Resumen: Almacenamiento de Datos en AWS

Fecha de entrega: *09/08/22 3:25 pm*

Estudiante: *David José Espinoza Soto - 2016012024*

Instrucciones: Del documento adjunto implementar un resumen de máximo 3 páginas.

Introducción

Los datos son el activo más valioso de una empresa porque para impulsar su crecimiento e innovación una empresa debe:

- Almacenar todos los datos relevantes sobre su negocio.
- Proporcionar el acceso de los datos.
- Analizar los datos de diferentes maneras.
- Destilar los datos hasta obtener conocimiento.

La mayoria de las grandes empresas tiene almacenes de datos destinados para informes y análisis. En el pasado esto era una inversión compleja y costosa, sin embargo actualmente se ha producido un cambio estratégico gracias al almacenamiento en la nube.

Introdución de Amazon Redshift

Los almacenamientos de datos en la nube como Amazon Redshift redujeron el costo y el esfuerzo de implementar un sistema de almacenamiento de datos, sin comprometer sus funciones, la escalabilidad y el rendimiento.

Amazon Redshift es una opción de almacenamiento de datos a escala de petabytes (facilmente puede llegar exabytes), rádida y completamente administrada que hace simple y rentable la tarea de analizar grandes volúmenes de datos. Amazon Redshift tambien puede obtener un excelente rendimiento de los motores de almacenamiento de datos en columnas realizando un procesamiento paralelo masivo (MPP).

Arquitectura Moderna de Análisis y Almacenamiento de Datos

Los datos fluyen hacia un almacén de datos desde sistemas transaccionales u 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 usuarios, científicos de datos, analistas de negocios y los que toman las decisiones acceden a los datos a través de herramientas de inteligencia empresarial (BI), clientes SQL u otras herramientas.

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.

Las canalizaciones 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

Opciones de Tecnología de Almacenamiento de Datos

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.

Son mas adecuadas para procedimientos transaccionales y no tanto para análisis.

Bases de Datos Orientadas a columnas

Las bases de datos orientadas a columnas organizan cada columna en su propio conjunto de bloques físicos. Esta funcionalidad les permite ser más eficiente en las operaciones 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 la memoria).

Dado que cada columna se empaqueta en su propio conjunto de bloques, cada bloque físico contiene el mismo tipo de datos. Cuando todos los datos son del mismo tipo de datos, la base de datos puede usar algoritmos de compresión extremadamente eficientes. Como resultado, necesita menos almacenamiento en comparación con una base de datos orientada a filas.

Ejemplos de bases de datos orientada a columnas: Amazon Redshift, Vertica, Greenplum, Teradata Aster, Netezza y Druid.

Arquitectura de Procesamiento Paralelas Masivas

Una arquitectura MPP le permite utilizar todos los recursos disponibles en el clúster para procesar datos, se mejora el rendimiento simplemente agregando más nodos al clúster.

Ejemplos de arquitectura MPP: Amazon Redshift, Druid, Vertica, Greenplum y Teradata Aster.

Código abierto compatible con MPP: Hadoop y Spark.

Análisis Profundo de Amazon Redshift

Amazon Redshift se basa en ANSI SQL, por lo que puede ejecutar consultas existentes con poca o ninguna modificación, volviendola la opción popular para los almacenes de datos empresariales.

Ofrece consultas rápidas y rendimiento de E/S para prácticamente cualquier tamaño de datos mediante el almacenamiento en columnas y la paralelización y distribución de consultas en varios nodos. Automatiza la mayoría de las tareas administrativas comunes de aprovisionamiento , configuración, supervisión, asegurar una copia de seguridad y la seguridad de los datos, lo que facilita su administración y la hace económica. Puede escalar el volumen de datos a petabytes en minutos gracias a estas facilidades.

Las consultas de esos exabytes almacenados en S3 se visualizan en Amazon Redshift Spectrum. También utiliza nodos RA3 con Redshift Managed Storage (RMS), los cuales aprovecha sus patrones de carga de trabajo y técnicas avanzadas de gestión de datos, como el desalojo automático de datos detallados y la obtención previa inteligente de datos.

Operaciones

Como servicio administrado, Amazon Redshift automatiza por completo muchas tareas operativas como:

- Clúster Performance: es un análisis automático para mantener estadísticas de tablas precisas.
- Cost Optimization: pausa y reanuda los clústeres que deben estar disponibles solo en el momento específico, para no cobrar tiempo muerto (cuando no se realizan cálculos).

Patrones de Uso Ideales

Las empresas utilizan Amazon Redshift para:

- 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.
- Ejecuta análisis según sea necesario en datos de eventos de gran volumen.
- Analizar tendencias sociales.
- Medir la calidad clínica, la eficiencia operativa y el desempeño financiero en el cuidad de la salud.

Antipatrones

Amazon Redshift no es ideal para los siguientes patrones:

- OLTP: Esto es para un sistema transaccional rápido, por lo cual puede ser una base de datos relacional o una no relacional.
- Datos no estructurados: Los datos deben llegar estructurados por un esquema definido, en el caso contrario primero deben pasar por un filtro.
- Datos BLOB: El almacenamiento de datos binarios como videos o imágenes se debe hacer por otros medios, y en Amazon Redshift guardar la referencia.