# Resumen 2

# Dillan Almendares Barrantes 2020033336

---

## Amazon Redshift

Amazon Redshift es uno de los sistemas de cloud data warehouse más importantes, es utilizado por grandes empresas por ofrecer un servicio más rentable para analizar grandes cantidades de datos, lo que antes significaba un sacrificio en rendimiento o una inversión en hardware, Amazon Redshift lo resuelve con almacenamiento columnar y procesamiento paralelo masivo, con soporte para escala de petabytes. Además del buen rendimiento que tiene por las bases de datos orientadas a columnas y el procesamiento paralelo masivo, automatiza ciertas tareas de configuraciones, monitoreo, respaldo y seguridad del data warehouse. Con Redshift Spectrum se puede consultar y escribir datos de un data lake con documentos de formato abierto.

Algunas funciones que ofrece Amazon Redshift para mejorar el rendimiento:

- AQUA (Advanced Query Accelerator): Una caché distribuida para ejecutar tareas demandantes como filtros y agregaciones más cerca de la capa de almacenamiento.
- Para optimizar el almacenamiento columnar, tiene el sistema de compresión AZ64 para datos numéricos y de fechas.
- Permite mantener vistas materializadas para las consultas pesadas frecuentes que podrían ralentizar el sistema.
- Con el uso de machine learning puede cambiar la configuración para adecuarse a las distintas cargas de trabajo o de actividad de los usuarios de manera que los recursos se empleen para el mayor rendimiento posible en cada momento.

## Estructuras de data warehousing

Que los datos se almacenen en data warehouses para ser analizados puede parecer innecesario al ser posible hacerlo en los procesos de los sistemas transaccionales o OLTP que generan los datos, pero estos sistemas están hechos operaciones de escritura continua y gran cantidad de pequeñas operaciones de lectura, por el otro lado, los data warehouse se optimizan para tener operaciones de escritura por lotes y lecturas de gran tamaño. Además, los data warehouses tienen esquemas menos normalizados como el de estrella o copo de nieve para que tengan un rendimiento superior en datos grandes. Para aprovechar la separación del data warehouse y los OLTP, es necesario tener un pipeline eficiente que se encargue de recolectar los datos del OLTP, los convierta al esquema requerido y los almacene en el data warehouse.

## Tecnologías de data warehouse

### Bases de datos orientadas a filas

Guardan la totalidad de la fila en bloques físicos, con índices secundarios alcalzan un alto rendimiento en operaciones de lectura. Rinden mejor en OLTP pero en caso de usarse en un data warehouse se optimiza con las siguientes técnicas:

- Usar vistas materializadas.
- Crear índices para cada posible combinación de predicados.
- Particionar datos.
- Utilizar joins basados en índices.

El problema que tienen es que las consultas leen todas las columnas de todas las filas seleccionadas, indiferentemente si la consulta no requiere de todas las columnas, provocando cuellos de botella.

**Bases de datos orientadas a columnas**

Cada columna tiene su bloque físico, así cada columna de una fila es independiente de las demás. Este formato tiene mayor rendimiento en las consultas de lectura por poder ahorrarse el leer las columnas que no fueron solicitas. Son mejor opción para data warehouse que las bases de datos orientadas a filas. Otra ventaja es que, al guardar las columnas en distintos bloques, cada bloque contiene un mismo tipo de datos que permite aplicar mejores algoritmos para comprimir dichos bloques y reducir el espacio que ocupan en disco.

**Arquitecturas de procesamiento paralelo masivo**

Utilizan todos los recursos disponibles del cluster para acelerar el procesamiento de datos.

## Operaciones

Amazon Redshift automatiza el proceso de analizar el rendimiento del cluster para asegurar que el almacenamiento es eficiente y detener clusters cuando la demanda operacional no los requiera.

Existe una función llamada Amazon Redshift Advisor que estudia las métricas del cluster para recomendar cambios que mejorarían el rendimiento.

**Seguridad**

Con Amazon Virtual Private Cloud (Amazon VPC) se puede ejecutar Amazon Redshift en una nube privada donde se puede configurar un firewall para que regule el tráfico de la red. Para encriptar los datos mientras se envían, soporta SSL. Como capa de protección adicional, los nodos de computo sólo pueden guardar datos, es el nodo líder del cluster el único con capacidad de acceder a ellos.

**Modelos óptimos**

Amazon Redshift es ideal para grandes cantidades de datos empresariales como:

- Analizar las ventas globales de productos.
- Datos del comercio bursátil.
- Analizar tendencias sociales.
- Análisis de los anuncios mostrados.

**Modelos desaconsejados**

Los modelos que no aprovechan las capacidades de Amazon Redshift son:

- OLTP: Para sistemas transaccionales es mejor utilizar bases de datos relacionales, además que desperdicia el potencial de análisis pensado para data warehouse.
- Datos desestructurados: No son compatibles con el funcionamiento de Amazon Redshift.
- BLOB data: Amazon Redshift sólo puede procesar los metadatos pero no los archivos de este tipo.