Diseñar un flujo ETL para resolver un caso real

Este documento describe el diseño de un flujo ETL utilizando AWS Glue para resolver el caso de un sistema de ventas multicanal en una empresa de productos para el hogar. El objetivo es integrar datos de distintos orígenes, aplicar transformaciones y consolidarlos en un destino común para habilitar reportes y análisis confiables.

Caso: Sistema de ventas multicanal

La empresa vende a través de tres canales principales:
- Tiendas físicas: exportan ventas en CSV desde cajas locales.
- E-commerce propio: guarda datos en una base de datos en la nube casi en tiempo real.
- Marketplace externo: entrega reportes semanales en hojas de cálculo.

Problemas identificados:
- Datos duplicados con nombres distintos para el mismo producto.
- Información de clientes fragmentada.
- Rezago en reportes del marketplace.
- Falta de visión consolidada para inventarios y campañas de marketing.

Diseño del flujo ETL

Orígenes de datos	Transformaciones necesarias	Destino	Resultado esperado
CSV (tiendas físicas)	Unificación de códigos, limpieza de duplicados	Amazon S3 (Parquet)	Datos estandarizados listos para análisis
BD en la nube (e-commerce)	Normalización de fechas y monedas	Amazon Redshift	Reportes en tiempo casi real
Excel (marketplace)	Conversión a Parquet, integración con catálogo de productos	Amazon S3 + Redshift	Consolidación semanal de ventas

Automatización y orquestación

- Crawlers en Glue para descubrir y catalogar fuentes.
- Jobs ETL en PySpark para aplicar las transformaciones.
- Triggers programados:
  * Diarios para CSV y BD.
  * Semanales para marketplace.

Conclusión

El flujo ETL propuesto con AWS Glue permite consolidar los datos de ventas en un data lake en Amazon S3 y cargarlos en Amazon Redshift para análisis avanzado. Esto asegura una visión unificada de clientes y productos, optimiza reportes financieros y soporta decisiones ágiles en marketing e inventarios.