



MODELO DE DATAWAREHOUSE

Carlos Adrián Mata Nevárez

Extracción de Conocimiento en Bases de
Datos

Maestro: Enrique Mascote
IDGS91N

CONTEXTO

Propósito de Redshift en un proyecto de Data Warehouse

- Amazon Redshift es un servicio de almacenamiento y análisis de datos en la nube diseñado para procesar grandes volúmenes usando SQL estándar.
- Se utiliza para construir Data Warehouses escalables, de alto rendimiento y orientados a análisis avanzados.
- Permite integrar, transformar y analizar datos provenientes de múltiples fuentes dentro de un modelo analítico centralizado.

CARACTERÍSTICAS CLAVE

1. Arquitectura MPP (Massively Parallel Processing)

Divide las consultas en partes y las ejecuta en paralelo para acelerar el análisis de datos masivos.

2. Nodo Líder y Nodos de Cómputo

El líder coordina las consultas y los nodos de cómputo almacenan y procesan los datos.

3. Escalabilidad elástica

Permite aumentar o reducir nodos según la demanda del proyecto sin interrumpir el servicio.

4. Integración con servicios AWS

Conecta fácilmente con S3, Glue, Data Pipeline, Kinesis, DynamoDB, etc.

5. Almacenamiento en columnas

Optimiza consultas analíticas, reduce I/O y mejora tiempos de respuesta.

VENTAJAS Y DESVENTAJAS

Ventajas

1. Rendimiento alto: gracias a MPP y almacenamiento columnar.
2. Totalmente administrado: no requiere instalación ni mantenimiento físico.
3. Escalable: ajusta capacidad bajo demanda sin tiempos muertos.
4. (opcional) Integración nativa con AWS: facilita la ingestión de datos.



Desventajas

1. Costo variable: puede ser elevado según el volumen de datos.
2. Dependencia de AWS: funciona mejor si todo el ecosistema ya usa Amazon.
3. Menor flexibilidad para ETL complejo: se apoya en Glue o herramientas externas.



PROCESO DE IMPLEMENTACIÓN

Flujo típico ELT en Redshift

1. Origen de datos

- Bases transaccionales, archivos CSV/JSON, APIs, logs, S3.

2. Staging (capa temporal)

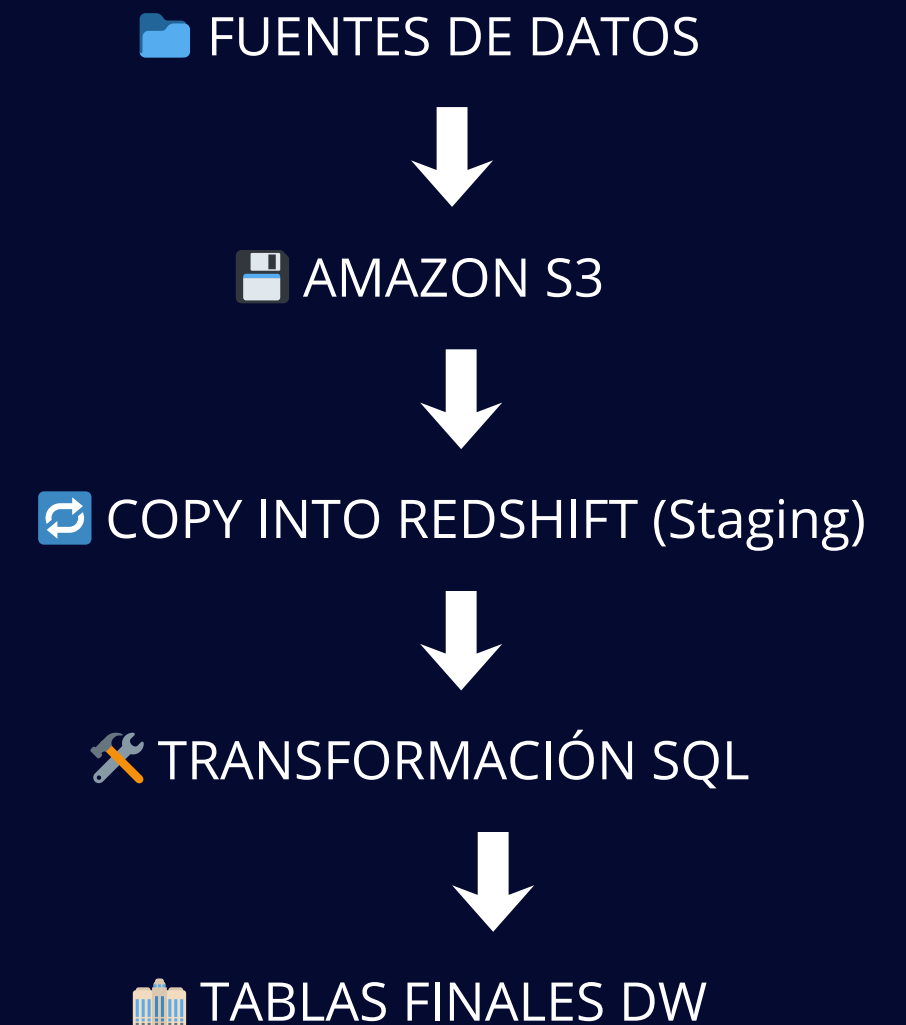
- Los datos se copian primero a tablas temporales para validación y limpieza.

3. Transformación (ELT dentro de Redshift)

- Se usan consultas SQL para limpiar, integrar y transformar datos.

4. Carga al Data Warehouse

- Se mueven los datos transformados a tablas finales optimizadas para análisis.



CONSIDERACIONES FINALES



Retos comunes

- Costo elevado si no se optimizan consultas.
- Problemas de rendimiento si las tablas no usan SORTKEY y DISTKEY adecuados.
- Necesidad de gobernanza de datos al trabajar con múltiples fuentes.

Recomendaciones prácticas



- Usar compresión automática para reducir almacenamiento.
- Automatizar cargas con AWS Glue o Lambda.
- Implementar monitoreo con CloudWatch y Redshift Advisor.



GRACIAS

REFERENCIAS

Amazon Web Services. (2024). Amazon Redshift Documentation. AWS.
<https://docs.aws.amazon.com/redshift>

Amazon Web Services. (2024). Amazon Redshift Best Practices. AWS
Whitepaper.

Mahmood, Z. (2021). Cloud Data Warehousing with Amazon Redshift.
Springer.

Amazon Web Services. (2023). AWS Data Analytics Blog. AWS.

Gupta, M. (2020). Modern Data Warehousing with AWS. O'Reilly Media.