Resúmen #1

Max Richard Lee Chung - 2019185076

Introduction

> *Almacén de datos (data warehouse) es un depósito central de información provenientes de una o más fuentes de datos.*

Los datos de una empresa deben de manterner datos relevantes de forma persistente para que cualquier persona con permisos pueda acceder, analizar y comprobar la información. Los data warehouse son muy costosos para mantener y complejos, ya que tienen una difícil escabilidad, largos tiempos de espera para mejoras de hardware, definir la cantidad de usuarios permitidos y no se pueden separa los datos.

Las empresas sufrían mucho cuando tenían que decidir si optar por una baja velocidad de extracción de datos o invertir más tiempo en mejores procesos. Amazon Redshift ofrece la solución que hace que los warehouses sean simples y rentables para analizar grandes volúmenes de datos utilizando herramientas de inteligencia empresarial.

Modern Analytics and Data Warehousing Architecture

Los almacenes de datos están optimizados para operaciones de escritura por lotes y leer grandes cantidades de datos. 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. A continuación se mostrarán los diferentes servicios que ofrece AWS para la arquitectura de las pipelines.

- AWS Analytics Services: Otorga ventajas de velocidad de respuesta aplicando los siguientes aspectos:

1. Fácil acceso a las data warehouses junto con la inicialización de diferentes análisis.
2. Una infraestructura de nube, cálculo y red segura.
3. Un conjunto de herramientas analíticas maduras completamente integradas.
4. El mejor rendimiento, la mayor escalabilidad y el costo más bajo para análisis.

- Analytics Architecture: Recolecta, almacena, procesa y analiza y vizualiza los datos.
- Data Collection

1. Transactional Data: Bases de datos relacionales (estructura compleja) o no relacionales (estructura no compleja).
2. Log Data: Historial de actividades de la base de datos.
3. Streaming Data: Recolección, almacenamiento y procesamiento de grandes cantidades de datos de forma continua.
4. Internet of Thungs (IoT) Data: manipular datos de dispositivos y sensores sin tener infraestructura.

- Data Processing

1. Batch Processing: Extract Transform Load (ETL): Proceso de extracción continua de datos de múltiples fuentes a un warehouse. Extract Load Transform (ELT): Proceso de cargar los datos almacenados a un sistema. Online Analytical Processing (OLAP): Almacena datos en esquemas multidimensionales.

2. Real-Time Processing: Procesa los datos en forma secuencial e incremental la información .

- Data Storage

1. Lake house: Conjunto de un lago y almacén de información. Permite consultar datos de la combinación y base de datos operacionales de forma rápida y eficiente.
2. Data warehouse: Rápido análisis de información en cantidades masivas.
3. Data mart: Parte o área de un almacén de datos (data warehouse).

- Analysis and Visualization: Se necesita de un buen programa de representación de información antes de recolectar y analizar los datos a adquirir.

Data Warehouse Technology Options

- Row-Oriented Databases: Almacenan filas completas en un bloque para tener un buen rendimiento de lectura por medio de índices secundarios. 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.
- Column-Oriented Databases: Organiza cada columna en un conjunto de bloques, esto permite que sea mucho más eficiente en las entradas y salidas de información. La compresión de datos es más eficiente porque cada bloque almacena un único tipo de dato.
- Massively Parallel Processing (MPP) Architectures: Permite usar todos los recursos de un clúster para procesar información. Aummenta la eficiencia y rendimiento por tener mayor cantidad de nodos del clúster como por ejemplo Amazon Redshift.

Amazon Redshift Deep Dive

- Integration with Data Lake: 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).
- Performance: 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.
- Durability and Availability: 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. Trata de mantener al menos 3 copias de los datos (original, réplica y un respaldo). Configuración simple.
- Elasticity and Scalability: Se obtiene la elasticidad y escalabilidad necesaria para la carga de trabajo actual, sin embargo, se pueden moldear de forma independiente. Dentro de dichas, por un lado, configuraciones 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.
- Amazon Redshift Managed Storage: 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

Automatiza completamente las tareas operacionales, tales como el rendimiento del clúster y la optimización de costo.

- Ideal Usage Patterns:

1. Ejecución de informes empresariales
2. Análisis de ventas globales para múltiples productos
3. Historial de ventas
4. Información juegos
5. Análisis de tendencias sociales
6. Medir la calidad clínica, la eficiencia operativa y el desempeño financiero en salud

- Anti-Patterns:

1. 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.
2. Unstructured data: Si no se quiere una estrucutra, se puede usar ETL con Amazon EMR para tener los datos listos para ser cargados.
3. BLOB data: Si se quiere almacenar largos objetos binarios, se guardan en datos S3 para referenciarlos en RedShift.