Investigación Grupal 1*
Bases de datos NoSQL: Familias de columnas

* **Repositorio en GitHub (acceso directo)**

1st Iván Daniel Rodríguez-Cruz
Escuela de Estadística y Sociología
Universidad de Costa Rica
San José, Costa Rica
ivan.rodriguezcruz@ucr.ac.cr

2nd Cesar Peñaranda-Chaves
Escuela de Estadística
Universidad de Costa Rica
San José, Costa Rica
cesar.penaranda@ucr.ac.cr

3rd Luis Carlos González
Escuela de Estadística y Psicología
Universidad de Costa Rica
San José, Costa Rica
luiscarlos.gonzalez@ucr.ac.cr

4th Joseph Elí Rivera Noguera
Escuela de Estadística y Administración de Negocios
Universidad de Costa Rica
San José, Costa Rica
joseph.rivera@ucr.ac.cr

5th Andrea Sánchez Corella
Escuela de Estadística
Universidad de Costa Rica
San José, Costa Rica
andrea.corella@ucr.ac.cr

***Abstract*—Este trabajo se enfoca en la utilización de bases de datos columnares para solucionar problemáticas empresariales y comparar dos de estas bases de datos. Se concluye que tanto Apache Cassandra como Amazon Redshift son útiles para el manejo de grandes volúmenes de datos, pero presentan diferentes fortalezas y dificultades de uso. La documentación de Apache Cassandra señala que el software está diseñado para tener un mejor rendimiento en tareas de gestión de aplicaciones en tiempo real o sistemas distribuidos a gran escala, mientras que Amazon Web Services sugiere que Amazon Redshift trabaja mejor en análisis empresariales. Se debe considerar que Amazon Redshift es de paga, lo cual puede salir del alcance de muchas personas o empresas, no obstante, se obtiene un servicio que está disponible en la nube y con sistema de guardado automático y una interfaz más amigable con el usuario lo que lo hace más sencillo de aprender y utilizar. En general, las bases de datos columnares son una solución que cumple con las expectativas de rendimiento en el manejo de grandes cantidades de datos y pueden ser una alternativa a las bases de datos relacionales tradicionales.**

***Index Terms*—NoSQL, columnar, Amazon Redshift, Apache Cassandra, bases de datos.**

I. Introducción

En la actualidad es común ver a empresas con grandes cantidades de datos almacenados y con diferentes necesidades para gestionar y darles el mayor provecho a estos datos, ante esta situación el mercado actual provee de diferentes herramientas para satisfacer las necesidades que puede presentar una empresa en cuanto a la gestión de datos. Una de estas herramientas son las bases de datos columnares. Por decirlo de una manera una base de datos columnar es similar a la transpuesta de una matriz de datos, lo que permite tener un acceso más rápido a las columnas de datos [1].

En este trabajo se planteará una posible problemática que pueda presentar una empresa y se utilizaran 2 bases de datos columnares para tratar de solucionar esta problemática, con la intención de comparar las 2 bases de datos columnares.

II. Problema de estudio

Un comercio de electrónicos, por ejemplo ExtremeTech, dada sus características de tienda física y virtual debe y necesita realizar un monitoreo de sus ventas a lo largo del tiempo para poder realizar análisis de mercado que lleguen a decisiones informadas y así mantener un buen seguimiento de sus finanzas. Esto supone que la tienda es muy probable que al momento de observar las ventas se pregunte cuáles fueron sus productos más cotizados (conocer la demanda y estar preparados en el sentido de *stock*) por los usuarios entre otras interrogantes.

Para todo esto es importante que los responsables del stock y las finanzas de la empresa tengan la habilidad de almacenar información sobre cada una de las ventas que se realizan día a día en sus locales (se puede hablar de qué producto compró el usuario, el precio, la marca de fabricación, las especificaciones más básicas del producto, etc.), con el fin de poder ir sabiendo cómo se comportan sus usuarios en las ventas.

A. Objetivo del estudio

En concreto, se pretende implementar una *base de datos columnar en NoSQL* para el seguimiento de ventas de un comercio electrónico. Esta base de datos permitiría realizar consultas muy concretas y eficientes sobre los datos, lo que llevaría a poder realizar análisis de ventas de una forma más robusta y rápida, en vez de realizarlo en programas como Excel (por el volumen de productos que vende la tienda, tener una hoja de Excel no es viable) [2].

B. Tipo de base de datos

Para la realización de este proyecto se pretenden usar las bases de datos *Columnares* de NoSQL debido a que ellas están diseñadas para almacenar y analizar grandes cantidades de

datos de forma eficiente. En este caso, la base de datos almacenaría los datos de ventas en columnas, una columna para cada atributo (variable) de datos. Conociendo esto, los encargados de mercadeo podrían hacer un seguimiento de su empresa con mayor eficiencia y escabilidad de los datos en el futuro. Por ejemplo, se podría obtener una tabla de los productos de la Marca X más vendidos durante un determinado lapso [1].

C. Software empleado

Para la implementación de este trabajo se hizo uso de los softwares Apache Cassandra (código abierto y gratuito) y Amazon Redshift (de paga y ligado a Amazon Web Services).

(a) Apache Cassandra (b) Amazon Redshift

Fig. 1: Softwares de bases columnares NoSQL.

III. Análisis

En el presente trabajo, se utiliza una base de datos columnar, estas organizan los datos en columnas en lugar de filas, ya que almacena los valores de una columna específica juntos. Esto hace que el acceso a los datos y la posibilidad de comprensión sea más eficaz, especialmente cuando se realizan consultas que sólo se requieren ciertas columnas. Desde 2004 ha tenido una evolución constante para implementaciones comerciales, porque las consultas en estas bases se vuelven sumamente rápidas, incluso en bases de datos muy extensas, ya que omiten los datos irrelevantes para el análisis y así se logra leer de inmediato lo que se necesita. También estas bases de datos columnares almacenan los datos en registros por lo que pueden contener un gran número de columnas dinámicas, esto a diferencia de las bases de datos relaciones que están optimizadas para almacenas filas de datos[3].

A. Apache Cassandra

Es un sistema de gestión de base de datos de código abierto para bases de datos muy extensas, pero estructuradas, pertenece a las bases de datos NoSQL columnares. Cuenta con un lenguaje de consulta propio con el nombre de Cassandra Query Language (CQL), es muy parecido al SQL, sin embargo, es preferido por los desarrolladores, ya que cuenta con un enfoque redundante que hace que se reduzca la probabilidad de fallo. Lo utilizan grandes empresas como Apple, Faceboock, Instagram, Uber, etc. Por su gran capacidad de manejar grandes volúmenes de datos que hace que sea beneficioso[4].

Además, Cassandra es una de las bases de datos más utilizadas, porque ofrece un servicio de alta disponibilidad y sin un único punto de fallo, a su vez, puede manejar con eficacia enormes cantidades de datos en múltiples servidores, también, este sistema es rápido y preciso para grandes cantidades de datos como para volúmenes más pequeños. Asimismo, una de las razones por la cual las empresas utilizan Cassandra es por su escalabilidad horizontal, ya que su estructura permite a sus usuarios hacer frente al aumento repentino de la demanda, porque permite añadir más hardware para incorporar clientes y datos adicionales, esto ayuda a las empresas a escalar sin hacer ajustes importantes.

Por otro lado, la estructura de este sistema también hace que los datos estén debidamente protegidos, esto a través por un registro de *commit*, este es un método de copia de seguridad para lograr garantizar que los datos no se pierdan, estos se indexan y se escriben en una tabla de memoria (memtable), que es una estructura de datos en la memoria y existe una activa por cada tabla[4].

B. Amazon Redshift

El servicio en la nube de Amazon conocido como Amazon Redshift representa una poderosa herramienta diseñada para el almacenamiento, procesamiento y gestión de datos, basada en la estructura columnar de bases de datos NoSQL [5]. La amplia aceptación que ha experimentado este servicio puede atribuirse en gran medida a la creciente cantidad de datos que las empresas manejan en la actualidad, donde estos datos se consideran uno de los activos más valiosos de cualquier organización y, por ende, requieren medidas efectivas para garantizar su confidencialidad y su procesamiento eficiente.

En este contexto, Amazon Redshift se destaca por ofrecer a los usuarios un entorno conveniente, aprovechando la infraestructura sólidamente establecida de AWS. Esto implica que la creación de un entorno personal o empresarial en los servidores de Amazon se realiza de manera sencilla. En comparación con otros sistemas, como Apache Cassandra, Amazon Redshift proporciona a los usuarios una amplia gama de herramientas y entornos que simplifican significativamente la gestión de datos, contribuyendo así a una experiencia más fluida y eficiente en su uso.

Además, al estar trabajando en un entorno ligado a AWS la escalabilidad de las bases de datos se hace mucho más cómodo por el simple hecho de aumentar o solicitar mayor almanenamiento en los *clubster* o **buckets** de Amazon (se incurren en más gastos). Este estilo de bases de datos en la nube, también permiten varias cosas: 1) creación de grupos de trabajos de forma más eficiente y práctica; 2) se cobra por almacenamiento y procesamiento de datos, esto puede ser útil al momento que una empresa crezca, en el sentido de la cantidad de data que guarda, para poder ir escalando y aumentando la capacidad de almacenamiento; 3) acceso más seguro y controlado por las alertas y servicios extras que ofrece todo el conjunto de AWS; 4) al ser un lenguaje en arquitectura columnar también facilita la flexibilidad de ir añadiendo, sin un esquema delimitado y fijo, más columnas que puedan servir

para entender y comprender los datos en cuestión; 5) la sintaxis de las consulta y creación de las bases de datos se asemeja en gran parte al lenguaje universal de SQL, esto facilita la migración o simbiosis entre sistemas de base de datos [5].

C. Comparaciones entre softwares

Al utilizar tanto Cassandra como Amazon Redshift en la creación de una base de datos diseñada para abordar los requerimientos de almacenar datos relacionados con un comercio electrónico, con el fin de permitir un análisis exhaustivo de ventas y gestión de inventario, es posible identificar varias diferencias entre estos dos productos. A continuación, se presentan algunos de sus principales contrastes, destacados en los siguientes apartados:

1) Consistencia: **A**pache Cassandra y Amazon Redshift son dos herramientas con características distintas. Apache Cassandra es un sistema de gestión de base de datos distribuida altamente escalable, diseñada para manejar grandes volúmenes de datos y está optimizada para escrituras rápidas y consultas flexibles. Por otro lado, Amazon Redshift es un servicio de almacenamientos de datos en la nube basado en columnas, y este está diseñado específicamente para el análisis de grandes conjuntos de datos. Es altamente escalable y ofrece un rendimiento rápido para consultas complejas en grandes volúmenes de datos. Ambos sistemas pueden manejar grandes volúmenes de datos, pero tienen diferencias significativas en términos de infraestructura y capacidades.

Tanto Apache Cassandra como Amazon Resdshift cuentan con mecanismos de tolerancia al fallo. Apache Cassandra está diseñado bajo un sistema de replicación de los datos en diferentes nodos. Los nodos son instancias individuales que forman parte del clúster de apache Cassandra. Al tener la información replicada en los diferentes nodos, si algún nodo falla no habría problema, ya que los datos podrían recuperarse desde otro nodo [6], se podría decir que los nodos representan la capacidad de expandirse que tiene Cassandra. Por su parte al utilizar Amazon Redshift se estaría haciendo uso de un almacén en la nube, administrado por Amazon Web Services [5]. Este almacén permite la creación de copias de seguridad en momentos específicos por lo que es posible restaurar la base de datos a una versión previa, por lo que ambos softwares garantizan la disponibilidad de los datos de diferentes formas.

2) Desempeño en inserción de datos: En lo que respecta a la sintaxis, Amazon Redshift utiliza la convencional sintaxis de SQL, en marcado contraste con el lenguaje específico de Apache Cassandra, conocido como *The Cassandra Query Language* (CQL). Esto significa que, al trabajar con bases de datos NoSQL, cada empresa y plataforma tiende a implementar sus propios lenguajes de consulta o adaptaciones de la sintaxis subyacente en SQL.

Amazon Redshift se destaca por su entorno más usuario-amigable, ofreciendo una interfaz gráfica con una amplia variedad de opciones que facilitan la creación e inserción de datos en diversos estilos. Por otro lado, Apache Cassandra presenta un enfoque más específico, con su propio lenguaje que, si bien inicialmente recurre a algunas funciones básicas de SQL, pone un énfasis más fuerte en aspectos como la replicación y una estructura de columnas más sólida que la proporcionada por SQL [7].

Para cassandra la sintaxis para subir los datos de una tabla tipo csv en una base para dicha plataforma se resume en lo siguiente:

COPY ventas(row_id, order_id, order_date, ship_date, ship_mode, customer_id, customer_name, segment, country, city, state, postal_code, region, product_id, category, sub_category, product_name, sales, quantity, discount, profit) FROM 'VentasElectronicos.csv' WITH delimiter =';' and HEADER = TRUE;

Podemos observar cómo se nombran las columnas a insertar también de donde provienen los datos además de otras características típicas de este tipo de código o similares (SQL), por otra parte, es importante destacar que al momento de nombrar de donde provienen los datos es necesario que el archivo nombrado este en armonía con las rutas especificadas con el código para que el programa encuentre el .csv en este caso sin mayor problema.

3) Desempeño en recuperación de datos: En cuanto a los tiempos de ejecución de las consultas, las páginas web oficiales de los 2 softwares dejan claro que cada sistema está diseñado con propósitos diferentes, si el interés de la empresa es hacer análisis de datos empresariales es muy probable que el rendimiento más óptimo se obtenga con Amazon Redshift [5]. En cambio si las necesidades están más enfocadas en la gestión de aplicaciones en tiempo real o sistemas distribuidos a gran escala Cassandra podría tener un mayor rendimiento [6].

4) Facilidad de uso: Cuando evaluamos la facilidad de uso de uno u otro sistema, la experiencia del individuo encargado de la administración de bases de datos juega un papel crucial. Sin embargo, en el contexto de este proyecto, se observó que la instalación inicial de Apache Cassandra resultó ser un desafío más significativo en comparación con Amazon Redshift. Esto se debe a que, en el caso de Cassandra, especialmente en un entorno como Windows 11, requiere la instalación a través de la línea de comandos (CMD) y la preexistencia de otros lenguajes de programación, como Java y Python, como se señala en [8]. Por otro lado, para utilizar Amazon Redshift, simplemente se necesita crear una cuenta en AWS para acceder a su entorno de consultas y consultas SQL.

Para una experiencia más óptima con Apache Cassandra, se aconseja su instalación y utilización en un entorno Linux o la utilización de software especializado, como Docker. Sin embargo, este enfoque implica que, para simplemente acceder a Apache Cassandra, los usuarios puedan necesitar familiarizarse con herramientas que, por lo general, no son tan amigables ni familiares para aquellos que carecen de experiencia en la materia.

5) Herramienta de gestión: Existen múltiples herramientas de gestión que se pueden emplear con Apache Cassandra, una opción es interactuar directamente con el clúster utilizando el lenguaje CQL, no obstante hay múltiples softwares libres que se pueden utilizar como herramientas de gestión, una de

las más empleadas es DataStax OpsCenter [9] [10]. DataStax Opscenter se puede utilizar para gestionar y supervisar los clústers de Cassandra en tiempo real [10].

Amazon Redshift al formar parte de los Amazon Web Services sus herramientas de gestión las provee Amazon, como puede ser el AWS Management Console con el que se puede administrar y supervisar Amazon Redshift desde una consola visual. Desde el Management console se pueden ejecutar y supervisar consultas y hasta monitorear el rendimiento de las consultas.

6) Método utilizado para hacer consulta: **E**n Apache Cassandra las consultas se realizan utilizando el lenguaje CQL (Cassandra Query Language), que es similar a SQL pero con algunas diferencias.

Además, en Amazon Redshift se utiliza el lenguaje SQL estándar para realizar consultas. Esto facilita el uso por parte de los desarrolladores y analistas que ya están familiarizados con lo que es el lenguaje de SQL.

7) Almacenamiento: **A**pache Casssandra utiliza un modelo de almacenamiento distribuido, donde los datos se distribuyen en múltiples nodos y cada nodo almacena una parte del conjunto de datos completo. Cassandra replica automáticamente los datos en varios nodos, para garantizar la protección de los datos. Esto permite una alta disponibilidad y escalabilidad horizontal.

Por otra parte, Amazon Redshift utiliza un modelo de almacenamiento columnar, donde los datos se almacenan por columnas en lugar de por filas, a su vez, usa técnicas para reducir el tamaño de los datos almacenados. Esto ayuda a ahorrar espacio en el almacenamiento y con ello mejora el rendimiento de las consultas, porque se reducen la cantidad de datos que se transfieren desde el disco.

D. Dificultades

A nivel general, trabajar con ambos softwares, es decir, Amazon Redshift y Apache Cassandra se pudieron observar las siguientes trabas:

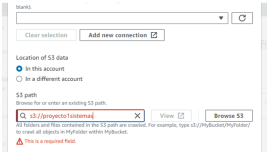
1) Amazon Redshift:

- Amazon Redshift es un sistema de administración de bases de datos que implica el pago de una tarifa por la utilización de sus servidores y sus servicios (Amazon Web Services). Para estudiantes universitarios, esto puede significar que el período de prueba gratuito (crédito gratuito de $300) no sea suficiente, lo que resulta en la acumulación de cargos que pueden superar los mil dólares en su factura.

Fig. 2: Factura generada por la realización de una tabla sin datos.

- Poder cargar una base de datos, por ejemplo en extensión .csv, en Amazon Redshift implica tener conocimientos sólidos en el servicio *Amazon S3* que corresponde al servicio de almacenamiento en la nube de objetos (imágenes, texto, etc.) que posee los mismos beneficios de escalabilidad y disponibilidad que todos los demás servicios de AWS [11]. En este caso, significa que para el trabajo en cuestión no se logró cargar la base de datos, que cuenta con alrededor de 10 mil filas, debido a errores persistentes en los permisos y roles que deben ser especificados en las configuraciones más internas de Amazon S3. A continuación, se enumeran los dos errores que surgieron al ejecutar el comando para cargar bases de datos en Amazon Redshift a través de Amazon S3:

(a) Error de sintaxis.



(b) Error en S3.

Fig. 3: Errores de carga de base de datos con Amazon S3.

- Complejidad de uso al ser la primera vez. Dado que Amazon Redshift se encuentra asociado a AWS la curva de aprendizaje para conocer todos sus requerimientos y herramientas es bastante engorroso de poder realizar de forma ordenada y con un sentido lógico para personas que tienen poca o casi nada de experiencia en este estilo de servicios de base de datos en la nube.

2) Apache Cassandra:

- Apache Cassandra cuenta con limitaciones en información de uso, ya que no es fácil encontrar tutoriales intuitivos ni simples de entender para un usuario aprendiz, ni siquiera se encontró tutoriales complejos a primera mano, fue necesario una ardua búsqueda para hallar información que fuese útil para lograr utilizar el programa a diferencia de programas similares que están mejor documentados y de fácil acceso a material didáctico, por lo que en el caso de Apache Cassandra se necesita una curva de aprendizaje aún más amplia y difícil de cubrir, ya que el material didáctico que se puede encontrar en la red está más dirigido a usuarios familiarizados con estos entornos y no para en el caso de estudiantes que tienen interés en implementar la herramienta para algún fin sin mucho conocimiento previo. Asimismo, la instalación de esta se puede considerar compleja, ya que como mencionamos el material es escaso, además, en el proceso de instalación se encontraron una amplia gama de problemas de compatibilidad con los diversos softwares necesarios para utilizar Apache Cassandra.Desde incompatibilidad con las versiones de Java, hasta corrupción entre programas ya instalados en Windows que no permitían el funcionamiento de este. Por otra parte,la interfaz utilizada

en este caso fue algo rústica ya que fue una terminal de python y tuvo que utilizarse de esta manera debido al no logro de un funcionamiento óptimo en alguna de las otras opciones, en este caso se intentó utilizar DevCenter sin mucho éxito donde se encontraron fallos en el desarrollo de la base de datos.

Por otro lado, podemos mencionar que el desarrollo y sintaxis del código para la creación de la base de datos y lo necesario para poner a prueba el software, fue algo compleja en algunos apartados, especialmente en algo tan simple como cargar la tabla los datos de un archivo tipo .csv, lo cual por lo general suele ser una tarea fácil, en este caso tomo mucho tiempo comprender el funcionamiento del software para poder afinar el código y lograr insertar los datos del archivo a la base, por lo que se puede decir que aunque el software cumplió con las funciones deseadas se presentaron las dificultades mencionadas, dejando claro que el uso de esta herramienta aunque muy útil, aprender a utilizarla desde cero no es tarea sencilla por toda la información necesaria para lograr utilizar de manera básica la herramienta.

IV. Conclusiones

Se determinó que ambos softwares son útiles para el manejo de grandes volúmenes de datos. Sin embargo, tanto Apache Cassadra como Amazon Redshift presentan diferentes fortalezas, así como, algunas dificultades de uso.

A pesar de que el procedimiento realizado en este trabajo no es suficiente para determinar los principales puntos fuertes de cada software, la documentación de Apache Cassandra señala que el software está diseñado para tener un mejor rendimiento en tareas de gestión de aplicaciones en tiempo real o sistemas distribuidos a gran escala, mientras que Amazon Web Services sugiere que Amazon Redshift trabaja mejor en análisis empresariales. Se debe de considerar que Amazon Redshift es de paga, lo cual se puede salir del alcance de muchas personas o empresas, no obstante, se obtiene un servicio que está disponible en la nube y con sistema de guardado automático y una interfaz más amigable con el usuario lo que lo hace más sencillo de aprender y utilizar.

References

[1] Amazon, "¿qué es una base de datos columnar?" 2022. [Online]. Available: https://aws.amazon.com/es/nosql/columnar/

[2] A. Williams, "Nosql database types explained: Column-oriented databases," 2021. [Online]. Available: https://www.techtarget.com/searchdatamanagement/tip/NoSQL-database-types-explained-Column-oriented-databases

[3] J. Hernández, "Base de datos columnares," 2020. [Online]. Available: https://gravitar.biz/bi/base-datos-columnar/#:~:text=Las%20bases%20de%20datos%20columnares%20almacenan%20datos%20en%20registros%20de,en%20cuanto%20a%20los%20requisitos.

[4] R. Cañadas, "Apache casssandra: Base de datos no relacional," 2021. [Online]. Available: https://abdatum.com/tecnologia/cassandra

[5] Amazon, "What is amazon redshift?" 2022. [Online]. Available: https://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html

[6] A. Cassandra, "Cassandra basics," 2023. [Online]. Available: https://cassandra.apache.org/_/cassandra-basics.html

[7] Cylla, "Cassandra query language (cql)," 2022. [Online]. Available: https://www.scylladb.com/glossary/cassandra-query-language-cql/

[8] A. Cassandra, "Installing cassandra," 2022. [Online]. Available: https://cassandra.apache.org/doc/latest/cassandra/getting_started/installing.html

[9] A. Infrabot, "Cassandra administration tools," 2019. [Online]. Available: https://cwiki.apache.org/confluence/display/CASSANDRA2/Administration+Tools

[10] DataStax, "Datastax enterprise opscenter," 2023. [Online]. Available: https://www.datastax.com/products/datastax-enterprise/dse-opscenter

[11] A. W. Services, "Amazon s3," 2022. [Online]. Available: https://aws.amazon.com/es/s3/