Resumen #2: "Data Warehousing on AWS"

Max Richard Lee Chung - 2019185076

Introduction

Antes, las empresas si llegaban al límite de escalado de sus motores de bases de datos, tenían que comprar o cambiar a otro motor del mismo vendedor con cambios sutiles en su semántica SQL. Amazon Redshift es un almacenamiento de datos en nube rápico, completamente administrado y gran capacidad de escalamiento que permite a las empresas disminuir los costos para desplegar sistemas de almacenamiento de datos sin tener que comprometer características, escalado y rendimiento. Tiene la opción de tener un sistema de almacenamiento por columnas que tiene un alto rendimiento paralelo.

Modern Analytics and Data Warehousing Architecture

Los datos típicamente fluyen con un sistema transaccional de un sistema a otro por medio de datos estructurados, semi estructurados o no estrucutrados para poder procesarlos, sin embargo, ¿por qué no se corren las transacciones en línea? Los almacenes de datos (data warehouse) están optimizados para operaciones de escritura por lotes y leer grandes cantidades de datos que usan esquemas desnormalizados. Las bases de datos de Procesamiento de Transacciones En Línea (OLTP) están optimizadas para operaciones de escritura continua y grandes cantidades de pequeñas operaciones de lectura con esquemas altamente normalizados. A continuación se mostrarán los diferentes servicios que ofrece AWS para la arquitectura de las pipelines.

Los servicios analíticos de AWS ayudan a las empresas tener mayor velocidad de respuesta por medio de un fácil acceso para crear almacenes y lagos de datos que inician con diferentes cargas de trabajo de análisis. Ofrece Una infraestructura de nube, cálculo y red segura dependiendo de los tipos de análisis. Un conjunto de herramientas analíticas maduras completamente integradas que cubren todos los usos comunes. Por último, otorga el mejor rendimiento, la mejor escalabilidad y el costo más bajo para realizar análisis.

Las arquitecturas analíticas normalmente siguen las fases de recolecta, almacenamiento, procesamiento y análisis y vizualización los datos. La fase recolección se realiza por medio de datos transaccionales (bases de datos relacionales (estructura compleja) o no relacionales (estructura no compleja)), registros de datos(actividades de la base de datos), flujo de datos (recolección, almacenamiento y procesamiento de grandes cantidades de datos de forma continua) e datos de Internet of Things (IoT) (dispositivos y sensores que envían señales de forma contínua para derivar esa información). Los datos pueden estar almacenados en lake house (conjunto de un lago y almacén de datos que permite consultar de forma rápida y eficiente en todos los almacenes), data warehouse (rápido análisis de información en cantidades masivas) y data mart (parte o área de un almacén de datos). El procesamiento de datos se realiza por medio del procesamiento en lotes (batch) y el procesamiento en tiempo real. EL procesamiento en lote se dividen en la extracción, transformación y carga (ETL); extraer, cargar y transformar (ELT) y el procesamiento analítico en línea (OLAP). EL primero es el proceso de extraer datos de múltiples fuentes para cargarlos en sistemas de almacenamiento de datos. El segundo es una variante del primero en donde se cargan al sistema y luego se transforman los datos cargados. Por último, los sistemas OLAP almacenan datos históricos agregados en esquemas multidimensionales

(usados para consultas, reportes y análisis). El procesamiento en tiempo real se utiliza para realizar una amplia variedad de análisis que uncluyen correlaciones, agregaciones, filtrado y muestreo. Esto permite que las empresas tengan mejor visibilidad a las actividades de los clientes y empresas.

Data Warehouse Technology Options

Las bases de datos orientadas a filas almacenan filas completas en un bloque para tener un buen rendimiento de lectura por medio de índices secundarios, el cual es preferible para procesamiento transaccional. Los data marts alivian el trabajo por medio de "sharding". Cada consulta realizada se tiene que leer todas las columnas y filas sin excepción si se quiere conocer solo una columna. Las bases orientadas a columnas organiza cada columna en un bloque, el cual permite que sea mucho más eficiente en las entradas y salidas de información de solo lectura, ya que cada bloque almacena un único tipo de dato. Arquitecturas de procesamiento massivo en paralelo (MPP) permite usar todos los recursos de un clúster para procesar información, el cual aumenta la eficiencia y rendimiento por tener mayor cantidad de nodos del clúster.

Amazon Redshift Deep Dive

Tiene la característica de Redshift Spectrum que permite una fácil entrada (escritura) y salida (consulta) de información al data lake con archivos abiertos (permite formatos como Parquet, ORC, JSON, CSV, entre otros tipos de ANSI SQL). Ofrece un rápido rendimiento con flexibilidad que incluyen un alto rendimiento en hardware, acelerador avanzada de consultas, almacenamiento eficiente y alto rendimiento en el procesamiento de consultas, vistas materializadas, gestión automática de la carga de trabajo (maxímizar el rendimiento) y vista de los resultados. Detecta automáticamente si un nodo falla para poder reemplazarlo inmediatamente y continuar con la última consulta para seguir el flujo de trabajo lo más rápido posible, en donde trata de mantener al menos 3 copias de los datos (original, réplica y un respaldo)con configuración simple. Se obtiene la elasticidad y escalabilidad necesaria para la carga de trabajo actual, sin embargo, se pueden moldear de forma independiente. Por un lado, se puede cambiar el tamaño de la elasticidad para añadir o eliminar nodos para algún recurso necesario o innecesario. Los cambios de tamaños se pueden ajustar dentro de un horario establecido. Por el otro lado, se puede ajustar el escalado de concurrencia, en donde se asigna automáticamente un cálculo extra por el aumento de solicitudes. Permite al usuario escalar y pagar por cálculos y almacenamiento de forma independiente, por lo que se puede personalizar el tamaño del clúster por medio de las necesidades.

Operations

Amazon Redshift es recomendado para ejecución de informes empresariales, análisis de ventas globales para múltiples productos, historial de ventas, información juegos, ánalisis de tendencias sociales, medir la calidad clínica, la eficiencia operativa y el desempeño financiero en salud. Amazon Redshift no es recomendado para OLTP (si se quiere un sistema transaccional rápido, mejor seleccionar otras bases de datos como Amazon Aurora o Amazon RDS o no transaccional como Amazon DynamoDB), datos no estrucutrados (si no se quiere una estrucutra, se puede usar ETL con Amazon EMR para tener los datos listos para ser cargados) y datos BLOB (si se quiere almacenar largos objetos binarios, se guardan en datos S3 para referenciarlos en RedShift).