MODULO 09 - EJERCICIO 05-A

ALEXIS YURI M.

Diseña un flujo ETL utilizando AWS Glue que resuelva una necesidad específica de una organización (por ejemplo: integrar datos de ventas dispersos para análisis en un data lake). Debe definir cláramente los origenes, las transformaciones necesarias y el destino.

Diseño de Flujo ETL para Ventas Multicanal.

El objetivo es crear un proceso que extraiga los datos de los tres canales (tiendas físicas, e-commerce y marketplace), los unifique y prepare, y los cargue en un destino listo para el análisis, utilizando servicios de AWS:

Extracción de datos.

El primer paso es consolidar los datos de las tres fuentes, a pesar de sus diferentes formatos y frecuencias de actualización.

- Tiendas físicas (CSV): Los archivos CSV se subirían a un bucket de Amazon S3, que actuaría como una capa de aterrizaje o staging area.

- E-commerce (base de datos en la nube): Se usaría un conector de AWS Glue para conectarse directamente a la base de datos y extraer la información de manera programada.

- Marketplace (hojas de cálculo): Los reportes en hojas de cálculo se convertirían a un formato estándar (como CSV) y se subirían a un bucket de Amazon S3.

Transformación de datos.

La etapa de transformación es la más crítica, ya que se debe resolver la duplicación de datos, las diferentes codificaciones y la falta de unificación. Para esto usaremos AWS Glue.

- Catalogación de datos: Se usaría un Crawler de AWS Glue para descubrir automáticamente el esquema de los archivos en S3 y de la base de datos del e-commerce, creando tablas en el Catálogo de Datos de AWS Glue.

- Limpieza y unificación: Se crearía un Job de ETL en AWS Glue para realizar las siguientes transformaciones:

- Unificar datos: Consolidar las tres fuentes en un solo conjunto de datos.

- Normalizar: Corregir las diferentes codificaciones de productos y clientes.

- Eliminar duplicados: Identificar y eliminar registros repetidos para obtener una visión única del cliente y los productos.

Carga de datos.

Finalmente, los datos limpios y transformados se cargarán en un destino optimizado para el análisis y la generación de informes.

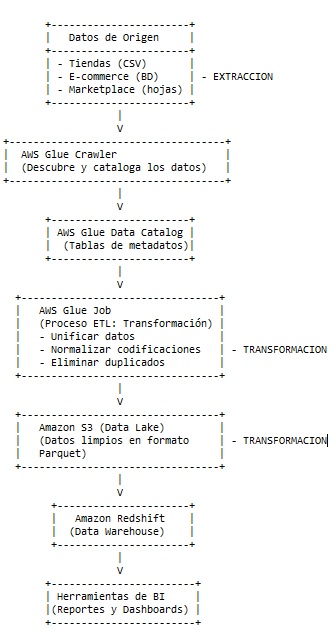
Destino: Se usaría un Data Warehouse en la nube, como Amazon Redshift, para almacenar el conjunto de datos final. Redshift está diseñado para *consultas analíticas* de alto rendimiento, lo que permitiría a la gerencia obtener una visión consolidada de su desempeño comercial y generar reportes en tiempo real.

Formato optimizado: Dentro del proceso de transformación, los datos se podrían convertir a un formato columnar (como Parquet), lo que optimizaría la velocidad y el costo del almacenamiento y las consultas en Redshift.

Automatización y monitoreo

El flujo se podría automatizar utilizando un Trigger en AWS Glue para ejecutar el Job de ETL de manera diaria o semanal, según la necesidad. Además, se configuraría una alarma en Amazon CloudWatch para notificar al equipo si el proceso falla.

Representación Visual del Flujo ETL.

****