Amazon Redshift

Amazon Redshift es una solución de almacenamiento de datos rápida, totalmente administrada y a escala de petabytes que simplifica y
rentabiliza el análisis de grandes volúmenes de datos mediante las herramientas de inteligencia empresarial existentes. Con Amazon Redshift,
puede obtener el desempeño de los motores de almacenamiento de datos en columnas que realizan un procesamiento paralelo masivo por una décima
parte del costo.

Arquitectura moderna de análisis y almacenamiento de datos

Los datos suelen fluir hacia un almacén de datos desde sistemas transaccionales y otras bases de datos relacionales, y suelen incluir datos
estructurados, semiestructurados y no estructurados.

* Data warehouses: están optimizados para operaciones de escritura por lotes y lectura de grandes volúmenes de datos.
* Bases de datos OLTP: están optimizados para operaciones de escritura continua y grandes volúmenes de pequeñas operaciones de lectura.

Servicios de analsis de AWS

Los servicios de análisis de AWS ayudan a las empresas a convertir rápidamente sus datos en respuestas proporcionando servicios de análisis
maduros e integrados, desde almacenes de datos en la nube hasta lagos de datos sin servidor. Lake Formation proporciona acceso fácil y
bajo demanda a recursos específicos que se ajustan a los requisitos de cada carga de trabajo de análisis. Los datos están curados y catalogados,
ya preparados para cualquier tipo de análisis. Los registros relacionados se emparejan y desduplican con aprendizaje automático.

Arquitectura de Analisis

Los pipelines analíticos están diseñados para manejar grandes volúmenes de flujos de datos entrantes procedentes de fuentes heterogéneas como
bases de datos, aplicaciones y dispositivos. Un canal de análisis típico consta de las siguientes etapas: Recoger los datos, almacenarlos, procesarlos,
analizarlos y visualizarlos.

Datos tradicionales

* Una base de datos NoSQL es adecuada cuando los datos no están bien estructurados para encajar en un esquema definido, o cuando el esquema cambia a menudo.
* Una solución RDBMS es adecuada cuando las transacciones se producen a través de múltiples filas de tablas y las consultas requieren uniones complejas.

Opciones tecnológicas para almacenes de datos

Bases de datos por filas

Las bases de datos orientadas a filas suelen almacenar filas enteras en un bloque físico. El alto rendimiento de las operaciones de lectura se consigue
mediante índices secundarios.

Bases de datos orientadas a columnas

Las bases de datos orientadas a filas suelen almacenar filas enteras en un bloque físico. El alto rendimiento de las operaciones de lectura se consigue
mediante índices secundarios.

Arquitecturas de procesamiento paralelo masivo (MPP)

Una arquitectura MPP permite utilizar todos los recursos disponibles en el clúster para procesar los datos, lo que aumenta drásticamente el rendimiento
de los almacenes de datos a escala de petabytes.

Amazon Redshift deep dive

Amazon Redshift ofrece un desempeño rápido de consultas y E/S para prácticamente cualquier tamaño de datos mediante el uso de almacenamiento en columnas
y la paralelización y distribución de consultas en varios nodos.

Integraciones con data lake

Amazon Redshift proporciona una característica llamada Redshift Spectrum que facilita tanto la consulta de datos como la escritura de datos de vuelta a
su data lake en formatos de archivo abiertos.

Rendimiento

Hardware de alto rendimiento, AQUA, almacenamiento eficiente y procesamiento de consultas de alto rendimiento, vistas materializadas, gestión automática
de la carga de trabajo para maximizar el rendimiento y el rendimiento, almacenamiento en caché de los resultados.

Durabilidad y disponibilidad

To provide the best possible data durability and availability, Amazon Redshift automatically detects and replaces any failed node in your data warehouse cluster.

Operaciones

* Rendimiento del clúster - Amazon Redshift realiza Auto análisis para mantener estadísticas de tablas precisas.
* Optimización de costos - Amazon Redshift le permite pausar y reanudar los clústeres que deben estar disponibles solo en un momento determinado, lo que le permite suspender la facturación bajo demanda mientras no se utiliza el clúster.