentes tecnologías asociadas a la administración de grandes volúmenes de datos, provenientes de diferentes fuentes y que se generan con rapidez (Li et al., 2015)

A pesar de que este término se asocia principalmente con cantidades de datos exorbitantes, se debe dejar de lado esta percepción, pues no solo va dirigido a esto, sino que abarca tanto volumen como variedad de datos y velocidad de acceso y procesamiento. En la actualidad, se ha pasado de la transacción a la interacción, con el propósito de obtener el mejor provecho de la información que se genera minuto a minuto (Mohanty, 2015).

Para almacenar y analizar Big Data, Hadoop es la herramienta más común para los investigadores y científicos. Este almacenamiento de una gran cantidad de datos en Hadoop, se realiza mediante el sistema de archivos distribuidos de Hadoop Distributed File System (HDFS), el cual divide un archivo muy grande en bloques pequeños y colocarlos en el clúster de forma distribuida. Básicamente, Hadoop y HDFS se diseñaron de tal manera que, funcionan eficientemente en el clúster homogéneo. Sin embargo, en esta era de redes, no se puede imaginar tener un grupo de nodos homogéneos solamente; por lo tanto, existe la necesidad de una política de almacenamiento que pueda funcionar de manera eficiente, tanto en clústeres homogéneos como heterogéneos, por lo que, las necesidades de aplicaciones que se pueden ejecutar de manera eficiente en el tiempo, basadas en entornos homogéneos y heterogéneos, pueden ser suficientes (Shah & Padole, 2018).

La localización de datos en Hadoop asigna el

bloque de datos para procesar en el mismo nodo; sin embargo, en Big Data, se necesita realizar esto mediante múltiples nodos. Hadoop tiene como funcion copiar el bloque de datos donde se ejecutan los claves y valor; lo que produce menos rendimiento, principalmente en clúster heterogéneo debido a retardo o congestiones en la entrada/salida de red (Shah & Padole, 2018).

## MATERIALES Y MÉTODOS

Se tomó como muestra 10 unidades de computadoras convencionales o básicas, que tuvieron las siguientes características de la tecnología homogénea, donde cada PC está integrado de un hardware con disco magnético, memoria RAM, CPU Intel y software de sistema operativo centOS 7, Hadoop, Java y compiladores e interconectado con la dirección IP clase C, quedando los 10 nodos que consolidan la arquitectura del clúster homogéneo, para el tratamiento Big data.

El método desarrollado en la utilización de Hadoop para Big Data siguió la siguiente metodología de esta exploración: instalación del sistema operativo CentOS 7, configuración de Hadoop versión 2.x, HDFS, MapReduce, la configuración para el clúster, se utilizó IP la clase C y finalmente se realizó la prueba con la ejecución de MapReduce en HDFS.

## METODOLOGÍA HOMOGÉNEA

Infraestructura homogénea de Hadoop.

El análisis, diseño e implementación del clúster homogéneo, se basa exclusivamente con computadoras de las mismas características (Tabla1).

**TABLA 1**

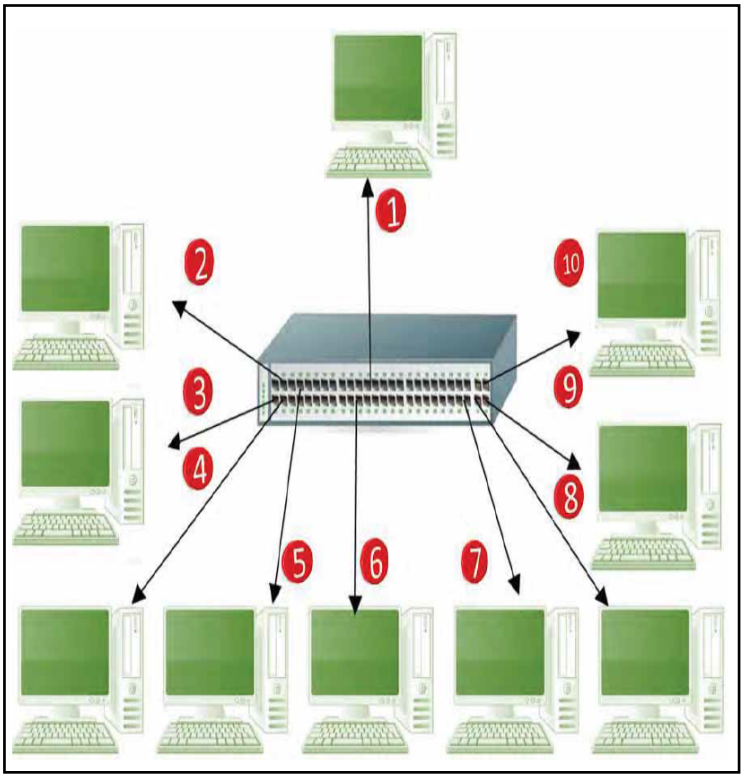
*Características de ordenadores homogéneas*

| Descripción | Tiempo de rendimiento |
|---|---|
| Sistema | Lenovo |
| RAM | 16 GB |
| Disco HDD | 1 TB |
| Procesador | Intel(R)Xeon(R) CPU E3-1270 v6 @ 3.8Ghz 3.79GHz |
| CPU | 64 bits |
| Sistema Operativo | CentOS 7 |
| Hadoop | 2.9 |
| Java | Versión JDK 1.8 |
| Dirección IP | Clase C |

La Figura 1, muestra 10 nodos, con sus características iguales y/u homogéneas, el primer nodo es denominado como maestro y los 9 nodos restantes son denominados esclavos, que constituyen el clúster homogéneo.

**FIGURA 1**

*Clúster homogéneo*




**Rendimiento de clúster homogéneo**

Para calcular el rendimiento, se trabajó con los datos de 12.8 GB y 6.4 GB con formato semi estructurados, para analizar, diseñar e implementar la arquitectura homogénea. Se procedió a ejecutar los datos semiestructurados, en una computadora conllevando su rendimiento en 5 minutos y 33 segundos, se adicionó horizontalmente a 3 nodos con mejor rendimiento de 2 minutos y 29 segundos y, llegando a 10 nodos con un rendimiento de 1 minuto y 2 segundos, llevando a que aumentando más PC-nodos, se tiene mejor rendimiento mediante el uso de la plataforma Open Source de Apache (Hadoop).

**TABLA 2**

*Características de ordenadores homogéneas*

| Nodos-PC | Tiempo de rendimiento |
|---|---|
| 1 | 5 minutos y 33 segundos. |
| 3 | 2 minutos y 29 segundos. |
| 5 | 1 minuto y 29 segundos. |
| 7 | 1 minuto y 18 segundos. |
| 10 | 1 minuto y 2 segundos. |

## METODOLOGÍA HETEROGÉNEA

### Infraestructura heterogénea de Hadoop

La Figura 2, vislumbra los 10 ordenadores con características diferentes y/o heterogéneas, el primer ordenador es denominado maestro y los nueve nodos son denominados esclavos, que constituyen parte del clúster heterogéneo, similar al homogéneo.

**FIGURA 2**

*Características del clúster heterogéneo*

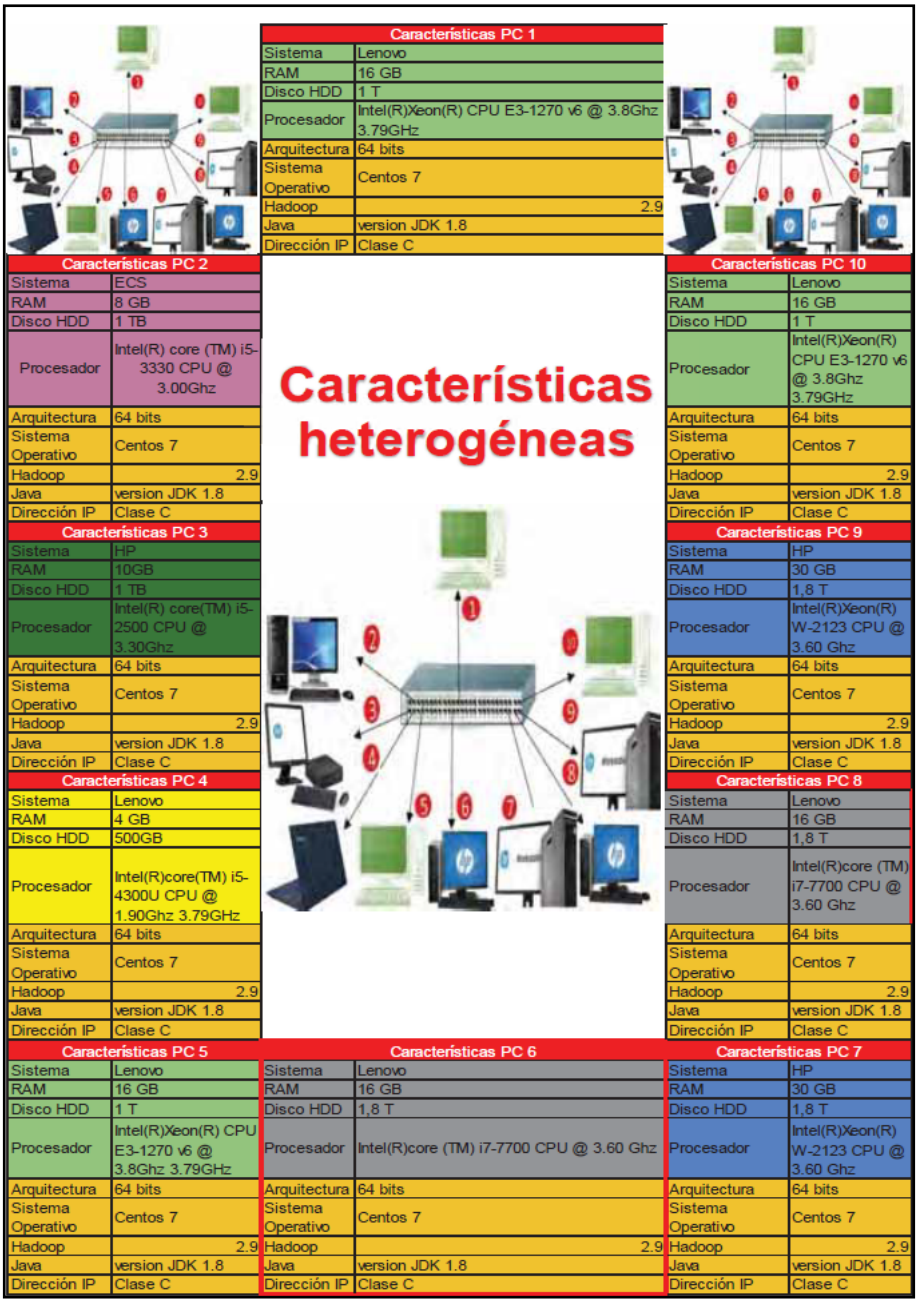
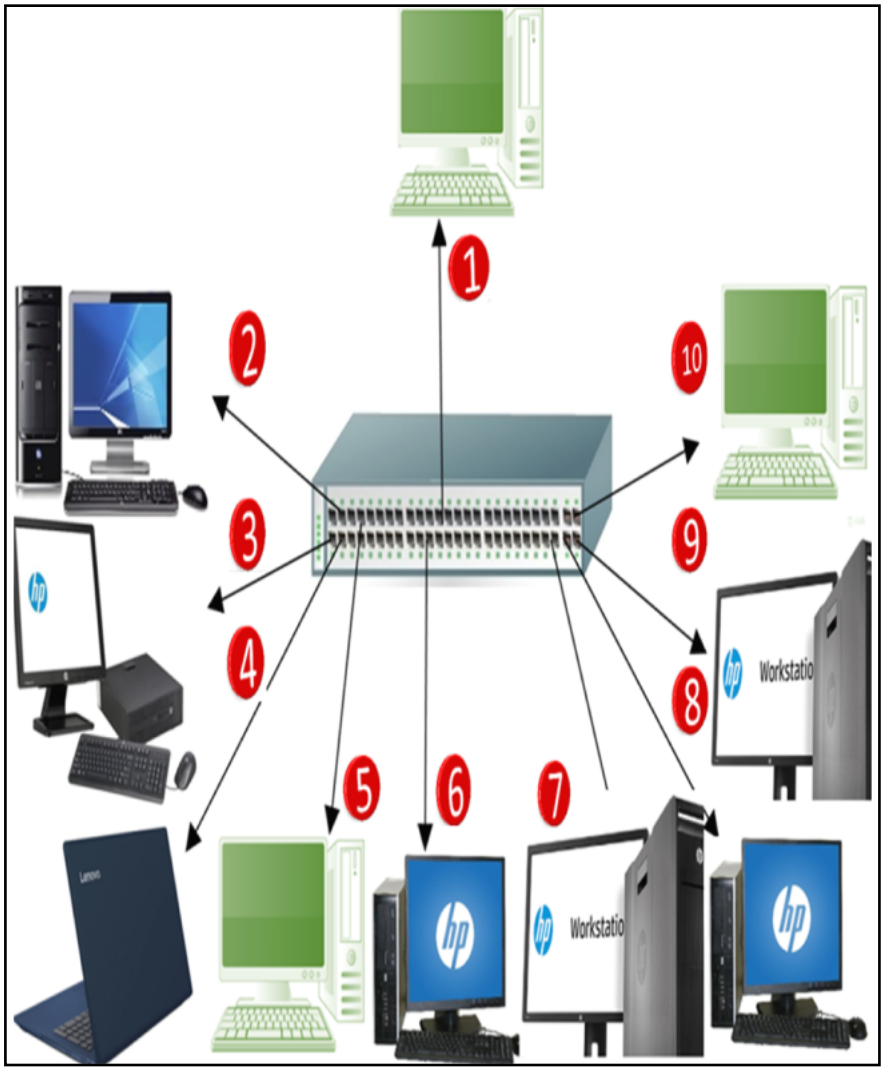


**FIGURA 3**

*Clúster heterogéneo*




**Rendimiento de clúster heterogéneo**

Las características de la tecnología heterogénea consisten en la conexión de 10 computadoras (PC o nodos) convencionales y módicas que cada PC está integrado de hardware con disco magnético, memoria RAM, CPU Intel diferente entre la PC-nodo (2), a la laptop-nodo (4), a las PC-nodos (1,5 y10) la PC-nodo (3,7 y9); así mismo a la PC-nodo (6 y 8) y software de sistema operativo centOS 7, Hadoop v2, Java JDK8, conformando la arquitectura heterogénea para el tratamiento de datos semiestructurados.

En primera instancia se realizaron pruebas de rendimiento con 6.4 GB en clúster heterogéneo, sin embargo, para mejorar la estimación o predicción se realizó en segunda instancia con 12.8 GB.

**TABLA 3**

*Tiempo de rendimiento en nodos heterogéneo*

| Nodos-PC | Tiempo de rendimiento |
|---|---|
| 1 | 5 minutos y 33 segundos. |
| 3 | 4 minutos y 31 segundos. |
| 5 | 2 minutos y 44 segundos. |
| 7 | 2 minutos y 44 segundos. |
| 10 | 1 minutos y 31 segundos. |

## RESULTADOS

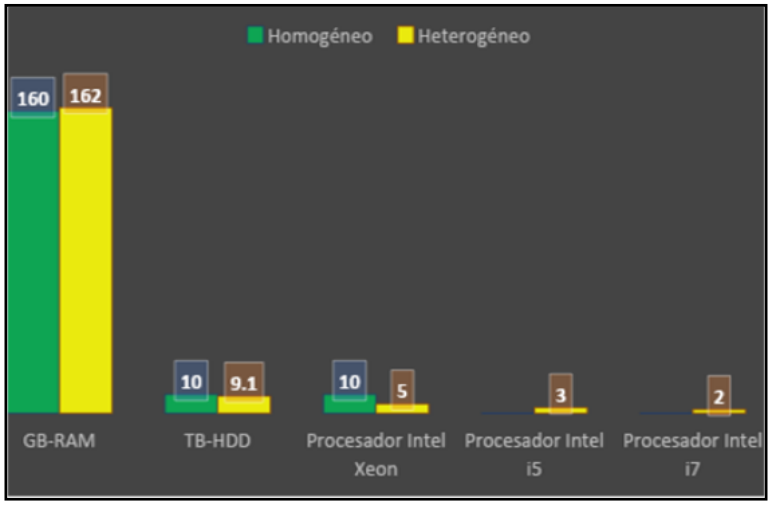
Se logró implementar el almacenamiento y procesamiento distribuido de datos no estructurados sobre HDFS y ejecutar el proceso paralelo con el modelo de programación MapReduce y administrar con YARN, sobre los clústeres de características homogéneas y heterogéneas.

Se hizo la comparación entre hardware homogéneo vs heterogéneo: memoria de 162 GB de RAM de clúster heterogéneo es mayor en 2GB RAM que clúster homogéneo.

Disco duro de10 TB HDD de clúster homogéneo es mayor en 0.9 TB HDD de clúster heterogéneo. 10 procesadores Intel Xeon integran clúster homogéneo, mientras que, el clúster heterogéneo está integrado por cinco procesadores Intel Xeon, tres procesadores Intel i5 y dos procesadores Intel i7 (Fig. 4).

**FIGURA 4**

*Comparación de clúster homogéneo vs heterogéneo*

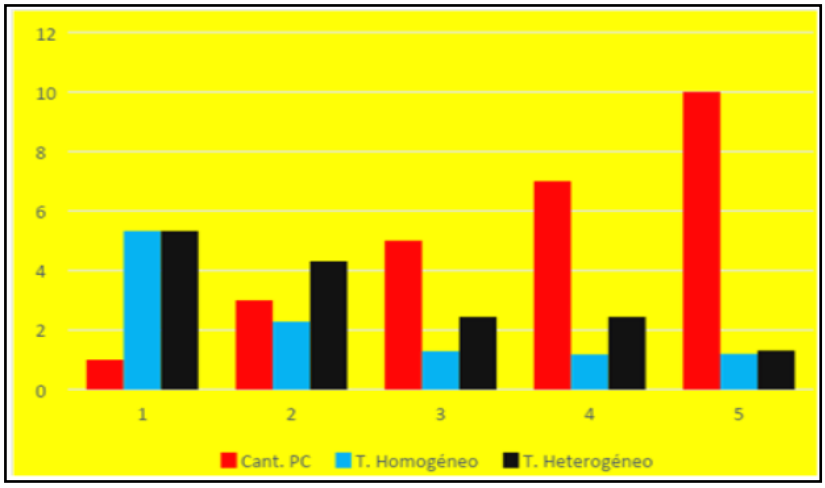



**Tiempo de rendimiento**

Mientras más cantidad de PC o nodos, menos el tiempo de rendimiento, como muestra la Figura 5. Las barras rojas indican la cantidad de PC o nodos, barras celestes y negras indican tiempo de rendimiento en clúster homogéneo y clúster heterogéneo respectivamente.

**FIGURA 5**

*Contraste de estimación mediante recta de regresión*



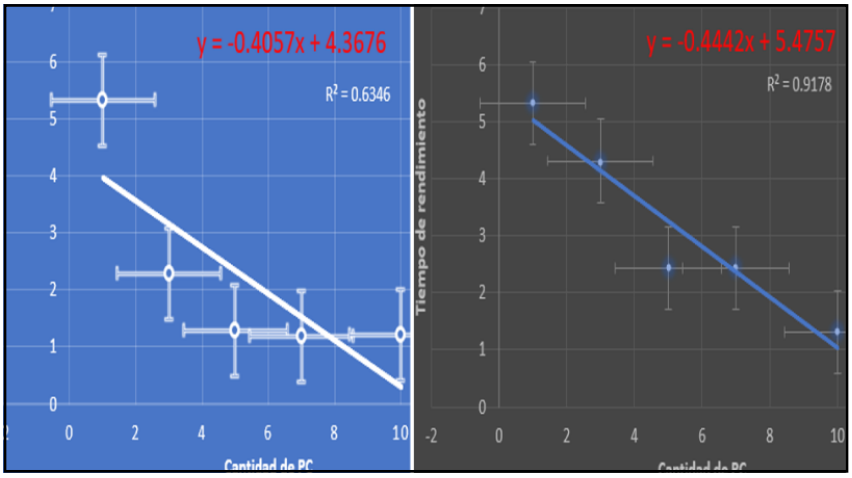

**Contraste, análisis de correlación**

En clúster homogénea, en base a su coeficiente de determinación, el coeficiente de correlación r es 0.7966, lo que significa que 80 % de datos se relacionan entre sí y su dirección depende del signo de su pendiente.

En clúster heterogéneo, en base a su coeficiente de determinación, el coeficiente de correlación r es 0.9580, es decir, 96 % de los datos se relacionan entre sí y su dirección depende del signo de su pendiente; el 96 % de la cantidad de PC o nodos están más estrechamente relacionados con su tiempo de rendimiento, que el clúster homogéneo (Fig. 6).

**FIGURA 6**

*Contraste, coeficiente de determinación*




**Contraste entre clúster homogéneo y heterogéneo**

El proceso de la diferencia de los tiempos de rendimiento entre clúster heterogéneo y homogéneo generados por cantidades de nodos, con el siguiente dato semiestructurado de 12.8 GB, entre clúster heterogéneo y clúster homogéneo, se detalla a continuación: con un 1 nodo la diferencia es 0 segundos, con tres nodos, 4 min. y 2 seg, con

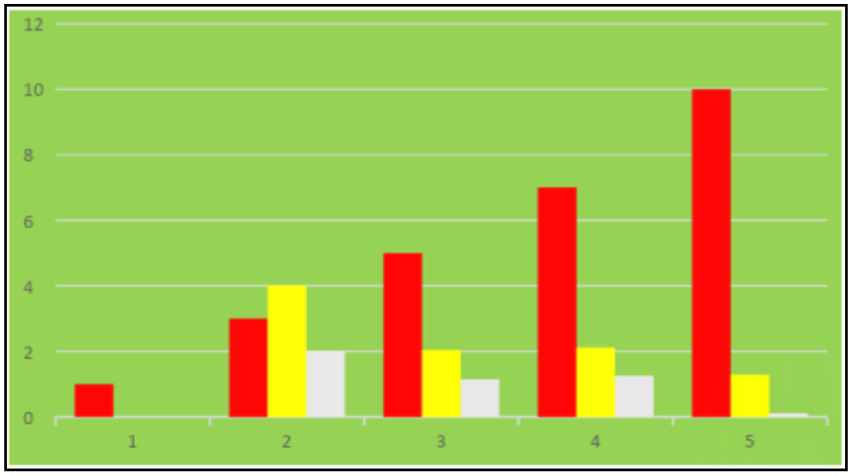
cinco nodos la diferencia es 2 min y 4 seg, con 7 nodos, 2 min y 12 seg y con 10 nodos, 1 min y 29seg. Se afirma la estimación, que el tiempo de rendimiento con 12.8 GB en clúster heterogéneo es casi el doble que en clúster homogéneo.

Por otra parte, con 6.4 GB la diferencia entre clúster heterogéneo y homogéneo, se puede visualizar que se mantiene el tiempo de rendimiento con mayor y menor grado de GB.

La estimación enmarcada general es que, mientras más nodos integran el clúster, el tiempo de rendimiento es óptimo (Fig. 7).

**FIGURA 7**

*Diferencia entre clúster heterogéneo y homogéneo*




## DISCUSIÓN

MapReduce de Hadoop y Apache Spark, son dos herramientas muy importantes que se utilizan para el procesamiento de Big Data. El procesamiento comenzó con MapReduce Framework de Hadoop, pero adolece de muchas desventajas debido a las múltiples operaciones de cesamiento de discos. Los inconvenientes del procesamiento tradicional de Big Data se han superado en un marco de manejo de memoria como Spark (Sharma & Kaur, 2019).

La arquitectura diseñada y aplicada en la presente investigación es para el tratamiento de trabajos por lotes, en arquitectura homogéneas y heterogéneas.

Hoy en día, con el desarrollo continuo de la tecnología moderna de Internet y comercio electrónico, los datos de la red están creciendo geométricamente y ha llegado la era de los grandes datos. En este trabajo se construye una nueva tecnología de minería y análisis de datos basada en el algoritmo de minería tradicional *a priori* y aplicando la última tecnología Hadoop. Esta tecnología se centra en los siguientes dos aspectos. El primero es la mejora del algoritmo tradicional, que se aplica principalmente a la tecnología Hadoop para actualizar el algoritmo de datos tradicional; la segunda parte es el procesamiento de análisis paralelo de datos de minería. Las dos partes principales constituyen la nueva tecnología de análisis de minería de datos grandes (Dong, 2022)

Las dos arquitectura homogéneas y heterogéneas son permisibles no solo para el tratamiento de la minería de datos, sino para ciencia de datos, ingeniería de datos, aprendizaje automático y cloud computing.

Los enfoques tradicionales de análisis y extracción no funcionan bien para Big Data porque estos datos son complejos y de gran volumen. Por lo tanto, existe la necesidad de diseñar un algoritmo de agrupamiento eficiente y altamente escalable. En este documento, presentamos un nuevo algoritmo de agrupamiento llamado agrupamiento híbrido para superar las desventajas de los algoritmos de agrupamiento existentes convencionales. A partir de los resultados experimentales, está claro que el algoritmo de agrupamiento híbrido propuesto es más preciso y tiene mejores valores de precisión, recuperación y medida (Kumar & Singh, 2019).

La arquitectura homogénea y heterogénea están prestos para los algoritmos híbridos, empero tiene superior rendimiento que la arquitectura homogénea, en consecuencia, se aplica el algoritmo hibrido sobre la arquitectura de mejor rendimiento.

## CONCLUSIONES

Se logró integrar un clúster de 10 nodos, con un único nodo maestro y 9 nodos esclavos, todos ellos con las mismas características de hardware básico.

Se logró integrar un clúster de 10 nodos, con un único nodo maestro y 9 nodos esclavos, con diferentes características de hardware básico.

Se estructuró el almacenamiento sobre HDFS, logrando obtener la tolerancia a fallas, con las réplicas generadas en los nodos esclavos.

Se implementó el procesamiento paralelo, con el modelo de programación MapReduce versión 2, con lenguaje de programación Java, con su empaquetado en Jar y compilador JDK compatible entre versiones.

Se obtuvo en el diseño una seguridad local y remota con el protocolo de comunicación SSH (Secure SHell) y con el sistema criptográfico RSA.

Se determinó la correlación estadística de los tiempos de procesamiento en función del número de nodos, en una propuesta de implementación de una arquitectura de clúster homogénea y heterogénea con Hadoop.

Permite estimar los tiempos de rendimientos del procesamiento distribuido de datos no estructurados con MapReduce de grandes volúmenes de datos; evitando así efectuar procesos experimentales, con el consecuente ahorro de tiempo y costos.

## REFERENCIAS BIBLIOGRÁFICAS

Dong, Z. (2022) *Research of Big Data Information Mining and Analysis: Technology Based on Hadoop Technology,* I*nternational Conference on Big Data, Information and Computer Network (BDICN)*, Sanya, China.

Kumar, S., & Singh, M. (2019). A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. *Big Data Mining and Analytics, 2* (4), 240-247 DOI: 10.26599/BDMA.2018.9020037

Li, K., Jiang, H., Yang, L., & Cuzzocrea, A. (2015). *Big data. Algorithms, Analytics, and Applications* https://doi.org/10.1201/b18050

Mohanty, H., Bhuyan P., & Chenthati D. (2015) *Big Data.* https://link.springer.com/book/10.1007/978-81-322-2494-5

Serrano, J. (2014). Big data y analítica web. Estudiar las corrientes y pescar en un océano de datos. *Profesional de la Información, 23*(6), 561–566. https://doi.org/10.3145/epi.2014.nov.01

Shah, A., & Padole, M. (2018). *Load Balancing through Block Rearrangement Policy for Hadoop Heterogeneous Cluster. International Conference on Advances in Computing, Communications, and Informatics (ICACCI)*, Bangalore, India.

Sharma, M., & Kaur, J. (2019). *A Comparative Study of Big Data Processing: Hadoop vs. Spark.* 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India.