

Introducing Amazon Redshift

Antes, cuando el volumen de datos crecía o se quería aceptar a más usuarios, se tenía dos opciones: pasar por un proceso de mejora (lo que era bastante costoso) o hacer consultas de manera que el performance disminuía (era lento). La llegada de cloud data warehouses cambió la forma en la que entidades piensan.

Amazon Redshift es un **data warehouse** que simplifica el análisis de datos con inteligencia de negocios (IB). Usa almacenamiento columnar con procesamiento masivo paralelo (MPP) a un bajo costo. Permite crecer al guardar datos en Amazon S3.

Modern Analytics and data warehousing architecture

Los datos entran a una data warehouse desde sistemas transaccionales y bases de datos relaciones, que incluyen datos estructurados, semi estructurados y no estructurados. Los usuarios pueden acceder estos datos mediante herramientas de business intelligence (BI) y clientes SQL

Data warehouse y Online transaction processing database (OLTP)

| Data warehouse | OLTP |
|--|--|
| Optimizada para lotes de escritura y lectura de muchos datos | Optimizada para escritura continua y operaciones de lectura pequeñas |
| Esquemas no normalizados | Esquemas altamente normalizados |
| Alto rendimiento de datos | Alto rendimiento de transacciones |

Es recomendable construir un data pipeline eficiente para extraer los datos del sistema fuente y convertirlos en un esquema funcional para data warehousing

AWS analytics services

Ayuda a convertir datos para **dar respuestas mediante el análisis de servicios integrado**. La rapidez con la que se den respuestas implica menor tiempo de configuración y conexión de los servicios de análisis en la nube. AWS da facilidad para construir data warehouses y data lakes, almacenamiento seguro en la nube, un stack integrado para el análisis, escalabilidad, un bajo costo y poco tiempo de producción.

Los datos están catalogados y listos para el análisis, usa **machine learning** para los records relacionados. También tiene un set de servicios para el análisis, estos se encuentran integrados en las capas de infraestructura, lo que permite aprovechar características brindadas que permiten reducir costos y aumentar velocidad.

Analytics architecture

Los **pipelines** se diseñan para hacer el manejo de volúmenes grandes de datos que ingresan de bases de datos, aplicaciones y otros. Estos tienen etapas:

1. Toman datos : Puede recolectar distintos tipos de datos, como: Datos transaccionales, datos Log, datos de Streaming y datos Entrada/salida.

2. Procesan datos : Los datos pueden ser analizados para extraer información valiosa. Hay 2 tipos:

- **Batch:** Esta el **Extract Transform Load** (ETL) que procesa datos extraídos de múltiples fuentes y la carga en un sistema warehouse, es continua y bien definida. También esta **Extract Load Transform** (ILT) que extrae los datos y los carga al sistema, para después hacer el análisis. Por último, esta el **Online Analytical Processing** (OLAP) que guarda datos de esquemas multidimensionales, permitiendo extraer datos de varias dimensiones.
- **Real-time:** Procesamiento de información de manera secuencial e incremental, brinda visibilidad en aspectos como actividad de clientes. Requiere alta concurrencia y escalabilidad.

3. Almacenamiento de datos :

- **Lake House:** Combinación de data warehouse y data lakes. permite hacer consultas en ambos y permite almacenar datos en archivos de formato abierto.
- **Data warehouse:** análisis rápido en grandes volúmenes de datos.
- **Data mart:** Es un data warehouse especializado en un área específica, son simples de diseñar y construir.

4. Visualizar datos : Se pueden ver los datos mediante las mismas herramientas usadas para procesarlos.

Data warehouse technology options

- **Row oriented databases:** Almacena los datos como un **bloque de filas**, son más usadas en procesamiento de transacciones con OLTP. Se pueden optimizar mediante el uso de vistas, uso de particiones, etc. Esta forma no es la mejor, pues en la lectura de datos se debe leer sobre todas las columnas de todas las filas, en lugar de solo en las columnas que necesito.
- **Column Oriented databases:** Almacena los datos como si cada **columna fuera un bloque**, esto hace que sea más eficiente para consultas de lectura, porque solo se lee la columna que me interesan, estas se usan más en el data warehousing.
- **Massively Parallel Processing (MPP) architectures:** Permite usar todos los recursos del cluster para el procesamiento de datos, permite mejorar el performance agregando nodos al cluster.

Amazon redshift deep dive

Ofrece beneficios para warehousing de buen rendimiento, además, incluye eficiencia en compresión y pocos requerimiento de almacenamiento. Permite consultas rápidas usando el almacenamiento columnar, distribuyendo consultas entre nodos y usando paralelismo. Automatiza tareas como configuración, monitoreo, backups y otros, que facilitan el manejo.

Integration with data lake

Facilita consultas de lectura y escritura, permitiendo consultas de archivos de formato abierto, se pueden exportar datos y automáticamente, redshift se hace cargo del formato de los datos, también permite datos de tablas externas, lo que da flexibilidad y estructura.

Performance

- Hardware de alto performance que permite múltiples nodos.

- Aqua, que permite la aceleración de consultas, mediante el filtrado y agregaciones.
- Vistas materializadas para el almacenamiento de calculos realizados previamente.
- Uso de machine learning que usan algoritmos para predecir consultas futuras.
- Result caching para responder rapidamente a consultas repetidas.

Durability and availability

Automaticamente detecta y remplace nbodos que fallan dentro del clúster, Redshift intenta mantener al menos tres copias de los datos: el plincipal, una replica y un backup. Se puede crear un mirror para que se gestione la replicacion y los failovers.

Elasticity and Scalability

Permite escalar procesamiento y almacenamiento y solo pagar por lo que se usa. Hay dos formas de procesar la escalabilidad, esta **Elastic resize** que agrega los nodos necesarios para la carga de trabajo y los remueve cuando este termina, este proceso pue ser automatizado mendiante un schedule. Tsmbien esta **Concurrency Scaling** que aumenta la capacidad de forma automática cuando es necesario aumentar la concurrencia.

Operations

- **Amazon Redshift Advisor:** Permite mejorar el performance y disminuir los costos en los clústers, esta herramienta ofrece recomendaciones basadas en las cargas de trabajo del clúster.
- **Interfaces:** Redshift brinda una consola en linea que permite correr consultas SQL, tambien hay drivers que permiten trabajar con clientes SQL.
- **Security:** Redshift solo permite el acceso a los datos desde el nodo lider del clúster, permitiendo una capa de seguridad. El manejo de seguridad en la base de datos se hace mediante usuarios que tienen privilegios.
- **Cost Model:** No requiere compromisos a largo plazo, los cargos se basan en el tamaño y la cantidad de nodos que hay en un clúster. Además no hay cargos extras por backups.
- **Ideal usage patterns:** Redshift es ideal para OLAP. Redshift le da soporte a datos semiestructurados y extiende datos entre data warehouse y data lake, lo que permite hacer análisis de volúmenes grandes de datos.
- **Anti-patterns:** OLTP (OLTP es mejor si lo que se requiere es un sistema transaccional rápido), datos no estructurados (Datos en redshift deben ser estructurados y definidos por un esquema) y BLOB data (Para almacenamiento de archivos binario grandes es preferible almacenar en S3).