Resumen 1

Bases de Datos II

Luis Diego Delgado Muñoz

Prof. Nereo Campos

Fecha de Entrega: 09/08/2022

Introducción

Los datos son el activo más importante de una empresa. Así tomando esto encuenta, una empresa debe:

- Almacenar cada punto de datos relevante acerca de su negocio.
- Dar acceso a los datos a quien lo necesite.
- Tener la habilidad de analizar los datos de maneras diferentes.
- Destilar los datos hasta obtener conocimientos.

Antes, construir y correr un almacen de datos era sumamente complicado y caro. Las empresas tenían que contratar un equipo de administradores de base de datos para mantener sus consultas funcionando rápidamente y protegerse contra la pérdida de datos.

Introducing Amazon Redshift

En el pasado, cuando los volúmenes de datos crecían o una empresa quería hacer análisis e informes disponibles para más usuarios, tenían que elegir entre aceptar consultas de bajo rendimiento o invertir tiempo y esfuerzo en un costoso proceso de actualización.

Los almacenes de datos en la nube como Amazon Redshift cambiaron la forma de pensar de las empresas almacenamiento de datos al reducir drásticamente el costo y el esfuerzo asociados con la implementación sistemas de almacenamiento de datos, sin comprometer las características, la escala y el rendimiento. Amazon Redshift es una solución de almacenamiento de datos a escala de petabytes, rápida y totalmente administrada, que hace que sea simple y rentable analizar grandes volúmenes de datos utilizando herramientas de inteligencia de negocios (BI).

Modern Analytics and Data Warehousing Architecture

Los datos generalmente fluyen hacia un almacén de datos desde sistemas transaccionales y otras bases de datos relacionales, y generalmente incluyen datos estructurados, semiestructurados y no estructurados. Estos datos se procesan, transforman e incorporan a una cadencia regular.

Los almacenes de datos generalmente emplean esquemas desnormalizados como el esquema Star y el esquema Snowflake debido a los requisitos de alto rendimiento de datos,

mientras que las bases de datos OLTP emplean esquemas altamente normalizados, que son más adecuados para los requisitos de alto rendimiento de transacciones.

Arquitectura analítica

Las tuberías de análisis están diseñadas para manejar grandes volúmenes de flujos de datos entrantes de fuentes heterogéneas, como bases de datos, aplicaciones y dispositivos. Una tubería de análisis típica tiene las siguientes etapas:

1. Recopilar datos
2. Almacenar los datos
3. Procesar los datos
4. Analiza y visualiza los datos

Data Warehouse Technology Options

Las opciones disponibles para crear almacened de datos son: **bases de datos orientadas a filas, bases de datos orientadas a columnas y arquitecturas de procesamiento paralelas masivas.**

Las bases de datos orientadas a filas suelen almacenar filas completas en un bloque físico. El alto rendimiento de las operaciones de lectura se logra a través de índices secundarios. Las bases de datos como Oracle Database Server, Microsoft SQL Server, MySQL y PostgreSQL son sistemas de bases de datos orientados a filas.

Las bases de datos orientadas a columnas organizan cada columna en su propio conjunto de bloques físicos en lugar de empaquetar todas las filas en un bloque. Esta funcionalidad les permite ser más eficientes en la entrada/salida (E/S) para consultas de solo lectura, porque solo tienen que leer aquellas columnas a las que se accede mediante una consulta desde el disco (o desde la memoria). Estas son una mejor opción que las bases de datos orientadas a filas para el almacenamiento de datos.

Una arquitectura MPP le permite utilizar todos los recursos disponibles en el clúster para procesar datos, lo que aumenta drásticamente el rendimiento de los almacenes de datos a escala de petabytes. Los almacenes de datos MPP le permiten mejorar el rendimiento simplemente agregando más nodos al clúster

Amazon Redshift Deep Dive

Como tecnología MPP en columnas, Amazon Redshift ofrece beneficios clave para un almacenamiento de datos eficiente y rentable, que incluye compresión eficiente, E/S reducida y menores requisitos de almacenamiento. Se basa en ANSI SQL, por lo que puede ejecutar consultas existentes con poca o ninguna modificación. Mediante el uso de almacenamiento en columnas, automatiza la mayoría de las tareas administrativas comunes asociadas con el aprovisionamiento, la configuración, el monitoreo, la copia de seguridad y la protección de un almacén de datos, lo que facilita su administración y la hace económica.

Operations

Ideal Usage Patterns

Amazon Redshift es ideal para OLAP utilizando sus herramientas de BI existentes. Las empresas utilizan Amazon Redshift para hacer lo siguiente:

- Ejecución de informes y BI empresarial
- Analizar datos de ventas globales para múltiples productos
- Almacenar datos históricos de operaciones bursátiles
- Analizar impresiones y clics de anuncios
- Datos de juego agregados
- Analizar tendencias sociales
- Medir la calidad clínica, la eficiencia operativa y el desempeño financiero en el cuidado de la salud

Amazon Redshift admite datos semiestructurados y extiende su almacén de datos a su lago de datos. Esto le permite:

- Ejecute análisis según sea necesario en datos de eventos de gran volumen, como análisis de registros y redes sociales
- Descargar datos de historial a los que se accede con poca frecuencia fuera del almacén de datos
- Unir el conjunto de datos externo con el almacén de datos directamente sin cargarlos en el almacén de datos

Anti-Patterns

Amazon Redshift no es ideal para los siguientes patrones de uso:

- OLTP: si necesita un sistema transaccional rápido, puede elegir un sistema de base de datos relacional como Amazon Aurora o Amazon RDS, o una base de datos NoSQL como Amazon DynamoDB.
- Datos no estructurados: los datos en Amazon Redshift deben estar estructurados por un esquema definido. Amazon Redshift no admite una estructura de esquema arbitraria para cada fila.
- Datos BLOB: si planea almacenar archivos binarios de objetos grandes (BLOB), como video digital, imágenes o música, es posible que desee almacenar los datos en S3 y hacer referencia a su ubicación en Amazon Redshift.