



**TRABAJO DE GRADO**  
**Opción Seminario-Diplomado.**

**Aportes de Big Data para la computación en la nube**

Corporación Universitaria Remington.  
Facultad de ingenierías  
Ingeniería de sistemas

Estudiantes:  
Yeifran Aponte Rubio  
Yenny Patricia Rodriguez Alarcon  
Cristian David Sanchez Cabrera  
Tutor:  
Roberto Carlos Guevara Calume  
Opción de Trabajo de grado Seminario-Diplomado:  
Big Data  
2024

### **Dedicatoria**

A nuestros familiares que nos brindaron el espacio y las herramientas para llevar a cabo nuestros proyectos de vida, la cual es obtener un título profesional.

### **Agradecimientos**

Extendemos nuestro agradecimiento a nuestro tutor Roberto Guevara por su pedagógica orientación en el trabajo de grado, la cual fue parte esencial para lograr el objetivo con la elaboración del trabajo.

**Tabla de Contenidos****Contenido**

Resumen.....	5
1. Marco conceptual y contextual .....	5
Marco contextual .....	7
Desarrollo e implementación del aprendizaje.....	8
Título 1 escalabilidad y flexibilidad en procesos de big data .....	8
disminucion de costos por uso .....	9
Figuras y tablas .....	11
Conclusiones .....	14
Referencias.....	14

## **Resumen**

Durante el desarrollo de este documento se evidenciará Como el Big Data y la computación en la nube permite impulsar el desarrollo de las investigaciones, gracias a sus múltiples herramientas que permiten investigar y tener información de primera a la mano, también es de considerar que gracias a estos grandes desarrollos la humanidad ha logrado recopilar y encajar información que se ha venido recolectando.

Aborda el impacto de la computación en la nube, en el procesamiento y análisis de grandes volúmenes de datos, comúnmente conocido como Big Data. A medida que las organizaciones generan y almacenan cantidades masivas de información, se vuelve imperativo contar con sistemas que no solo manejen, sino que también extraigan valor de esos datos de manera eficiente.

El Big data y la computación en la nube aporta una gran parte esencial en dichas infraestructuras gracias a sus mecanismos para almacenar y recolectar información, su facilidad de interpretar datos permite tener una velocidad, escalabilidad, y flexibilidad a la hora de procesar y analizar la información.

Lo que se conoce hoy en día como Nube ya sea Cloud, Gmail, Dropbox, icloud. Es un contenedor inmenso de información que se almacena de cuartos que contienen miles de servidores, que a su vez son alimentados por la información que vaga en el mundo.

***Palabras Clave:*** Análisis de macrodatos, Almacenamiento, Servidores, Infraestructura en la Nube

## **Marco Conceptual y Contextual**

El presente informe técnico se sitúa en la confluencia de dos paradigmas contemporáneos de gran relevancia: la computación en la nube y el fenómeno del Big Data. La computación en la nube se define como un modelo de prestación de servicios informáticos que permite el acceso on demand a un conjunto compartido de recursos computacionales -que incluye servidores, almacenamiento, aplicaciones y servicios- que pueden ser provisionados y desmantelados con celeridad mediante la red.

### **Cloud Computing**

Se refiere la aplicación de técnicas en los datos, una poderosa combinación de tecnología y análisis de datos es un campo que implica extraer conocimientos y obtención de información valiosa a partir de grandes conjuntos de datos.

Los datos suelen ser estructurados y se generan de diversas fuentes, como lo son redes sociales, transacciones en línea, registros en servidores, entre otros.

### **Nube Publica**

Modelo del cual su infraestructura TI, ofrecen recursos virtuales accesibles a través de internet que son compartidos dentro de una misma red en una organización, una de sus características de la nube publica es que cuenta con amplias redes de centro de datos físicos repartidos por todo el mundo.

### **Nube Privada**

Es donde una única organización controla y gestiona la infraestructura de los recursos decidiendo así que departamento puede gestionar y visualizar los datos ejemplo departamento de marketing y departamento de finanzas.

### **Nube Híbrida**

Es donde una organización utiliza los recursos tanto de una nube privada y una publica, la organización utiliza el diseño de su infraestructura TI internos con los servicios de proveedores de nubes de terceros. Con la nube híbrida se pueden almacenar datos y ejecutar aplicaciones en varios entornos, como lo constata (Cloud, Oracle, 2024)

## **Marco Contextual**

Antes de las primeras apariciones de la computación en la nube uno de los grandes problemas que afrontaban la compañías o mismos usuarios eran las limitadas herramientas para almacenar información, también la manera que podían gestionarla y la manera en que podían acceder a ella. Pues todos los datos se encontraban por así decirlo desordenados.

El concepto fundamental de los recursos informáticos tiene raíces desde los años sesenta, cuando JCR Licklider introdujo la idea de una “red de computadores intergalácticas”, cuya visión era que todo el mundo pudiese estar interconectado y poder acceder a los programas y datos desde cualquier lugar.

Fue cuando Amazon Web Services en 2002 introdujo servicios basados en la nube, incluyendo almacenamiento, computación e incluso inteligencia humana.

Permitiendo así un paso al desarrollo del big data junto con la computación en la nube.

La web 2.0 es la evolución más reciente. Sin embargo, como internet no comenzó a ofrecer ancho de banda significativo hasta los años noventa, la computación en la nube ha sufrido algo así como un desarrollo tardío como aparece en. (Aws.Amazon, 2023)

### **Desarrollo e Implementación del Aprendizaje**

En el siguiente informe se observa y se analiza la unificación y aplicación de BID DATA en entornos de computación en la nube exponiendo como estas tecnologías admiten el manejo eficiente de gran cantidad de datos.

La cantidad de datos que generamos y almacenamos es de forma veloz, es por ello que estos datos pueden provenir de muchas fuentes, tales como las redes sociales, registros médicos, transacciones financieras, etc.

A continuación, vamos a observar la gestión y análisis de la capacidad de datos, con el objetivo de tener información valiosa para la toma de decisiones, como (Singh, 2020)

### **Escalabilidad y Flexibilidad en Procesos de Big Data**

Permite que las empresas puedan almacenar y procesar grandes cantidades de datos, gracias a la nube es posible incrementar la capacidad de procesamiento, se realizaron prácticas en plataformas como Amazon web services y Google cloud, demostrando a su vez la capacidad de estas grandes plataformas para almacenar y procesar una gran cantidad de datos. (Nephosit, 2024)

### **Caso de Éxito**

RetailCorp, dedicada a la comercialización de productos electrónicos a través de múltiples canales, que incluyen tanto establecimientos físicos como ventas en línea. Ante un crecimiento exponencial en la generación de datos provenientes de sus transacciones y de la interacción con su clientela, RetailCorp se enfrenta al formidable desafío de extraer valor significativo de esta información, con el fin de enriquecer su estrategia comercial y optimizar la experiencia del consumidor, como lo indica (Es.Wikipedia, 2024)

### **Disminución de Costos por Uso**

Reduce la necesidad de la inversión en hardware como los servidores y almacenamiento físico, ya que las cantidades de almacenamiento en la nube permite que las grandes empresas paguen por los recursos que utilizan, el uso de la nube redujo un 40% en los mantenimientos de servidores.

Mediante la implementación de un sistema fundamentado en la computación en la nube, RetailCorp persigue la finalidad de emplear herramientas analíticas que le permitan segmentar a sus clientes, anticipar la demanda de productos y perfeccionar su gestión de inventario. En este contexto, el trabajo se concentrará en la configuración de un entorno de procesamiento de datos utilizando la plataforma Amazon Web Services (AWS), donde se llevará a cabo un análisis exhaustivo de datos históricos relativos a ventas y comportamientos de consumo. El objetivo de este ejercicio es demostrar de manera fehaciente cómo la sinergia entre la computación en la nube y Big Data puede conferir a RetailCorp una ventaja competitiva, facilitando la toma de decisiones informadas y sustentadas en datos, tal como lo indica (Aws.Amazon, 2023)

### **Descripción del Procedimiento**

El proceso de ejecución se estructuró en varias etapas meticulosamente planificadas, cuya descripción es fundamental para comprender los resultados alcanzados

1. Configuración del entorno de nube: Se desplegó un clúster mediante el servicio Amazon EMR (Elastic MapReduce), con una configuración que incluía nodos maestros y nodos esclavos, ajustados para maximizar la eficiencia en el procesamiento de grandes volúmenes de datos. Las instancias EC2 de tipo m5.xlarge fueron seleccionadas para equilibrar el costo con la capacidad de procesamiento.
2. Ingesta y almacenamiento de los datos: El conjunto de datos de ventas fue almacenado en Amazon S3, que facilitó tanto la ingesta como el posterior acceso eficiente a los datos. Este servicio de almacenamiento distribuido permitió la manipulación fluida de los datos a lo largo de todo el proceso.
3. Procesamiento de los datos: El procesamiento de los datos se efectuó mediante Apache Spark, cuya arquitectura distribuida posibilita un manejo óptimo de tareas de gran envergadura. Se realizaron operaciones de limpieza de datos, así como análisis agregados que incluyeron la segmentación de ventas por regiones geográficas. Para llevar a cabo predicciones de demanda, se implementaron algoritmos de aprendizaje automático entrenados sobre los datos, como lo constata (Aws.Amazon, 2023)
4. Visualización de los resultados: Los datos procesados fueron almacenados en Amazon Redshift y, a través de la herramienta de visualización Amazon QuickSight, se generaron representaciones gráficas que ofrecieron un panorama claro del rendimiento de

ventas, así como de las previsiones de demanda futura, lo cual constituyó el pilar para la toma de decisiones fundamentadas.

### **Análisis de los Resultados Obtenidos**

El análisis de los resultados manifiesta una clara superioridad de la infraestructura en la nube frente a la ejecución local, tanto en términos de velocidad como de escalabilidad.

Las observaciones más significativas se detallan a continuación:

1. Reducción en los tiempos de procesamiento: En el entorno local, el procesamiento del conjunto de datos de 100 GB tomó alrededor de 120 minutos. Sin embargo, empleando el clúster de Amazon EMR, el mismo volumen de datos fue procesado en apenas 40 minutos, lo que representa una reducción del 66% en el tiempo de ejecución.
2. Escalabilidad: Al incrementar progresivamente el tamaño del dataset, el entorno local mostró una relación exponencial entre el tamaño de los datos y el tiempo requerido para procesarlos. En cambio, el sistema en la nube exhibió una elasticidad notable, manteniendo tiempos de procesamiento relativamente constantes, incluso para volúmenes de datos de hasta 200 GB, como lo indica (Zhang, 2021)

### **Figuras y Tablas**

#### **Tabla 1**

*Tiempos de procesamiento en función del tamaño del dataset.*

Tamaño del Dataset (GB)	Tiempo Local (min)	Tiempo en la Nube (min)
50 GB	80	25
100 GB	120	40
200 GB	240	55

*Nota.* La Tabla 1 y el Gráfico 1 exponen de manera clara la ventaja competitiva de la nube en términos de procesamiento de datos a gran escala. La capacidad de añadir nodos al clúster de Amazon EMR sin comprometer la integridad del procesamiento permitió reducir significativamente el tiempo de ejecución, lo cual no fue posible replicar en un entorno local limitado por su infraestructura

## Tabla 2

*Ventajas Y Desventajas Del Big Data En Computación En La Nube.*

VENTAJAS	DESVANTAJAS
<p><b>Flexibilidad y Escalabilidad:</b> Puede aumentar o reducir la capacidad de grandes inversiones.</p>	<p><b>Dependencia de la conectividad:</b> Requiere una conexión a Internet</p>
<p><b>Pago por uso:</b> No hay necesidad de comprar hardware caro ni mantener centros de datos locales.</p>	<p><b>Preocupaciones sobre la privacidad:</b> Aunque los proveedores en la nube tienen altos estándares de seguridad, algunas industrias prefieren mantener datos sensibles.</p>
<p><b>Acceso a tecnologías avanzadas:</b> Integración con servicios de IA, ML, análisis en tiempo real, etc.</p>	

## FIGURA 1

*Big data y servicio en la nube*



*Nota.* La imagen muestra un flujo de datos interconectado a través de una red de nodos, con un enfoque en los servicios en la nube como almacenamiento, análisis e inteligencia artificial, reflejando la naturaleza escalable y flexible de esta tecnología.

## **Conclusiones**

Los resultados obtenidos corroboran lo expuesto en la literatura sobre las bondades de la computación en la nube, particularmente cuando se trata de procesar grandes volúmenes de datos. La reducción del tiempo de procesamiento en más del 60% al migrar a la nube pone de relieve la eficiencia inherente de los sistemas distribuidos, en contraposición a la rigidez de las infraestructuras locales, que, por su naturaleza, sufren de cuellos de botella conforme se incrementa el tamaño de los datos.

Dichos hallazgos son consonantes con investigaciones previas, tales como las llevadas a cabo por Zhang et al. (2021), quienes también reportaron mejoras sustanciales en la eficiencia de procesamiento de datos al hacer uso de plataformas de nube escalables. En consonancia con este estudio, la presente ejecución en AWS no solo reafirma la superioridad de la nube en términos de tiempo, sino que también resalta su flexibilidad para adaptarse a necesidades variables de procesamiento, característica especialmente valiosa en escenarios de Big Data.

## **Referencias Bibliográficas**

- Aws.Amazon. (2023). *aws.amazon.com*. Obtenido de <https://aws.amazon.com/es/what-is/public-cloud/>
- Bastidas, L.R. (2007). *El inicio del siglo XXI*. Planeta. Sitio web: <http://www.rbastidasl.com/libro-inicio-del-sigloxxi>.
- Borges, J.L. (2013). *Ficciones*. Buenos Aires, Argentina: Debolsillo.
- Campusbigdata. (2024). *www.campusbigdata.com*. Obtenido de <https://www.campusbigdata.com/blog/cloud-big-data-science-poder-de-los-datos-en-la-nube/>
- Cloud, O. (2024). *www.oracle.com*. Obtenido de <https://www.oracle.com/co/cloud/what-is-cloud-computing/>
- Es.Wikipedia. (3 de Octubre de 2024). *es.wikipedia.org*. Obtenido de [https://es.wikipedia.org/wiki/Computaci%C3%B3n\\_en\\_la\\_nube#Comienzos](https://es.wikipedia.org/wiki/Computaci%C3%B3n_en_la_nube#Comienzos)
- Neposit. (2024). *www.neposit.com*. Obtenido de <https://www.neposit.com/big-data-con-cloud-computing/>
- Singh, A., & Verma, R. (2020). Optimizing Big Data Processing in Cloud-Based **Systems**. *Journal of Cloud Computing*, 8(3), 123-140.
- Zhang, Y., Chen, H., & Liu, J. (2021). Scalability and Performance of Big Data Analytics in the Cloud. *IEEE Transactions on Cloud Computing*, 9(1), 67-78.