**UNIVERSIDAD TÉCNICA DE BABAHOYO**

**FACULTAD DE ADMINISTRACIÓN, FINANZAS E INFORMÁTICA**

**PROCESO DE TITULACIÓN**

MAYO 2023 – SEPTIEMBRE 2023

**EXAMEN COMPLEXIVO DE GRADO O DE FIN DE CARRERA**

**PRUEBA PRÁCTICA**

**PREVIO A LA OBTENCIÓN DEL TÍTULO DE:**

INGENIERO EN SISTEMAS DE INFORMACION

**TEMA:**

# ESTUDIO DE LA HERRAMIENTA MAP REDUCE Y SU UTILIZACIÓN EN LA BIG DATA

**ESTUDIANTE:**

AARON RAMIREZ PALMA

**TUTOR:**

ING. CARLOS SOTO VALLE

**AÑO 2023**

**ÍNDICE GENERAL**

PLANTEAMIENTO DEL PROBLEMA .....5

JUSTIFICACIÓN.....9

OBJETIVOS.....11

OBJETIVO GENERAL .....11

OBJETIVOS ESPECIFICOS.....11

LÍNEAS DE INVESTIGACIÓN .....12

MARCO CONCEPTUAL.....13

MARCO METODOLOGICO.....25

RESULTADOS.....26

DISCUSION DE RESULTADOS .....38

CONCLUSIONES .....40

RECOMENDACIONES .....41

BIBLIOGRAFÍA .....42

**Índice de Figuras**

Figura 1 .....26
Figura 2 .....26
Figura 3 .....27
Figura 4 .....27
Figura 5 .....28
Figura 6 .....28
Figura 7 .....29
Figura 8 .....30
Figura 9 .....30
Figura 10 .....30
Figura 11 .....31
Figura 1 .....31
Figura 13 .....32
Figura 14 .....32
Figura 15 .....32
Figura 16 .....33
Figura 17 .....33
Figura 18 .....34
Figura 19 .....34
Figura 20 .....35
Figura 21 .....35
Figura 22 .....36
Figura 23 .....36
Figura 24 .....37

**Índice de Tablas**

Tabla 1. .....20

Tabla 2. .....45

Tabla 3. .....46

**Resumen**

Existe un paradigma que va de la mano con lo que se conoce como programación y computación paralela, la misma que establece una inspirada respuesta en la organización de datos en volúmenes enormes. En este contexto se incorpora en esta investigación un framework que permita el procesamiento de grandes cantidades de datos a través de computación paralela en ambientes distribuidos.

En consecuencia, a lo expresado anteriormente, la presente investigación está enfocada en explorar como se utiliza la herramienta Map Reduce en la administración de Big Data y con ello aportar información significativa a los estudiantes de pregrado en la utilización de esta herramienta para el procesamiento de grandes conjuntos de datos. Esta investigación se desarrolla utilizando el método descriptivo, enmarcado en la obtención de información de distintas fuentes que permitan cumplir con el propósito de la investigación, además es importante mencionar que el estudio se centra en el funcionamiento de la herramienta Map Reduce, con un punto de vista más pragmático.

**Palabras claves:** Big Data, Map Reduce, Procesamiento distribuido, Volumen de datos, Paralelismo.

**Summary**

There is a paradigm that goes hand in hand with what is known as programming and parallel computing, which establishes a response inspired by the organization of data in enormous volumes. In this context, a framework that allows the processing of large amounts of data through parallel computing in distributed environments is incorporated into this research.

As a result of what was expressed above, this research is focused on exploring how the Map Reduce tool is used in the management of Big Data and thereby providing significant information to undergraduate students in the use of this tool for the processing of large sets. . of data. This research is developed using the bibliographic method, framed in obtaining information from different sources that allow the purpose of the research to be fulfilled. It is also important to mention that the study focuses on the operation of the Map Reduce tool, with a point of More pragmatic view.

**Keywords:** Big Data, Map Reduce, Distributed Processing, Data Volume, Parallelism.

## PLANTEAMIENTO DEL PROBLEMA

En la última década en Ecuador, la tecnología ha experimentado una evolución significativa con relación al procesamiento y análisis de datos en proporciones elevadas, lo cual contribuye al fortalecimiento y adiestramiento de los profesionales y demás empresas que forman parte del sector productivo del país.

Al hablar de la Big Data, se tiene como propósito principal, la administración de grandes cantidades de datos, con el objetivo de experimentar una nueva metodología para el análisis de información que genere una creciente demanda de soluciones tecnológicas con la capacidad de obtener conocimientos que fortalezcan las habilidades del ser humano. El crecimiento exponencial de los datos generados en la era digital da lugar al desafío de procesar, analizar y abordar un lenguaje computacional que adicione intrínsecamente técnicas y herramientas como Map Reduce.

El contexto de Big Data aporta un principal problema en el procesamiento eficiente de datos, debido a este principio, se requiere una herramienta con versatilidad para dividir y distribuir la carga de trabajo de manera efectiva, asegurando un procesamiento paralelo y escalable., como la herramienta Map Reduce, la cual utiliza una técnica y modelo de programación desarrollada por Google y con el pasar del tiempo se ha convertido en una solución popular para abordar esta problemática, sin embargo, a pesar de su popularidad y amplia adopción en la industria, la utilización de Map reduce en el procesamiento de Big Data presenta una serie de desafíos que aún no han sido resueltas.

Map Reduce emplea un algoritmo basado en el modelo cliente-servidor y define cada nodo como un punto de intersección que permita la interacción entre el framework y el usuario, de igual forma se categoriza como una lista de espera, en la que el modelo de negocio trabaja con la mentalidad de que el primero en llegar será el primero en ser atendido.

Es así como se introduce en el escenario de un método denominado JobTracker o servidor maestro y Task Traker. La diferencia entre los dos se enfoca en la parametrización de las listas de carga que son enviadas desde el lado del cliente hacia el servidor para su respectivo análisis.

## JUSTIFICACIÓN

El presente trabajo de investigación se justifica debido a la importancia que sigue a nivel mundial el área del conocimiento con respecto al análisis de información, el modelamiento y el debido procesamiento de todo tipo de información sin importar el orden y el tamaño de la muestra.

El presente caso de estudio se centra en estudiar una herramienta informática que posee una funcionalidad que incrementa la potencialidad del ámbito investigativo y en ese aspecto, el estudio de la herramienta Map reduce servirá para capacitar a profesionales y académicos en la selección y aplicación eficiente de herramientas tecnológicas para el procesamiento distribuido en un mundo cada vez más orientado al procesamiento de grandes cantidades de datos.

La investigación busca ampliar el conocimiento en el procesamiento de Big Data, ya que en la actualidad se presenta como una técnica altamente innovadora para la creación de aplicaciones web destinadas a empresas de gran complejidad y en constante cambio.

Actualmente las grandes empresas requieren a personal especializado para manejar los grandes volúmenes de datos que poseen en sus sistemas de información, ya que un buen procesamiento de esos datos pueden ser aprovechados para realizar estrategias de marketing exitosas, crear nuevos productos y servicios e inclusive poder ofrecer publicidad personalizada para cada usuario, pero para ello el primer paso es contar con el conocimiento adecuado sobre herramientas para procesar estos conjuntos de datos y las utilidades en las que pueden ser aprovechadas.

El estudio de Map Reduce en el contexto de Big Data es esencial para comprender y aprovechar plenamente su potencial, lo que conlleva ventajas competitivas y oportunidades

de desarrollo en la gestión y análisis de datos a gran escala, este conocimiento puede ser aprovechado por los distintos profesionales en el área de Big Data para ahorrar recursos en el procesamiento de estos grandes volúmenes de datos.

## OBJETIVOS

### OBJETIVO GENERAL

- Realizar un estudio que identifique situaciones óptimas para el uso de Map Reduce.

### OBJETIVOS ESPECIFICOS

- Analizar los fundamentos teóricos de la herramienta Map Reduce y su utilización en la Big Data.
- Identificar las ventajas y desventajas del uso de la herramienta Map Reduce en el contexto de la Big Data.
- Establecer un ambiente de prueba para determinar el ecosistema ideal para la utilización de la herramienta Map Reduce.

## LÍNEA DE INVESTIGACIÓN

La línea de investigación que corresponde al presente estudio se denomina: Sistemas de información y comunicación, emprendimiento e innovación, la misma que va enmarcada con la sub línea de investigación de la carrera, la cual es Redes y Tecnologías inteligentes de hardware y software.

La presente investigación surge a partir de los conocimientos adquiridos durante todo el proceso académico en la carrera de sistemas de información; este tema se encuentra relacionado a la sub línea de investigación de la carrera, la cual se denomina Redes y Tecnologías inteligentes de Hardware y software, porque profundiza sobre la temática de Big Data, lo cual no solo forma parte de la carrera, sino que su utilización es sumamente importante en la actualidad, así que abordarlo puede ser muy significativo para la sociedad.

El presente estudio titulado "Estudio de la herramienta Map Reduce y su utilización en Big Data" tiene como objetivo principal llevar a cabo un análisis exhaustivo de la herramienta Map Reduce y su aplicación en el contexto de Big Data. El propósito central de este estudio es comparar y evaluar cómo se utiliza la herramienta Map Reduce en el procesamiento de grandes volúmenes de datos en el ámbito de Big Data. El objetivo es determinar la eficacia y eficiencia de Map Reduce para abordar los desafíos asociados con el procesamiento y análisis de grandes conjuntos de datos.

En resumen, este estudio de caso tiene como objetivo brindar un análisis de la herramienta Map Reduce y su aplicación en el procesamiento de Big Data. Al explorar sus características, ventajas y limitaciones, se proporcionará un recurso valioso para aquellos que se dedican a abordar desafíos de procesamiento de datos a gran escala.

## MARCO CONCEPTUAL

## CONCEPTO Y CARACTERÍSTICAS DE BIG DATA.

La Big Data son grandes cantidades de datos que son tan vastos y complejos que las herramientas tradicionales de administración y análisis de datos no son suficientes para manejarlos eficazmente. Estos conjuntos de datos se caracterizan por su volumen masivo, alta velocidad de generación y diversidad en términos de formatos. Cuando se realiza un análisis relacionado a la Big Data, esto busca extraer información valiosa y conocimientos significativos que puedan conducir a la toma de decisiones informadas y a la identificación de patrones y tendencias (Montoya & Gil, 2018).

Según (Gómez, 2021), entre las características más importantes sobre la Big Data, se tienen los siguientes elementos:

- **Volumen:** El volumen son los datos numerosos que se recopilan. La Big Data se distingue por la inmensa magnitud de los conjuntos de datos involucrados, que a menudo superan las capacidades de las bases de datos y herramientas tradicionales.
- **Variedad:** La variedad se refiere a la diversidad de tipos y formatos de datos. La Big Data abarca datos como bases de datos y hojas de cálculo, texto, imágenes, videos y datos como datos JSON o XML.
- **Velocidad:** La velocidad hace referencia a la rapidez con la que se generan y recopilan los datos. En entornos de Big Data, la información puede fluir a una velocidad exponencial, lo que requiere sistemas capaces de procesar y analizar datos en tiempo real.

- **Veracidad:** La veracidad se refiere a la confiabilidad y precisión de los datos. Dado que los datos pueden provenir de múltiples fuentes y ser generados en diversos contextos, es crucial garantizar su calidad y exactitud para obtener resultados válidos.
- **Valor:** El valor es la capacidad de extraer información significativa y conocimientos útiles de los datos. Los análisis que se realizan a estos conjuntos de datos, buscan identificar patrones, tendencias y relaciones que puedan generar ventajas competitivas y permitir la toma de decisiones informadas.
- **Volatilidad:** La volatilidad se refiere a la naturaleza cambiante de los datos a lo largo del tiempo. Los datos pueden ser efímeros y evolucionar rápidamente, lo que requiere un enfoque dinámico en su captura y análisis.

## IMPORTANCIA DEL PROCESAMIENTO DISTRIBUIDO EN EL CONTEXTO DE BIG DATA.

El procesamiento distribuido en el contexto de Big Data es sumamente importante para enfrentar los desafíos que ha ocasionado la nueva era digital (Almora, 2018). Esta actividad permite dividir tareas en partes más pequeñas para poder gestionar los grandes volúmenes de datos.

La escalabilidad proporciona la capacidad de adaptarse al crecimiento de datos y la eficiencia en tiempo de procesamiento se logra al realizar tareas en paralelo, reduciendo el tiempo necesario para análisis complejos (Campo & Cruz, 2019).

En resumen, el procesamiento distribuido es muy importante en el procesamiento de Big Data, proporcionando la capacidad de extraer conocimientos significativos y tomar decisiones en diversos sectores y aplicaciones.

## DEFINICIÓN Y ORIGEN DE LA HERRAMIENTA MAPREDUCE.

Map Reduce es un paradigma de programación y un modelo de procesamiento distribuido diseñado con el propósito de manipular y analizar eficientemente los grandes volúmenes de datos.

Map Reduce fue presentado por Google en un artículo seminal en 2004, el término "Map Reduce" refleja la estructura y las operaciones fundamentales involucradas en el procesamiento: la fase de "Map" implica aplicar una función a cada elemento de los datos y la fase de "reduce" combina los resultados intermedios para obtener la salida final (Cisneros & Liliana, 2019).

La génesis de Map Reduce surgió de la necesidad de abordar los desafíos de procesamiento distribuido y escalable en el entorno de Big Data, permitiendo a las organizaciones procesar, analizar y obtener conocimientos de datos masivos de manera efectiva (Linares, 2019).

## PRINCIPIOS BÁSICOS DE FUNCIONAMIENTO DE MAPREDUCE.

Map Reduce es una herramienta de procesamiento de datos ampliamente utilizado en el contexto de Big Data, gracias a las ventajas que ofrece y a su eficiencia y escalabilidad, los principios básicos de funcionamiento de esta herramienta están centrados en la distribución de tareas y el procesamiento paralelo para manejar grandes conjuntos de datos de manera efectiva.

La herramienta Map Reduce se beneficia de la tolerancia a fallos al realizar copias redundantes de los datos y tareas en varios nodos, lo que garantiza la integridad y disponibilidad de los datos, incluso en caso de fallos de hardware o software (Sarabia, 2020).

Esta herramienta ofrece múltiples ventajas en términos de escalabilidad, ya que se puede agregar fácilmente más nodos para manejar conjuntos de datos aún más grandes.

En resumen, Map Reduce se basa en la distribución de tareas, procesamiento paralelo y tolerancia a fallos para abordar eficazmente el procesamiento de Big Data, dividiendo las tareas complejas en tareas más pequeñas que se ejecutan de manera eficiente.

## ESTRUCTURA Y FLUJO DE TRABAJO DE MAPREDUCE: ETAPAS DE MAPEO Y REDUCCIÓN.

(Padilla, 2019) indica que Map Reduce se compone de dos fases principales: la fase de mapeo (Map) y la fase de reducción (reduce).

**Fase de Mapeo (Map):** En esta etapa, los datos de entrada se dividen en fragmentos más pequeños y se asignan a múltiples nodos del clúster. Cada nodo realiza una función de "Map" que procesa los datos de entrada y produce pares clave-valor intermedios. Estos pares representan la transformación de los datos originales en formatos que serán utilizados en la fase de reducción.

**Fase de Reducción (Reduce):**

En la fase de reducción, los pares clave-valor intermedios generados en la etapa de mapeo se agrupan por clave y se envían a nodos de reducción específicos. Cada nodo de reducción realiza una función de "reduce" que procesa los valores asociados a una clave particular. El resultado final que se obtendrá es una serie de pares clave-valor reducidos que representan la salida final del proceso.

**Flujo de Trabajo de Map Reduce:**

El flujo de trabajo de Map Reduce sigue una secuencia específica:

**División de Datos:** Los datos de entrada se dividen en fragmentos más pequeños para permitir el procesamiento paralelo en nodos individuales del clúster.

**Fase de Mapeo (Map):** Los nodos de mapeo procesan sus fragmentos de datos, aplicando una función definida por el usuario. Cada nodo genera pares clave-valor intermedios.

**Ordenamiento y Agrupamiento:** Los pares clave-valor intermedios se ordenan y agrupan según sus claves, lo que permite que los nodos de reducción procesen valores relacionados juntos.

**Fase de Reducción (Reduce):** Los nodos de reducción procesan grupos de pares clave-valor asociados a una clave específica. El usuario va a definir una función de "reduce" que toma los valores asociados a una clave y produce un conjunto reducido de pares clave-valor, ejemplo: 1-" Zapato".

**Generación de Resultados:** Los resultados finales de las fases de reducción se combinan para formar la salida final del proceso Map Reduce.

**CURVA DE APRENDIZAJE INICIAL Y NECESIDAD DE EXPERTISE.**

El uso de Map Reduce implica una curva de aprendizaje inicial. Los usuarios deben familiarizarse con los conceptos de "Map" y "reduce", entender cómo diseñar tareas que se adapten al modelo y aprender a trabajar con los sistemas distribuidos.

(Zeebaree, y otros, 2020), mencionan que, dado que Map Reduce es una herramienta poderosa pero especializada, su uso eficaz demanda un conocimiento profundo. La necesidad de su experiencia se origina en varios aspectos:

- **Diseño y Optimización de Tareas:** Los usuarios deben comprender cómo diseñar tareas de "Map" y "reduce" que se ajusten a las necesidades específicas de procesamiento y minimicen el movimiento de datos entre nodos.
- **Gestión de Datos Distribuidos:** Se requiere conocimiento sobre cómo manejar y distribuir los datos en el clúster, garantizando su acceso eficiente y reduciendo los cuellos de botella en el procesamiento.
- **Identificación de Cuellos de Botella:** Expertise en el monitoreo y diagnóstico de problemas de rendimiento, como cuellos de botella en el clúster o en las tareas individuales, es esencial para optimizar el rendimiento.
- **Selección de Algoritmos:** Elegir el algoritmo adecuado para aprovechar al máximo las capacidades de Map Reduce y alcanzar los objetivos de procesamiento es una habilidad clave.
- **Resolución de Problemas:** Ante desafíos inesperados, los expertos en Map Reduce deben tener la capacidad de diagnosticar y resolver problemas de manera eficiente para evitar interrupciones en el procesamiento.

## POSIBLES PROBLEMAS DE LATENCIA Y OVERHEAD EN EL PROCESAMIENTO.

Aunque Map Reduce ha demostrado ser efectivo en el procesamiento paralelo y distribuido de datos, también puede enfrentar problemas de latencia y overhead en ciertas situaciones. Aquí hay algunos posibles problemas relacionados con la latencia y el overhead en el procesamiento de Map Reduce en Big Data, según lo que mencionan (Rajput & Mehta, 2018):

**Latencia en el inicio del trabajo:** Cuando se inicia un trabajo de Map Reduce, puede haber una latencia significativa debido a la necesidad de distribuir el código, los datos y configurar

el entorno de ejecución en los nodos de procesamiento. Esto puede ser especialmente problemático para trabajos pequeños que no aprovechan completamente la capacidad de procesamiento paralelo.

**Latencia de la fase de Map:** La fase de Map implica la ejecución de tareas de mapeo en paralelo en diferentes nodos. Sin embargo, si hay desequilibrios en la distribución de datos o en la complejidad del proceso de mapeo, algunos nodos pueden terminar antes que otros, lo que aumenta la latencia total.

**Overhead de comunicación:** Durante las fases de Map y Shuffle, los nodos deben comunicarse constantemente para intercambiar datos intermedios. Esta comunicación puede generar overhead significativo, especialmente cuando los datos intermedios deben ser transferidos entre nodos a través de la red. Un exceso de comunicación puede ralentizar el procesamiento y aumentar la latencia general.

**Latencia de la fase de Reduce:** Similar a la fase de Map, la fase de Reduce también puede sufrir latencia debido a desequilibrios en los datos de entrada o a la complejidad de las funciones de reducción. Si un grupo de nodos Reduce termina antes que los demás, se subutiliza el potencial de procesamiento paralelo.

**Overhead de lectura/escritura de datos:** La lectura y escritura de datos en sistemas de almacenamiento distribuido puede generar overhead significativo. Map Reduce implica movimientos intensivos de datos entre los nodos de procesamiento y los sistemas de almacenamiento, lo que puede resultar en latencia adicional.

**Fallas y reintentos:** En entornos distribuidos, los nodos pueden fallar por diversas razones, lo que puede causar retrasos debido a reintentos y reubicación de tareas en nodos sanos. Esto introduce latencia adicional y puede afectar el rendimiento general del trabajo.

**Overhead de planificación:** El administrador de recursos y planificación debe asignar tareas a los nodos de manera eficiente. Si la planificación no se realiza de manera óptima, puede haber overhead adicional debido a la asignación ineficiente de recursos, ocasionando un problema gravísimo.

**Escalabilidad limitada:** A medida que el tamaño de los datos y la complejidad del trabajo aumentan, Map Reduce puede enfrentar problemas de escalabilidad. El procesamiento de grandes cantidades de datos puede llevar a latencias más largas debido a la necesidad de distribuir y coordinar tareas en un gran número de nodos (Marchant, 2018).

## CASOS DE USO EN DIFERENTES INDUSTRIAS

**Tabla 1.**
Comparativa entre las ventajas y desventajas de utilizar la herramienta Map Reduce en diferentes industrias.

| Segmento | Casos de éxito | Deficiencias |
|---|---|---|
| Industria financiera | Detección de fraudes financieros. | "Es algo lento y en el área financiera puede ser un inconveniente, ya que se necesita la información lo más rápido posible." [a] |
| Salud | Análisis de genes y enfermedades para mejorar los medicamentos. | "En el área de la salud se requiere la integración de distintas variedades de datos y para map reduce puede ser complicado." [a] |
| Comercio electrónico | Permite conocer al cliente y personalizar ofertas exclusivas. | "Los datos recolectados de este sector son muy variables, lo cual puede dificultar su procesamiento en Map Reduce." [a] |
| Entretenimiento | Permite analizar a los espectadores para recomendar contenido que mejore la experiencia de usuario. | "No puede procesar datos en tiempo real, solo datos estáticos." [a] |
| Agricultura | Optimización de riego y fertilización de cultivo. | "Los datos recolectados de la agricultura son muy variables, lo cual puede dificultar su procesamiento en Map Reduce." [a] |
| Transporte | Análisis de rutas menos problemáticas y más eficientes | "Los enormes volúmenes de datos de este sector hacen que el mantenimiento de la infraestructura donde opera Map Reduce sea costosa." [a] |
| Telecomunicaciones | Predecir picos de llamadas, para controlar esta demanda. | "No puede procesar datos en tiempo real, solo datos estáticos." [a] |
| Educación | Permite conocer al estudiante para personalizar los cursos y su experiencia educativa. | "Requiere infraestructura que muchas instituciones no pueden pagar por sus recursos limitados." [a] |

Nota. [a]Recalde, (2018). Esta tabla muestra como la herramienta Map Reduce ha sido aprovechada por ciertos sectores, sin embargo, también muestra las limitaciones que posee para estas industrias.

En la tabla 1, se muestran algunos casos de uso de Map Reduce en diversas industrias, en las cuales se demuestra su versatilidad y su capacidad para abordar ciertos desafíos, esta

herramienta ha permitido a las organizaciones en diferentes sectores aprovechar al máximo el potencial de Big Data para tomar decisiones adecuadas y obtener ventajas competitivas en sus respectivos campos.

Map Reduce, pese a ofrecer grandes ventajas, tiene un problema común en muchos sectores, la cual está ligado a los altos costos de mantenimiento en la infraestructura que necesita para funcionar correctamente, es por ello que se requiere una evaluación exhaustiva de la industria donde se desea adoptar la herramienta en relación a sus objetivos específicos y sus necesidades en particular, para decidir si el uso de esta herramienta es adecuado o no para sus actividades.

## EJEMPLOS CONCRETOS DE CÓMO MAPREDUCE HA SIDO UTILIZADO PARA ABORDAR DESAFÍOS ESPECÍFICOS.

La herramienta Map Reduce ha demostrado su versatilidad en el procesamiento de Big Data al abordar desafíos específicos en diversas industrias y aplicaciones. A continuación, se presentan ejemplos concretos que ilustran cómo Map Reduce ha sido utilizado para superar retos particulares.

### 1. Análisis de Redes Sociales:

En plataformas de redes sociales como Facebook y Twitter, Map Reduce se utiliza para analizar patrones de interacción entre usuarios. El proceso de "Map" puede analizar millones de mensajes para extraer menciones, hashtags y conexiones entre usuarios, mientras que la fase de "reduce" consolida esta información para identificar tendencias, influenciadores y relaciones entre comunidades.

### 2. Motor de Búsqueda en Línea:

Los motores de búsqueda, como Google, utilizan Map Reduce para indexar y clasificar páginas web. La fase de "Map" puede procesar la información de las páginas y extraer palabras clave, mientras que la fase de "reduce" calcula la relevancia de las páginas para las consultas de búsqueda, mejorando la calidad de los resultados.

**3. Detección de Fraude Financiero:**

En la industria financiera, Map Reduce se utiliza para analizar transacciones en tiempo real y detectar patrones de comportamiento sospechoso que podrían indicar fraude. El procesamiento paralelo permite identificar transacciones anómalas entre millones de registros, agilizando la detección y reduciendo el riesgo.

**4. Optimización de Publicidad en Línea:**

En la publicidad en línea, Map Reduce se utiliza para analizar el comportamiento del usuario y ajustar las estrategias de publicidad en tiempo real. El procesamiento distribuido permite analizar datos de clics y preferencias para personalizar las campañas publicitarias y maximizar el retorno de la inversión.

**5. Análisis de Texto y Sentimiento:**

En el análisis de texto, Map Reduce se aplica para procesar grandes cantidades de contenido textual y determinar el sentimiento asociado con él. Las fases de "Map" y "reduce" pueden analizar palabras clave y patrones lingüísticos para identificar tendencias y opiniones en los datos.

**IMPACTO EN LA TOMA DE DECISIONES Y EFICIENCIA OPERATIVA EN LAS ORGANIZACIONES.**

Map Reduce ha tenido un impacto significativo en las organizaciones al permitir el procesamiento eficiente y escalable de grandes volúmenes de datos. Su influencia se extiende más allá del ámbito técnico, afectando directamente la toma de decisiones y la eficiencia operativa. (Díaz, 2019), menciona que se exploran los efectos de Map Reduce en estos aspectos:

**Información Accesible y Accionable:** Map Reduce permite analizar y extraer información significativa de datos masivos en tiempo real. Esto habilita la generación de conocimientos más profundos y precisos, lo que respalda una toma de decisiones informada y basada en datos.

**Análisis en Tiempo Real:** La capacidad de procesamiento distribuido y paralelo de Map Reduce facilita el análisis de datos en tiempo real. Las organizaciones pueden responder rápidamente a cambios en el entorno, lo que es crucial para la toma de decisiones ágiles y estratégicas.

**Detección de Patrones y Tendencias:** Map Reduce permite identificar patrones, tendencias y correlaciones en datos masivos. Esto puede llevar a insights previamente ocultos que respalden la identificación de oportunidades y desafíos.

**Personalización y Segmentación:** En la toma de decisiones relacionadas con marketing y ventas, Map Reduce posibilita la personalización de ofertas y la segmentación de clientes basada en análisis exhaustivos, mejorando la eficacia de las estrategias.

**Procesamiento Escalable:** Map Reduce permite escalar horizontalmente para manejar la creciente carga de trabajo sin degradar el rendimiento. Esto mejora la eficiencia operativa al evitar cuellos de botella y retrasos en el procesamiento.

**Reducción de Costos:** Al procesar grandes volúmenes de datos de manera eficiente, las organizaciones pueden reducir los costos de almacenamiento y procesamiento, ya que se aprovechan al máximo los recursos disponibles.

**Optimización de Recursos:** La capacidad de distribuir tareas en múltiples nodos permite utilizar eficientemente la capacidad de cómputo y memoria disponibles en el clúster, optimizando la utilización de recursos.

**Automatización de Procesos:** La automatización de procesos a través de Map Reduce puede eliminar la necesidad de tareas manuales repetitivas, aumentando la eficiencia y reduciendo la posibilidad de errores humanos.

**Rápida Generación de Informes:** Map Reduce acelera la generación de informes y análisis, permitiendo a las organizaciones tomar decisiones más rápidas y basadas en datos actualizados.

## INTEGRACIÓN DE MAPREDUCE EN ECOSISTEMAS DE BIG DATA (HADOOP, SISTEMAS DE ALMACENAMIENTO DISTRIBUIDO).

La integración de Map Reduce en los ecosistemas de Big Data, como Hadoop y otros sistemas de almacenamiento distribuido, es esencial para procesar y analizar eficientemente enormes conjuntos de datos. Map Reduce es un paradigma de programación que divide las tareas en etapas de "Map" y "Reduce", permitiendo la ejecución paralela en múltiples nodos (Vaca, 2018).

El framework de procesamiento Map Reduce coordina la distribución de tareas y datos entre los nodos del clúster, aprovechando su capacidad de procesamiento (Blanco, 2018). A

medida que los datos crecen exponencialmente, esta integración garantiza un rendimiento escalable y tolerante a fallos, permitiendo un análisis eficaz de Big Data.

## MARCO METODOLOGICO

La presente investigación tiene la necesidad de utilizar la metodología descriptiva, esta metodología tiene como objetivo la búsqueda y revisión de fuentes bibliográficas confiables y relevantes para respaldar un estudio y se utilizara debido a que la investigación necesita consultar diversas fuentes de información que se encuentran esparcidas en todo el internet, como las revistas científicas y trabajos de investigación académicas de todas partes del mundo.

La metodología a utilizar en esta investigación servirá para recopilar los trabajos realizados por diversos autores en relación al tema de la Big Data y la utilización de la herramienta Map Reduce dentro de ese contexto, con el propósito de tener información científica y comprobada que permita conocer exactamente la influencia de esta herramienta en la ya mencionada Big Data y asimismo poder realizar comparativas entre las ventajas y desventajas que produce esta herramienta mientras se la utiliza en la Big Data.

Cabe indicar que en esta fase de la investigación se describe la metodología que forma parte del estudio de la herramienta Map Reduce y su utilización en la Big Data, donde se investigarán las fortalezas y debilidades de la herramienta para generar conocimientos que permitan evaluar y considerar todos los aspectos posibles antes de decidir o no la implementación de Map Reduce, ya que, aunque permite procesar datos a grandes escalas, no es la solución ideal para todos los casos.

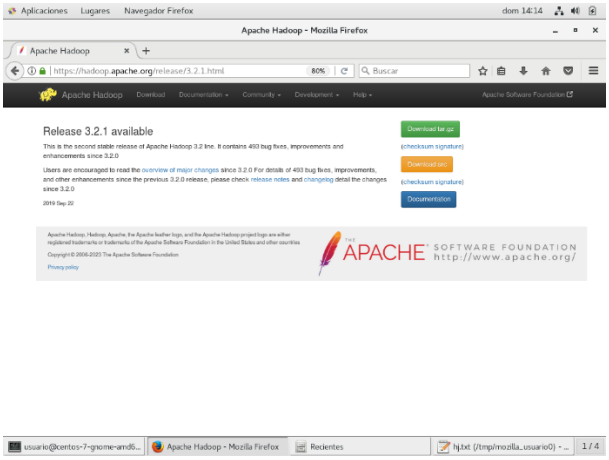
## RESULTADOS

En esta investigación se mostrará la instalación y un ejemplo de Map reduce. El sistema operativo donde se instalará todo será Centos7.

### Instalaciòn

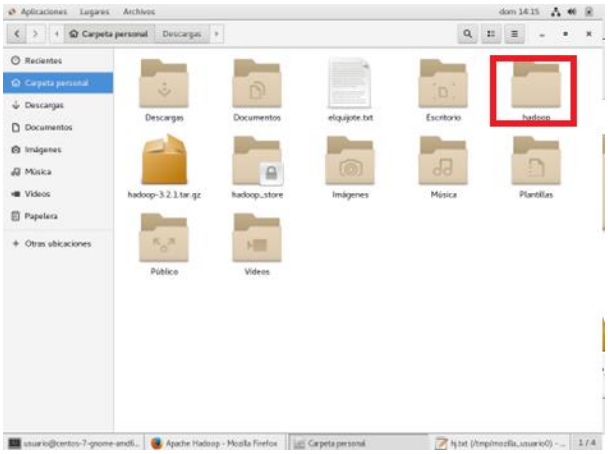
En primer lugar, serà necesario ir a la pagina de Apache Hadoop 3.2.1 y dar click en el boton verde de descargar tar.gz.

**Figura 1**
Descarga de Hadoop 3.2.1



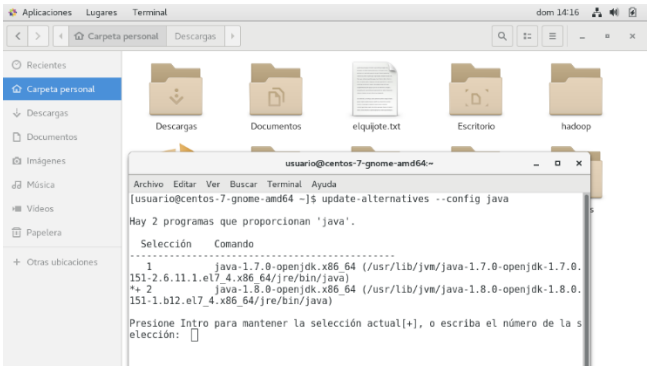

Posteriormente ese archivo descargado lo descomprimimos en la carpeta principal de nuestro sistema operativo.

**Figura 2**
Descompresión de Hadoop



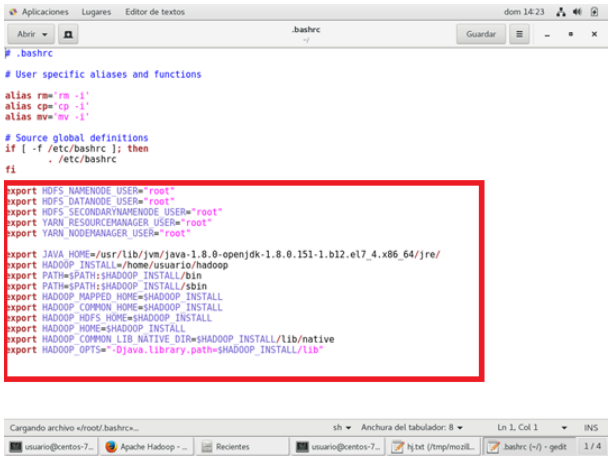

Luego, usamos el comando update-alternatives –config java para ver que programas de nuestro sistema operativo proporciona java, el cual será necesario para usar Map Reduce.

**Figura 3**
Verificar programa de Java



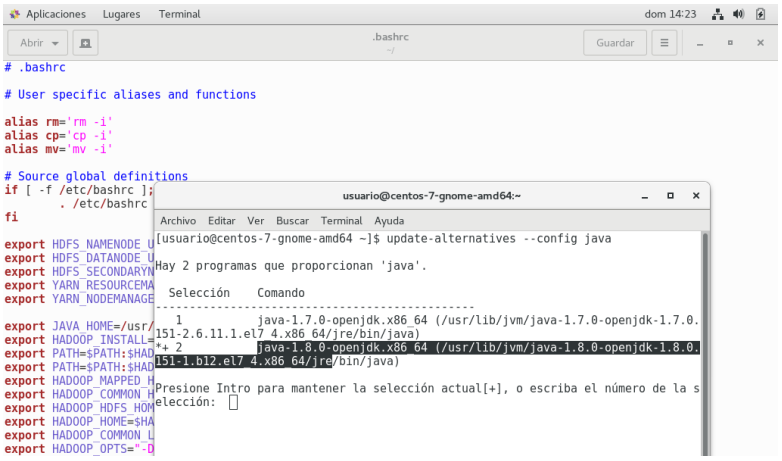

Abrimos otra terminal y ejecutamos el comando gedit ~/. bashrc, el cual nos abrirá el siguiente archivo en donde colocaremos las siguientes líneas.

**Figura 4**
Establecer las rutas para los archivos



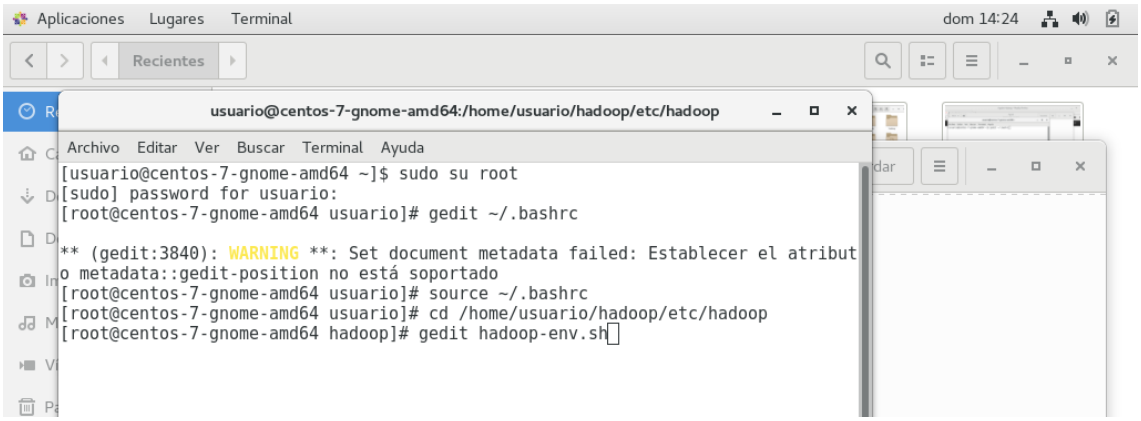

En JAVA_HOME se debe colocar la ruta de la aplicación de java (lo que se encuentra sombreado) que anteriormente habíamos consultado.

**Figura 5**
Versión de Java en Centos7



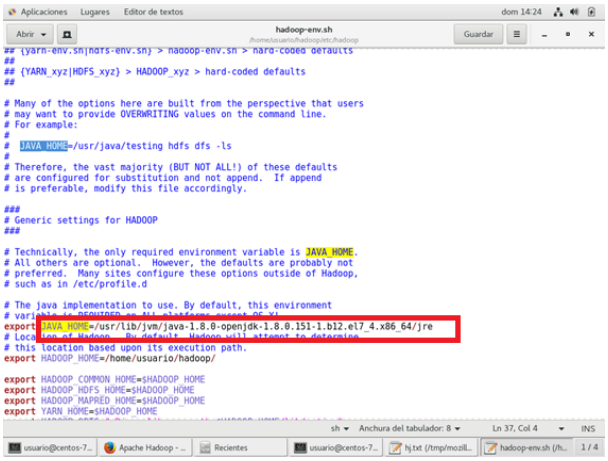

Ejecutaremos los siguientes comandos para que posteriormente se nos habrá un apartado para ingresar la ruta de JAVA_HOME.

**Figura 6**
Apertura del archivo hadoop-env.sh



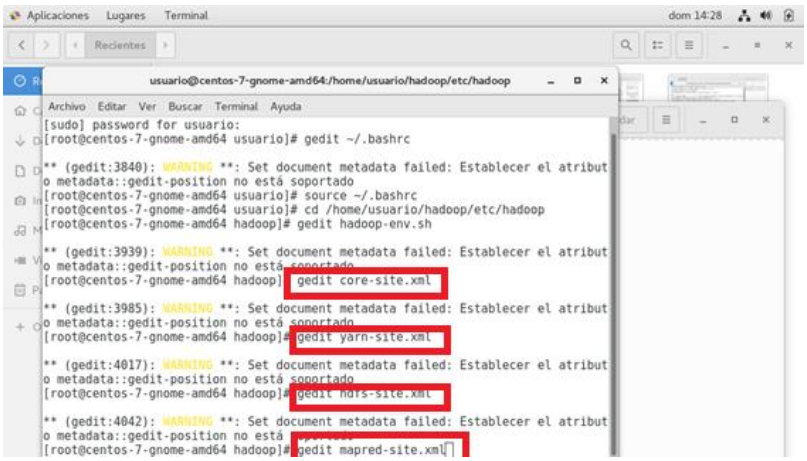

En este documento colocamos la misma ruta del Java_home anteriormente mostrado y guardamos el documento.

**Figura 7**
Cambio de ruta del Java_home



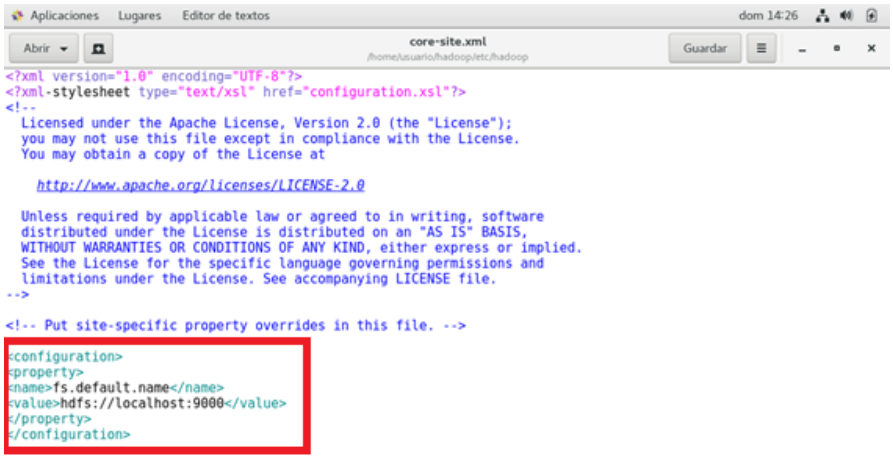

En la terminal, se modificarán algunos elementos de hadoop, con el comando: gedit y el nombre del archivo. Cuando se ejecuta el comando se abrirán los archivos correspondientes para modificarlos.

**Figura 8**
Modificación de archivos .xml



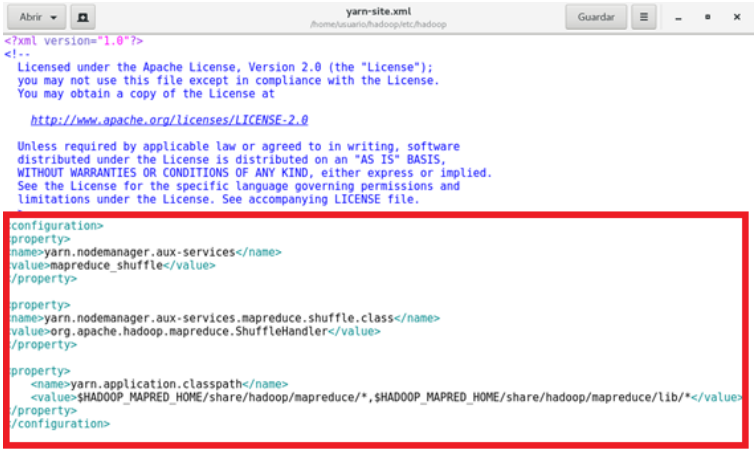

En core-site.xml, debe integrarse el siguiente código:

**Figura 9**
Archivo core-site.xml



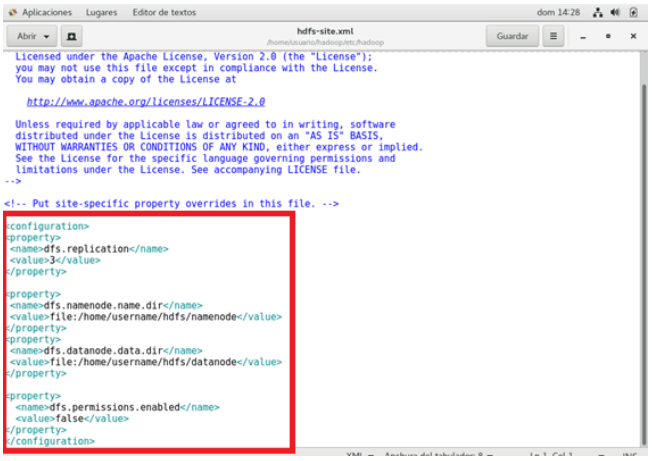

En yarm-site.xml, debe integrarse el siguiente código:

**Figura 10**
Archivo yarn-site.xml



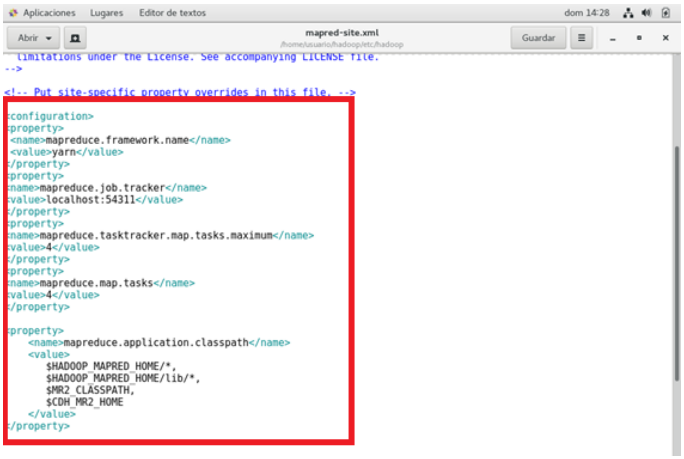

En hdfs-site.xml, debe integrarse el siguiente código:

**Figura 11**
Archivo hdfs-site.xml



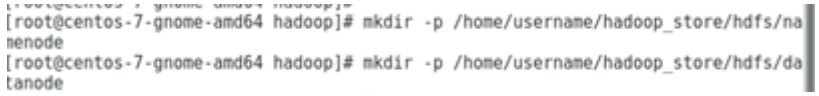

En mapred-site.xml, debe integrarse el siguiente código:

**Figura 122**
Archivo mapred-site.xml




Ejecutamos estos comandos para preparar los archivos de namenode y datenode.

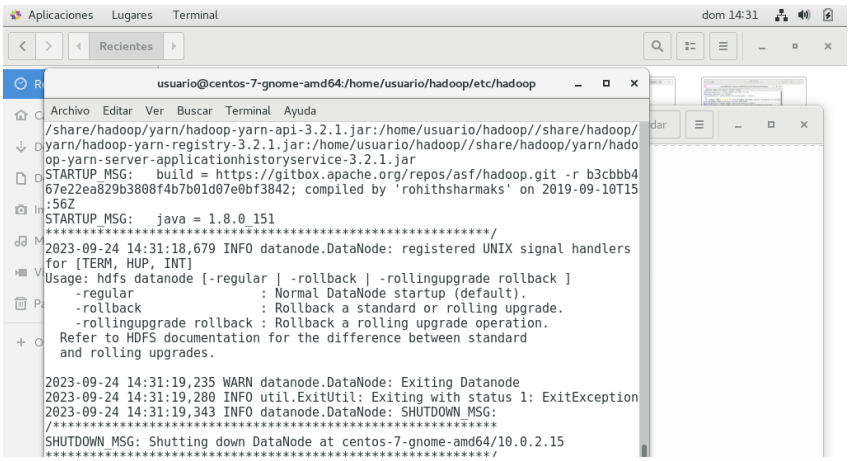
**Figura 13**
Ejecución de comandos namenode y datanode




Con los siguientes comandos activamos los servicios namenode y datanode.

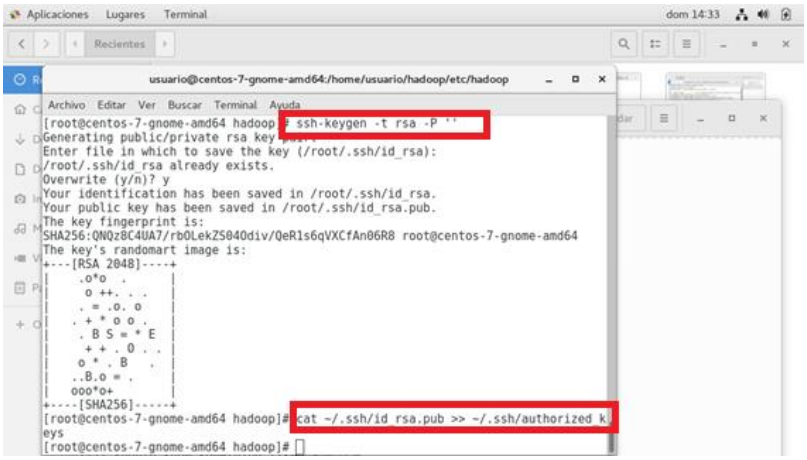
- hdfs namenode -format
- hdfs datanode -format

**Figura 14**
Activación de servicios Namenode y Datanode




Ejecutamos los comandos de ssh

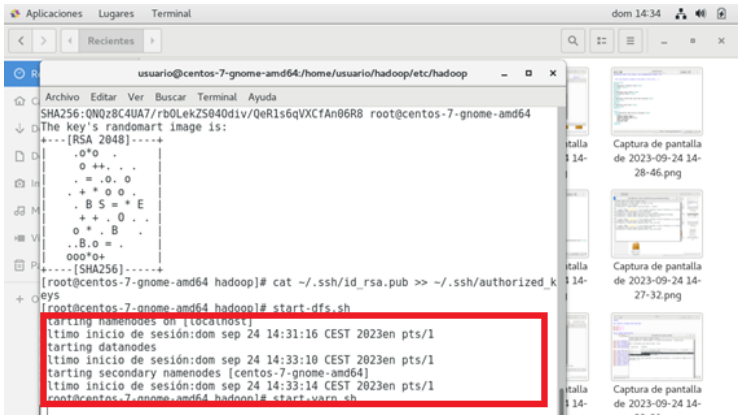
**Figura 15**
Ejecución de comandos ssh




Iniciamos los servicios con:

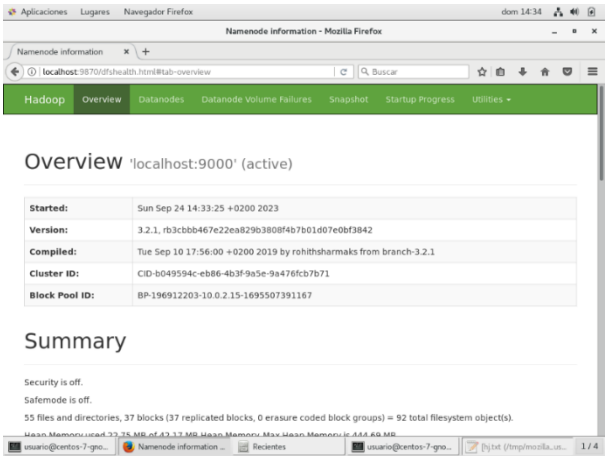
- start-dfs.sh
- start-yarn.sh

**Figura 16**
Activación de servicios yarn y dfs




Para verificar que, si funciona, se debe ir a un navegador y entrar al siguiente enlace: http://localhost:9870 y debe aparecer la siguiente pantalla.

**Figura 17**
Pantalla principal informativa de Hadoop



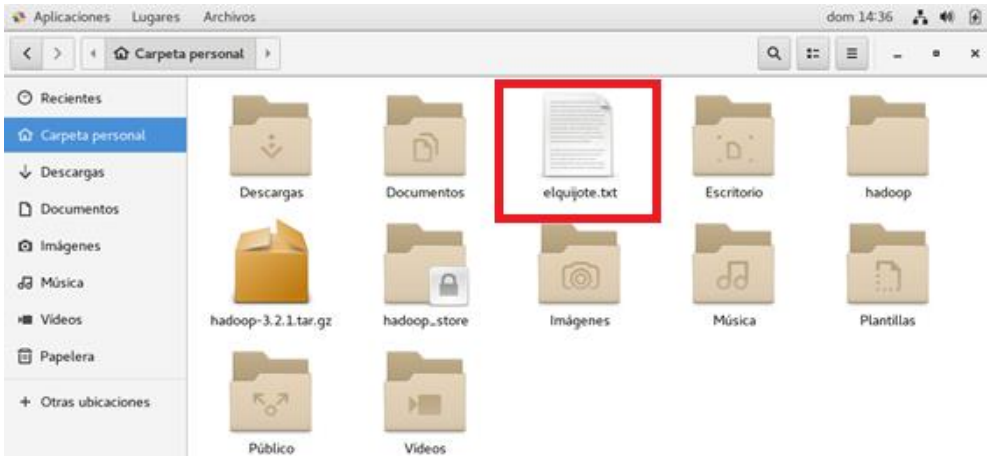

## Ejemplo usando Map Reduce.

**Figura 18**
Creación de directorio libros


Con este comando se crea una carpeta en el sistema de archivos de hadoop.

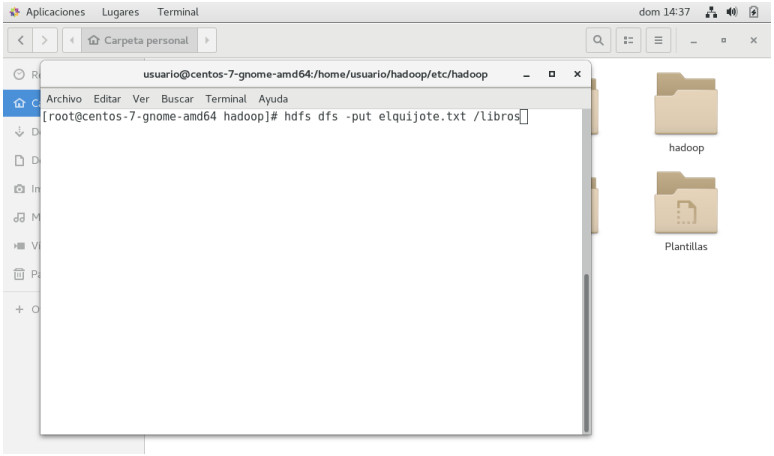
Para este ejemplo se necesita un archivo de texto, ya que con Hadoop verificaremos cuantas palabras se repiten dentro del archivo. En este caso se tiene el archivo llamado, “El quijote” en la carpeta personal del sistema.

**Figura 19**
Archivo a usar para el ejemplo



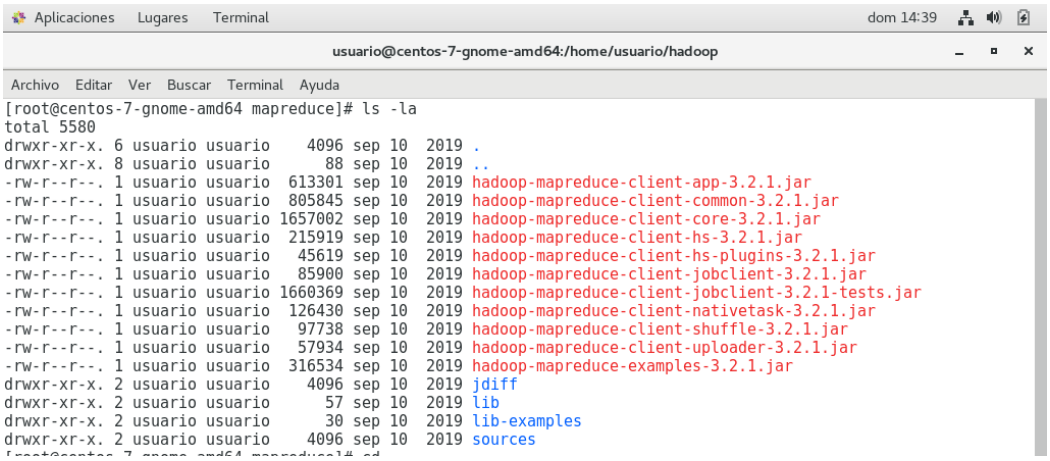

Ahora se sube el archivo usando el siguiente comando:

**Figura 20**
Subida de archivo al directorio



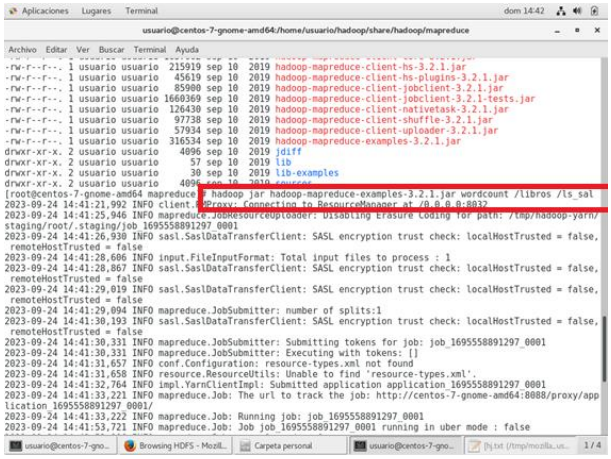

Usamos el comando cd share/hadoop/Map reduce para cambiar de directorio y usamos el comando ls -la para ver los archivos que tenemos en dicha carpeta. Los archivos de color rojo significan que son archivos de java.

**Figura 21**
Archivos de la carpeta Map reduce




Usamos el comando hadoop jar hadoop-mapreduce-examples.3.2.1.jar wordcount /libros /ls_sal, para procesar el archivo que tenemos en la carpeta libros, la palabra “wordcount”, significa que usaremos ese ejemplo de los archivos de jara y el procesamiento saldrá en la carpeta ls_sal.

**Figura 22**
Procesamiento de archivos



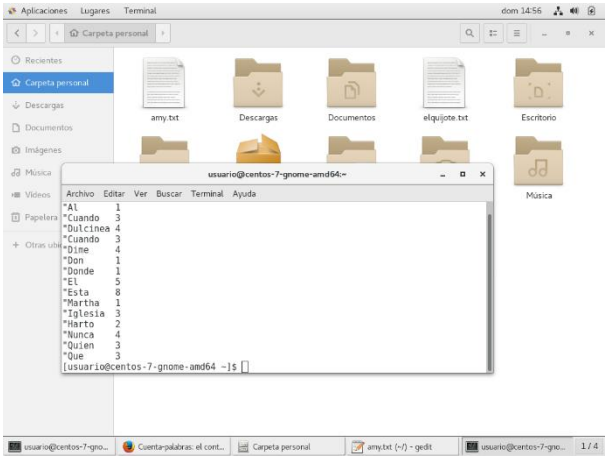

Para visualizar el procesamiento usamos el siguiente comando:

**Figura 23**
Salida del archivo previamente procesado

```
[root@centos-7-gnome-amd64 mapreduce]# hdfs dfs -ls /ls_sal/part-r-00000
```

Finalmente, podemos ver el conteo de palabras, y es así como se pueden procesar los datos utilizando Map Reduce.

**Figura 24**
Visualización de resultados.




El hecho de procesar esta información usando Map Reduce, reduce tiempos y recursos que pueden ser destinados a otras actividades, sin embargo, se puede observar que la instalación y ejecución de actividades usando la herramienta de Map Reduce es algo compleja y requiere ciertos conocimientos especiales.

## DISCUSION DE RESULTADOS

En la etapa de prueba de Map Reduce se pudo evidenciar el funcionamiento de la herramienta, en donde se observó el gran trabajo que realiza en el procesamiento de los datos y en la rapidez que lo realiza, sin embargo, el código necesario para realizar dicha actividad fue un poco complejo para ser un pequeño ejemplo.

Esta herramienta funciona procesando la información en tareas más pequeñas que trabajan en conjunto para obtener el resultado esperado, pero en la fase de prueba se puede analizar que para usar esta herramienta se requiere conocimientos algo avanzados sobre el funcionamiento de Map Reduce para llevar a cabo la actividad de procesamiento de grandes volúmenes de datos, sin embargo, los recursos y el tiempo que reduce en procesar los datos es de mucha utilidad.

Identificar las ventajas y desventajas del uso de Map Reduce es esencial para comprender su impacto en el procesamiento de Big Data. Las ventajas incluyen escalabilidad horizontal, tolerancia a fallos, programación sencilla y eficiente utilización de recursos. Estas características hacen que Map Reduce sea ideal para aplicaciones que requieren el procesamiento de grandes conjuntos de datos de manera eficiente y confiable. Sin embargo, las desventajas incluyen un alto costo de inicio para tareas pequeñas, complejidad en ciertos algoritmos y limitaciones de latencia. Estos aspectos hacen que Map Reduce no sea la elección óptima para todas las situaciones, especialmente cuando se trata de tareas que no se ajustan bien a su modelo por lotes.

Diversas industrias han adoptado Map Reduce para abordar los desafíos de procesamiento de Big Data. Por ejemplo, en la industria de la publicidad en línea, Map Reduce se utiliza para analizar y procesar los datos de clics y visualizaciones de anuncios, permitiendo la personalización de la publicidad dirigida. En la atención médica, se aplica para analizar datos

clínicos y de pacientes para identificar patrones y tendencias que puedan mejorar los diagnósticos y tratamientos. La industria financiera también se beneficia al analizar transacciones financieras para detectar fraudes y predecir movimientos del mercado.

En resumen, Map Reduce es una herramienta donde su ecosistema principalmente se basa más para las grandes industrias en donde la latencia no sea un problema y en donde los datos en su mayoría sean estáticos y puedan ser procesados con calma y tranquilidad, para obtener los mejores resultados.

## CONCLUSIONES

- El estudio de la herramienta Map Reduce y su uso en el contexto de Big Data revela una gran capacidad de dividir tareas complejas en etapas más manejables, lo cual ha demostrado ser muy eficaz para agilizar el procesamiento y análisis de grandes volúmenes de datos.
- La herramienta Map Reduce en la Big Data ha permitido identificar un conjunto de ventajas y desventajas cruciales. Su eficiencia en el procesamiento escalable y su adaptabilidad son puntos fuertes que pueden llevar a mejoras significativas en el manejo de datos a gran escala. Sin embargo, la curva de aprendizaje inicial y la necesidad de optimización pueden representar desafíos que los profesionales deben abordar para maximizar los beneficios de esta herramienta.
- El escenario ideal de Map Reduce se centra en las grandes industrias en donde no se requiera procesar datos en tiempo real y en donde la latencia no sea un problema grave, ya que esta herramienta no está destinada para industrias donde se requiera la información en cuestión de segundos, sino más bien en lugares donde se requiera analizar datos históricos que permitan predecir lo que sucedería en un futuro.

## RECOMENDACIONES

- Se recomienda que los investigadores se sumerjan en una formación sólida sobre los conceptos clave de Map Reduce y cómo se aplican en escenarios de Big Data. Además, es importante combinar esta formación teórica con ejercicios prácticos y proyectos de implementación para comprender mejor cómo funciona la herramienta en situaciones reales.
- Se recomienda recopilar más ejemplos concretos de cómo Map Reduce ha sido aplicado en sectores como finanzas, salud, marketing, entre otras. Esta diversidad de casos proporcionará una comprensión más completa de las aplicaciones prácticas y los beneficios reales que la herramienta puede ofrecer en diferentes contextos.
- Es recomendable que en otras investigaciones se puedan realizar ambientes de prueba en el procesamiento de Map reduce en comparación con otras herramientas similares para analizar los factores de tiempo y recursos en relación a aplicaciones análogas.

**BIBLIOGRAFÍA**


Almora, N. (2018). *ANALISIS Y USOS DEL BIG DATA APLICADO EN LA UNIVERSIDAD NACIONAL "SAN LUIS GONZAGA" DE ICA: CASO FACULTAD DE INGENIERÍA DE SISTEMAS.* UNIVERSIDAD NACIONAL "SAN LUIS GONZAGA DE ICA", Perú. Obtenido de https://repositorio.unica.edu.pe/bitstream/handle/20.500.13028/3095/An%C3%A1lisis%20y%20usos%20del%20Big%20data%20aplicado%20en%20la%20universidad%20nacional%20%E2%80%9CSan%20Luis%20Gonzaga%E2%80%9D%20de%20Ica%20caso%20Facultad%20de%20Ingenier%C3%ADa%20de

Blanco, C. (2018). *MARCO DE TRABAJO PARA LA IMPLEMENTACIÓN DE BIG DATA ANALYTICS EN EL CONTEXTO ESPECÍFICO DEL ÁREA DE SALUD.* Universidad de Palermo, Argentina. Obtenido de https://dspace.palermo.edu/dspace/bitstream/handle/10226/2128/Tesis-BlancoC-2015.pdf?sequence=1&isAllowed=y

Campo, J., & Cruz, J. (2019). *DISEÑO E IMPLEMENTACIÓN DE UNA BASE DE DATOS DISTRIBUIDA HOMOGÉNEA EN EL PROTOTIPO DE UN SISTEMA DE CONTROL DE ACCESO.* UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS, Bogotá. Obtenido de https://repository.udistrital.edu.co/bitstream/handle/11349/24712/Campo%20Romero%20Juli%E0n%20David%202019.pdf?sequence=1

Cisneros, G., & Liliana, S. (2019). *a influencia del Big data en el desempeño del contador: Análisis para el caso ecuatoriano.* UNIVERSIDAD ESTATAL DE MILAGRO, Milagro. Obtenido de https://repositorio.unemi.edu.ec/bitstream/123456789/5081/1/2.TESIS%20BIG%20DATA.pdf

Díaz, E. (2019). DESARROLLO DE UN SISTEMA DOMÓTICO BASADO EN IOT PARA LA SEGURIDAD RESIDENCIAL Y MEJORAMIENTO DEL CONSUMO ENERGÉTICO, APLICANDO CONCEPTOS DE BIG DATA. *Trabajo de grado.* UNIVERSIDAD MILITAR NUEVA GRANADA. Obtenido de https://repository.unimilitar.edu.co/handle/10654/32454

Gómez, Á. (2021). *BIG DATA, UN SISTEMA DE GESTIÓN DE DATOS.* Tecana American University, New York. Obtenido de https://tauniversity.org/sites/default/files/articulo_big_data_de_angel_gomez_degraves.pdf

Linares, C. (2019). *IMPLEMENTACIÓN DE UN SISTEMA DE BIG DATA APLICADO A LA MIGRACION DE DATOS BAJO LA DISTRIBUCION CLOUDERA CON APACHE HADOOP, EN EL BANCO I N TERBANK.* Universidad Tecnologica del Perú, Lima. Obtenido de https://repositorio.utp.edu.pe/handle/20.500.12867/1944

Marchant, T. (2018). *MÉTODOS DE ANÁLISIS PARA BIG DATA Y SU PARTICIPACIÓN EN LA INDUSTRIA: ESTUDIO APLICADO A LA PREVENCIÓN DE FALLOS EN EMPRESAS FERROVIARIAS.* UNIVERSIDAD TECNICA FEDERICO SANTA MARIA, Valparaíso. Obtenido de https://repositorio.usm.cl/bitstream/handle/11673/43410/3560900257438UTFSM.pdf?sequence=1&isAllowed=y

Montoya, L., & Gil, G. (2018). Actualidad e importancia de la implementación de Big Data utilizando las herramientas Hadoop y Spark. *Lámpsakos*, 67-72.

Padilla, C. (2019). Big Data, una herramienta para apoyar en decisiones del sector hotelero en Quito-Ecuador. *INNOVA Research Journal*, 80-88.

Rajput, K., & Mehta, V. (2018). An Analytical Study of Hadoop and Its Components. *International Journal for Scientific Research & Development*. Obtenido de https://www.researchgate.net/publication/331535483_An_Analytical_Study_of_Hadoop_and_Its_Components

Recalde, S. (2018). ANÁLISIS Y PROPUESTA DE UNA HERRAMIENTA BUSINESS INTELLIGENCE QUE PERMITA MEJORAR LA TOMA DE DECISIONES GERENCIALES EN LA EMPRESA SOLDENEG SOLUCIONES DE NEGOCIOS CÍA. LTDA. *Tesis.* UNIVERSIDAD CENTRAL DEL ECUADOR, Quito. Obtenido de http://www.dspace.uce.edu.ec/handle/25000/16053

Sarabia, D. (2020). *Arquitectura de análisis de datos generados por el Internet de las cosas (IoT) en tiempo real.* Universidad Politécnica de Valencia, Valencia. Obtenido de https://riunet.upv.es/bitstream/handle/10251/149398/Sarabia%20-%20Arquitectura%20de%20an%C3%A1lisis%20de%20datos%20generados%20por

%20el%20internet%20de%20las%20cosas%20IoT%20en%20tiempo%20....pdf?sequence=4

Vaca, R. (2018). *EL USO DE BIG DATA Y SU INCIDENCIA EN LA CALIDAD DE LOS SERVICIOS ACADÉMICOS DE LA UNIVERSIDAD TÉCNICA DE AMBATO.* UNIVERSIDAD TÉCNICA DE AMBATO, Ambato. Obtenido de https://repositorio.uta.edu.ec/handle/123456789/25796

Zeebaree, S., Shukur, H., Haji, L., Zebari, R., Jacksi, K., & Abass, S. (2020). Characteristics and Analysis of Hadoop Distributed Systems. *Technology Reports of Kansai University*, 1555-1564.

## Anexos

La herramienta Map Reduce, pese a ser de mucha ayuda en la Big Data, tiene también sus ventajas y desventajas, lo cual es una información muy importante para las industrias, ya que pueden tomar la decisión de adoptar o no esta herramienta dentro de la organización.

A continuación, se presenta los aspectos positivos y negativos de la herramienta Map Reduce para una mejor comprensión:

**Tabla 2.**
Ventajas de la herramienta Map Reduce

| CARACTERISTICAS | DEBATE |
|---|---|
| Escalabilidad | "Map Reduce es altamente escalable, lo que significa que puede manejar grandes volúmenes de datos al distribuir tareas en clústeres de computadoras. "[a] |
| Tolerancia a fallos | "Map Reduce es resistente a fallos. Si un nodo falla, los datos y las tareas se redistribuyen automáticamente a otros nodos disponibles. "[b] |
| Utilización eficiente de recursos | "Map Reduce permite aprovechar al máximo los recursos de hardware al ejecutar tareas en paralelo en múltiples nodos del clúster. Esto puede mejorar el tiempo de procesamiento y el rendimiento general." [a] |

Nota. aAlmora, (2018) bGómez, (2021). Esta tabla muestra los puntos fuertes de Map Reduce al momento de procesar grandes cantidades de datos.

**Tabla 3.**
Desventajas de la herramienta Map Reduce

| CARACTERISTICAS | DEBATE |
|---|---|
| Rendimiento en tareas no estructuradas | "Si bien Map Reduce es altamente eficiente para tareas de procesamiento por lotes y procesamiento de datos estructurados, puede no ser la mejor opción para aplicaciones más interactivas." [a] |
| Overhead en procesos pequeños | "El modelo Map Reduce puede tener un alto costo de inicio en términos de tiempo y recursos, lo que lo hace menos eficiente para procesar conjuntos de datos pequeños." [b] |
| Complejidad para ciertos algoritmos | "Algunos algoritmos no se ajustan naturalmente al modelo de Map Reduce y pueden requerir transformaciones y adaptaciones complejas para funcionar de manera eficiente en este entorno." [b] |
| Limitaciones de latencia | "Debido a la naturaleza por lotes del proceso Map Reduce, puede haber cierta latencia en la obtención de resultados, especialmente en comparación con sistemas de procesamiento en tiempo real." [a] |

Nota. [a]Cisneros & Liliana, (2019) [b]Vaca, (2018). Esta tabla muestra los puntos débiles de Map Reduce al momento de procesar grandes cantidades de datos.

**Tabla 4.**
Características de la herramienta Map Reduce

| **Característica** | **Hadoop Map Reduce** | **Observación** |
|---|---|---|
| Procesamiento en Lote | Excelente | Map reduce es una herramienta que permite un buen procesamiento por lotes, tiene una amplia comunidad en soporte y puede ser fácilmente integrada con otros ecosistemas, sin embargo, el procesamiento en tiempo real es limitado y no sería de gran utilidad en empresas relacionadas al streaming. |
| Procesamiento en Tiempo Real | Limitado | |
| Escalabilidad | Excelente | |
| Facilidad de Uso | Moderada | |
| Velocidad de Procesamiento | Moderada | |
| Tolerancia a Fallos | Sí | |
| Comunidad y Soporte | Amplia | |
| Integración con Otros Ecosistemas | Sí | |
| Herramientas de Administración | Amplias | |

Nota. Esta tabla muestra las características recogidas en el marco conceptual para conocer a simples rasgos detalles importantes de Map Reduce.