Introducción de Amazon Redshift

Antes las empresas debían de elegir entre las opciones de aceptar un rendimiento de consulta lento o bien, invertir tiempo y esfuerzo en los procesos de actualización, sin embargo, cuando entra a escena el formato de almacenes de datos en la nube de como lo es Amazon Redshift las demás empresas cambiaron la manera de pensar sobre el almacenamiento de datos que se reducía drásticamente en costo y esfuerzo de implementación, esto con la característica que no ponía en riesgos ninguno de los atributos como la escalabilidad y el rendimiento. Siendo esta una solución con almacenamiento de petabyte escalable a los exabytes, no obstante, esto tenia un costo, precios por hora y hasta por anualidad, pero con una buena funcionalidad. Esto provoco que desde su lanzamiento oficial ha tenido un crecimiento rápido, los cuales muchas empresas y clientes han adoptado este sistema de almacenamiento.

Arquitectura moderna de análisis y almacenamiento de datos.

Alguna de las principales diferencias entre data *warehouse* y las bases de datos *OLTP* son:

- Las *warehouse* están para operaciones de escritura por lotes y lectura de grandes volúmenes de datos. Mientras que las *OLTP* están optimizadas para operaciones de lectura continua y de grandes volúmenes de operaciones de lecturas pequeñas.
- Los *warehouse* emplean un esquemas desnormalizados por el requisito de alto rendimiento de datos, sin embargo, los *OLTP* emplean esquemas altamente normalizados, los cuales son más adecuados para requisitos de alto rendimiento de transferencias.

AWS analytics services

Funcionan para convertir rápidamente los datos en respuestas proporcionadas por servicios de análisis maduros e integrados. Un camino fácil para construir un conjunto de datos y almacenarlos, infraestructura segura de almacenamiento, cuenta con una pila de análisis completamente integrada con un conjunto de herramientas de análisis, mejor rendimiento, mayor escalabilidad y menor costo para el análisis.

Analytics architecture

Los pipelines estan diseñados para manejar grandes flujos de datos de fuentes heterogéneas como bases de datos, aplicación o dispositivos, estos realizan las siguientes etapas: 1) Recoger datos 2) Guardar datos. 3) Procesar los datos. 4) Analizar y visualizar los datos.

Data collection: Se contemplan los siguientes tipos de datos: *transactional data, log data, streaming data* y *IoT data.*

Data processing: Procesa los datos y los recompila proporcionando información útil, se puede analizar la información obtenida para adquirir inteligencia (a base de toda la información que se analiza) para lograr un crecimiento en el proceso que se ejecute.

Data storage: Procesa los datos y los recompila proporcionando información útil, se puede analizar la información obtenida para adquirir inteligencia (a base de toda la información que se analiza) para lograr un

crecimiento en el proceso que se ejecute. Para el procesamiento en tiempo real se tiene que, se puede procesar los datos de manera secuencial e incrementalmente por registros a registro. Ayuda a la visibilidad de muchos aspectos de la actividad en la empresa, actividades de los servidores, los clics en la web y la geolocalización de dispositivos, sin embargo esto requiere una capa de procesamiento altamente recurrente y escalable.

Analysis and visualization: Una vez se hayan procesado los datos ponerlos a la disposición del respectivo análisis se deben utilizar herramientas para este propósito, se pueden utilizar las herramientas como MySQL Workbench para el análisis de datos en Amazon Redshift con ASNI SQL por dar el ejemplo que se menciona.

Data warehouse technology options

Manera para construir los data warehouse orientadas a filas o por columnas.

Orientada a filas: Se almacenan en filas enteras en un bloque físico, además que, el alto rendimiento para las operaciones de lectura se logra a través de los índices secundarios. Para la optimización del rendimiento se realizan las siguientes técnicas: vistas materializadas, tablas acumulativas pre-agregadas, índices por cada combinación de predicado, particiones de datos, uniones basadas en índices.

Orientada a columnas: organizan cada columna en su propio conjunto de bloques físicos en lugar de empaquetar las filas completas en un bloque. Es más eficiente en *I/O* (*Input/Output*) para consultas de lectura

Amazon Redshift Deep dive

Integration with data lake: Esto es una flexibilidad para almacenar los datos altamente estructurados y de acceso frecuente en un *data warehouse* de Redshift con la función llamada *Spectrum* el cual facilita la escritura de datos en el *data Lake*.

Performance: Este ofrece un rendimiento rápido y flexible, además ofrece, hardware de alto rendimiento, **AQUA** (*advanced Query Accelerator*) el cual es una caché distribuida y acelerada por hardware.

Durability and availability: La major durabilidad y disponibilidad. Detecta si hay un nodo fallido y lo reempleza automáticamente en el clúster para que esté disponible lo antes posible.

Elasticity and scalability: Puede escalar cómputo y almacenamiento de manera independiente y esto solo pagando por lo que se utilice. Además, que puede ejecutar cargas de trabajo de almacenamiento no uniformes e imprescindibles. Amazon Redshift proporciona dos formas de elasticidad de cómputo: *Elastic resize* y *concurrency scalling*.

Operations

El servicio de Amazon Redshift incluye las siguientes operaciones: *Amazon Redshift advisor, interfaces, security, cost model, ideal usege patterns y anti-patterns.*