E-commerce Sales Data Pipeline for Olist Dataset

Project Overview

This project implements an end-to-end ETL (Extract, Transform, Load) pipeline to process and analyze the Olist Brazilian E-commerce public dataset. The primary goal is to transform raw, multi-file transactional data into a clean, unified, and aggregated format suitable for business intelligence and strategic analysis.

The pipeline extracts data from multiple CSV files, cleans and joins them, performs key business aggregations, and loads the final tables into a data warehouse. This provides clear, actionable insights into sales performance, customer behavior, and product trends.

Business Problem & Objectives

The Olist dataset is distributed across nine different CSV files, making direct analysis difficult. Business stakeholders cannot easily answer fundamental questions about performance. This project solves that by creating a single source of truth to answer critical business questions, such as:

- What is our monthly sales revenue and order volume over time?
- Which product categories are the most profitable?
- Who are our most valuable customers based on spending and order frequency?
- Which geographical regions are driving the most sales?

Final Aggregated Tables

As per the client request, the pipeline generates the following four key aggregated tables, which serve as the final output for BI dashboards:

1. **monthly_revenue**: Tracks top-line metrics over time.
    - **Columns**: report_month, total_revenue, total_orders
2. **category_performance**: Ranks product categories by revenue.
    - **Columns**: product_category, total_revenue, unique_products_sold
3. **customer_value**: Identifies the most valuable customers.
    - **Columns**: customer_unique_id, total_spend, total_orders, first_order_date, latest_order_date
4. **regional_sales**: Summarizes sales by customer location.
    - **Columns**: customer_state, total_revenue, total_customers

ETL Architecture

The pipeline follows a standard ETL process:

1. Extract

- Raw data is extracted from the various source CSV files provided in the Olist dataset.

2. Transform

- **Data Cleaning**: Handles missing values, corrects data types, and removes duplicates.
- **Data Joining**: Combines data from orders, payments, products, and customers tables to create a unified view of each transaction.
- **Feature Engineering**: Translates product categories from Portuguese to English for better readability.
- **Aggregation**: Performs the required calculations (SUM, COUNT, etc.) to create the final analytical tables as specified above.

3. Load

- The final, aggregated tables are loaded into a structured data warehouse, making them available for querying by BI tools like Tableau, Power BI, or Google Looker Studio.

Tech Stack

- **Language**: Python, SQL
- **Data Manipulation**: pyspark, pysparksql
- **Data Storage (Source)**: CSV files

Dataset

This project uses the **Olist Brazilian E-commerce Public Dataset**, which can be downloaded from Kaggle.

The key source files used are:

- olist_orders_dataset.csv
- olist_order_items_dataset.csv
- olist_order_payments_dataset.csv
- olist_products_dataset.csv
- olist_customers_dataset.csv
- product_category_name_translation.csv