

Resumen 2 kenneth Chacón Rivas 2020054320

Introducing Amazon Redshift

Este es un servicio de almacenes de datos en la nube a escala de petabyte brindado por AWS. Cuenta con miles de clientes de muchas industrias gracias a que las empresas pueden ahorrar en implementar sistemas de almacenamiento de datos. En vez de ello hacen uso de Amazon Redshift el cual por sus características permite analizar grandes volúmenes de datos de manera sencilla y rentable gracias a herramientas de inteligencia comercial.

Modern Analytics and data warehousing architecture

Por medio de la tecnología y técnicas avanzadas que se cuentan al alcance de la mano se logra poder analizar grandes volúmenes de datos con una gran robustez y eficiencia, a eso podemos denominarlo como analítica moderna.

Si nos referimos a la arquitectura de data warehousing podemos describirla como la manera en que se gestionan y almacenan los datos. Se basa en una nube flexible y escalable como lo proporciona AWS que proporciona la facilidad de poder hacer el análisis de grandes cantidades de datos, donde se implementa Amazon S3 que es de gran ayuda.

Data warehouse technology options

Encontramos tres opciones disponibles.

Bases de datos orientadas a filas: Estas almacenan filas completas en un bloque físico, a través de índices secundarios se logra el alto rendimiento de las operaciones de lectura. Son adecuadas para el procesamiento transaccional (OLTP). Este tipo de base de datos presenta un inconveniente debido a que se encuentran limitados por los recursos disponibles en una sola máquina, pero los Data marts alivianan un poco con el uso de fragmentación funcional. Bases tradicionales como Oracle Database Server, Microsoft SQL Server, MySQL y PostgreSQL son ejemplos de este tipo.

Bases de datos orientadas a columnas: Este tipo organiza cada columna en su propio conjunto de bloques físicos. Lo que le permite mayor eficiencia en la entrada/salida para consultas de lectura, una ventaja es que se evita el caso de las bases de datos orientadas a filas donde se debe leer todas las filas en cada consulta generando un cuello de botella de rendimiento significativo. También otra ventaja es una mejor visualización de la información obtenida de las consultas. Algunos ejemplos de este tipo de bases de datos son Vertica, Greenplum, Teradata Aster, Netezza y Druid.

Arquitecturas de procesamiento paralelas masivas: Esta arquitectura (MPP) permite hacer uso de todos los recursos disponibles en el clúster para procesar los datos favoreciendo el rendimiento de los almacenes de datos a escala de petabyte. Se puede simplemente agregar más nodos al clúster para mejorar el rendimiento.

Amazon Redshift, Druid, Vertica, Greenplum y Teradata Aster están creados bajo la arquitectura de procesamiento paralelo masivo (MPP).

Amazon redshift deep dive

Amazon Redshift ofrece grandes ventajas para almacenar datos de manera eficaz con un gran rendimiento. Hace uso del almacenamiento en columnas, paralelización y distribución de consultas en múltiples nodos.

Integration with data lake: Mediante la herramienta Redshift Spectrum se facilita la escritura y consulta de datos en formato de archivos abiertos como Parquet, ORC, JSON, Avro, CSV.

Performance: se ofrece múltiples funciones para lograr un alto rendimiento a través de Hardware de alto rendimiento y AQUA, esta última es una memoria caché distribuida y acelerada permitiendo hasta 10 veces más rápido la ejecución de una tarea. también proporciona durabilidad, disponibilidad, elasticidad y escalabilidad.

Operations

Amazon Redshift automatiza por completo muchas tareas operativas como optimización de costos y rendimiento del clúster.

Amazon Redshift Advisor: Mediante el análisis de la carga de trabajo y las métricas de uso de su clúster brinda recomendaciones específicas sobre los cambios que deberían realizarse Interfaces.

Interfaces: Se puede usar una amplia gama de familiar SQL clients por medio de JDBC y ODBC. En la consola web se proporciona una interfaz de editor de consultas, y hay una gran cantidad de integraciones validadas por proveedores de BI y ETL.

Seguridad: En cuanto a la seguridad se puede ejecutar Amazon Redshift dentro de una nube privada virtual que está basada en Amazon Virtual Private Cloud. (Amazon VPC).

también brinda modelo de costo, patrones de uso ideales y anti patrones.