Project Overview

Tuesday, March 18, 2025 10:58 AM

Building a Retailer Data Lake

Introduction

In the modern retail industry, managing and analyzing large volumes of data is essential for staying competitive. A data lake provides a centralized repository to store, process, and analyze vast amounts of structured and unstructured data efficiently.

Why a Data Lake?

A data lake enables retailers to store and analyze data efficiently, unlocking valuable insights for business growth.

Key benefits include:

- Centralized Data Access: Provides a single repository for all retailer data, accessible across departments.
- Scalability: Handles large volumes of data seamlessly without performance issues.
- Data Flexibility: Supports multiple data formats (structured, semi-structured, unstructured) and various query engines.
- Advanced Analytics & AI: Empowers businesses to perform deep analytics, uncover trends, and enhance decision-making.

Data Sources

The system ingests data from three main sources:

i. **MySQL Retailer Database (mysql-retailer):**
   Contains key business data such as:
   i. products
   ii. categories
   iii. customers
   iv. orders
   v. order_items

ii. **MySQL Supplier Database (mysql-supplier):**
   Contains supplier-related information:
   i. suppliers
   ii. product_suppliers

iii. **API Reviews (api-reviews):**
   Captures customer feedback from external sources:
   a. customer_reviews

Data Landing in GCS

Once extracted, the data is landed into Google Cloud Storage (GCS) under separate folders for easy organization:

- retailer-db (Retailer-related data)
- supplier-db (Supplier-related data)
- customer_reviews (Reviews data from APIs)

Data Analysis in BigQuery

The data moves from GCS to BigQuery using the Medallion Architecture, which consists of three layers:

- Bronze Layer: Raw data from GCS is ingested into BigQuery as-is, without transformations.
- Silver Layer: Data is cleaned, standardized, and transformed to improve quality.
- Gold Layer: Final, business-ready tables are created for analytics and reporting.

Visualization in Looker BI

After Analysis, Looker BI is used to generate dashboards and reports based on gold-layer tables.

Workflow Orchestration with Airflow

All processes (data extraction, loading into GCS, transformation in BigQuery) are managed using Apache Airflow, ensuring automation, scheduling, and monitoring.

Key Benefits

✅ Scalability – Handles large datasets efficiently.
✅ Data Quality – Ensures clean and enriched data at the Gold Layer.
✅ Automation – Airflow orchestrates workflows efficiently.
✅ Real-time Insights – Looker BI provides dashboards for business decision-making.