Estudiante: *Jose Pablo Arias Navarro - 2021024635*

Resumen 2

Introducing Amazon Redshift

Amazon Redshift es un cloud data warehouse el cual cambió la forma en que las empresas pensaban sobre el almacenamiento de datos, ya que logró reducir el costo y el esfuerzo relacionado con la implementación de sistemas de almacenamiento de datos sin comprometer las características, la escalabilidad y el rendimiento de estas. Amazon Redshift es una solución de data warehousing a escala de petabytes la cual es rápida y administrada completamente, esta simplifica y hace rentable el análisis de grandes volúmenes de datos con las herramientas de inteligencia empresarial (BI).

Modern analytics and data warehousing architecture

Los datos suelen entrar a un data warehouse desde sistemas transaccionales y otras bases de datos relacionales, normalmente incluyen datos no estructurados, semiestructurados y estructurados. Los usuarios pueden acceder a estos datos a través de herramientas de BI, clientes SQL y otras herramientas.

Data warehouses y OLTP databases:

- **Data warehouses:** Estan optimizados para operaciones de escritura por lotes y lectura de grandes volúmenes de datos. (Utilizan esquemas desnormalizados como el Star y el Snowflake)
- **OLTP databases:** Están optimizadas para operaciones de escritura continua y una gran cantidad de operaciones de lectura pequeñas. (Utilizan esquemas altamente normalizados, más adecuados para requisitos de alto rendimiento de transacciones)

AWS analytics services

Estos servicios ayudan a las empresas a convertir rápidamente sus datos en respuestas al proporcionar servicios de análisis integrados. El objetivo de obtener respuestas rápidamente es que da como resultado un menor gasto de tiempo haciendo las conexiones y configurando los servicios de análisis en la nube. AWS nos proporciona una forma fácil para crear data lakes y data warehouses, un almacenamiento seguro en la nube, un stack de análisis totalmente integrado, un buen rendimiento, escalabilidad y un bajo costo.

Analytics architecture

Los pipelines de análisis son diseñados para manejar grandes volúmenes de datos entrantes de fuentes como bases de datos, aplicaciones y dispositivos. Tiene 4 etapas: recopilar datos, almacenar los datos, procesar los datos y analizar y visualizar los datos.

1. ***Recopilación de datos:*** Puede almacenar diferentes tipos de datos, algunos de estos son:

- **Datos Transaccionales:** Normalmente son almacenados en sistemas bases de datos relacionales o en sistemas de bases de datos NoSQL.

- **Datos Log:** Ayudan a solucionar problemas, realizar auditorías y realizar análisis utilizando la información almacenada en los logs.
- **Datos de Streaming:** Estos por lo general deben recopilarse, almacenarse y procesarse continuamente. (Se puede usar Amazon Kinesis o Amazon MSK)
- **Datos IoT:** Datos enviados por dispositivos y sensores, con AWS IoT los dispositivos conectados interactúan de forma fácil y segura con la nube de AWS.

2. ***Procesamiento de datos:*** Los datos recolectados pueden ser analizados para hacer crecer un negocio, hay 2 tipos de flujos de trabajo:

- **Batch processing:**
    - *Extract Transform Load (ETL):* Extrae datos de múltiples fuentes para cargarlos en data warehosing systems. (Es continua y bien definida)
    - *Extract Load Transform (ELT):* Extrae datos y se cargan primero en el sistema destino.
    - *Online Analytical Processing (OLAP):* Almacenan datos en esquemas multidimensionales. (Utilizados para consultas, informes y análisis)
- **Real-time processing:** Puede procesar los datos de forma secuencial e incremental registro por registro o durante ventanas de tiempo deslizantes (sliding time windows). Requiere una capa de procesamiento altamente concurrente y escalable.

3. ***Almacenamiento de datos:***

- **Lake house:** Permite consultar datos de su data warehouse, data lake y de bases de datos operativas para obtener información de manera rápida y profunda.
- **Data warehouse:** Puede ejecutar análisis rápidos en grandes volúmenes de datos y descubrir patrones ocultos en sus datos.
- **Data mart:** Es un data warehouse centrado en un área funcional o tema específico. (Fáciles de diseñar, construir y administrar)

4. ***Análisis y visualización:*** Muchas veces se puede realizar el análisis de datos utilizando las mismas herramientas que se utilizan para procesar datos, como se puede hacer con MySQL Workbench. Por otra parte, Amazon QuickSight es un servicio de BI y basado en la nube que permite crear visualizaciones, realizar análisis y obtener información empresarial de los datos.

Data warehouse technology options

- ***Row-oriented databases:*** Suelen almacenar filas enteras en un bloque físico y son adecuadas para el procesamiento transaccional (OLTP). En un row-based data warehouse cada consulta tiene que leer todas las columnas de todas las filas de los bloques que cumplen el predicado de la consulta, incluidas las columnas que no eligió, esto puede llegar a crear un cuello de botella de rendimiento en los data warehouses.
- ***Column-oriented databases:*** Estas organizan cada columna en su propio conjunto de bloques físicos en lugar de empaquetar las filas enteras en un bloque, esto les permite ser más eficientes en la entrada/salida para consultas de solo lectura. Este tipo de bases de datos son una mejor opción que las row-oriented databases para el data warehousing.
- ***Massively Parallel Processing (MPP) architectures:*** Estas permiten usar todos los recursos disponibles en un cluster para procesar datos, lo que aumenta en gran medida el rendimiento de los

warehouses a escala de petabytes. (Permiten mejorar el rendimiento agregando nodos al cluster)

Amazon Redshift deep dive

Redshift ofrece beneficios clave para un almacenamiento de datos rentable y de alto rendimiento, incluida una compresión eficiente, E/S reducida y menores requisitos de almacenamiento. Se obtiene un rendimiento rápido de consultas y E/S para cualquier tamaño de datos gracias al almacenamiento en columnas, la paralelización y la distribución de consultas en diferentes nodos. Este automatiza la mayoría de las tareas administrativas comunes como configuración, monitoreo, respaldos y protección de un data warehouse.

Integration with data lake

Amazon Redshift provee una característica llamada Redshift Spectrum que facilita la consulta de datos y la reescritura de datos en su data lake en formatos de archivo abiertos como Parquet, ORC, JSON y otros más.

Performance

Amazon Redshift ofrece varias características para lograr su alto rendimiento, entre las que se incluyen:

- *Hardware de alto rendimiento*.
- *AQUA (Advanced Query Accelerator)*, es una caché distribuida y acelerada por hardware que permite que Amazon Redshift se ejecute rápidamente.
- *Almacenamiento eficiente y procesamiento de consultas de alto rendimiento*, el almacenamiento en columnas, la compresión de datos y los mapas de zona reducen la cantidad de E/S necesaria para realizar consultas.
- *Vistas materializadas*, estas se pueden usar para almacenar cálculos previos utilizados con frecuencia, con el fin de acelerar algunas consultas.
- *Gestión automática de la carga de trabajo para maximizar el rendimiento*, se utiliza machine learning para predecir y clasificar las consultas entrantes en función de sus tiempos de ejecución y recursos necesitados.
- Utiliza *result caching* para ofrecer tiempos de respuesta muy cortos en consultas repetidas.

Durability and availability

Redshift detecta y reemplaza automaticamente cualquier nodo con errores en el clúster. Hace que el nodo de reemplazo este disponible de inmediato y carga los datos a los que se accede con más frecuencia. Siempre trata de tener al menos 3 copias de los datos: el original, la réplica y el backup. Además, este también nos permite configurar un entorno sólido de recuperación ante desastres (DR) de una manera sencilla.

Elasticity and scalability

Amazon Redshift permite escalar el proceso y el almacenamiento de forma independiente y pagar solo por lo que se esta usando. Este proporciona dos formas de elasticidad informática:

- ***Elastic resize:*** Permite cambiar rápidamente el tamaño del cluster agregando nodos para obtener los recursos necesarios para cargas de trabajo exigentes y para eliminar nodos cuando se complete el trabajo con la finalidad de ahorrar costos.
- ***Concurrency Scaling:*** Permite admitir usuarios simultáneos virtualmente ilimitados y consultas simultáneas, con un rendimiento de consultas rápido y constante.

Operations

Amazon Redshift Advisor

Nos ayuda a mejorar el rendimiento y disminuir los costos operativos de un cluster. Este ofrece recomendaciones específicas sobre los cambios que se deberían de realizar, estas recomendaciones son personalizadas mediante el análisis de la carga de trabajo y las métricas de uso del cluster.

Interfaces

Amazon Redshift tiene controladores personalizados de Java Database Connectivity (JDBC) y Open Database Connectivity (ODBC), esto permite que se pueda utilizar un amplio rango de clientes SQL conocidos. Redshift proporciona un editor de consultas integrado en la consola web, en el cual se pueden realizar consultas SQL en clusters de Amazon Redshift directamente desde la consola de administración de AWS.

Security

Se puede ejecutar Amazon Redshift dentro de una nube privada virtual basada en el servicio Amazon VPC para brindar seguridad de datos, pues con este podemos definir un firewall y sus diversas reglas. También, Redshift admite conexiones habilitadas para SSL entre la aplicación y el data warehouse cluster, lo que permite cifrar los datos en tránsito. Además, los nodos informáticos de Redshift almacenan los datos, pero solo se puede acceder a ellos desde el nodo principal del cluster.

Cost model

Los cargos se basan en el tamaño y la cantidad de nodos del cluster. Además, no hay ningún cargo adicional por el almacenamiento de una copia de seguridad.

Ideal usage patterns

Algunas de las razones por las que las empresas utilizan Amazon Redshift son: analizar datos de ventas, almacenar datos históricos de comercio de acciones, analizar impresiones de anuncios y clicks, y analizar las tendencias sociales.

Anti-Patterns

Amazon Redshift no es ideal para estos patrones de uso:

- ***OLTP:*** Si necesitamos un sistema transaccional rápido, lo mejor sería trabajar con Amazon Aurora, Amazon RDS o una base de datos NoSQL como Amazon DynamoDB.
- ***Unstructured data:*** Si los datos no están estructurados se puede utilizar ETL o EMR para preparar los datos y poderlos cargar en Redshift.
- ***BLOB data:*** Si se quieren almacenar archivos de objetos binarios grandes (BLOB), como videos, imágenes o música lo mejor sería almacenarlos en S3 y referenciarlos a su ubicación en Amazon Redshift.

Información recuperada del documento "***Data Warehousing on AWS: AWS Whitepaper***"