FMCG Supply Chain Data Pipeline Project Documentation

Project Github Link

FMCG Medallion Architecture Data Pipeline Using Apache Airflow and PostgreSQL

Project Description

This project builds an **end-to-end modular and scalable data pipeline** for a **Fast Moving Consumer Goods (FMCG) supply chain** analytics system. It uses the **Medallion Architecture** (Bronze, Silver, Gold) and orchestrates workflows using **Apache Airflow** with data stored and processed in **PostgreSQL**. The pipeline is fully containerized using **Docker Compose** and supports export to **Power BI** for reporting.

The pipeline handles raw data ingestion, cleansing, schema normalization, aggregation, and business metric computation to support real-time and scheduled analytics needs.

Tech Stack

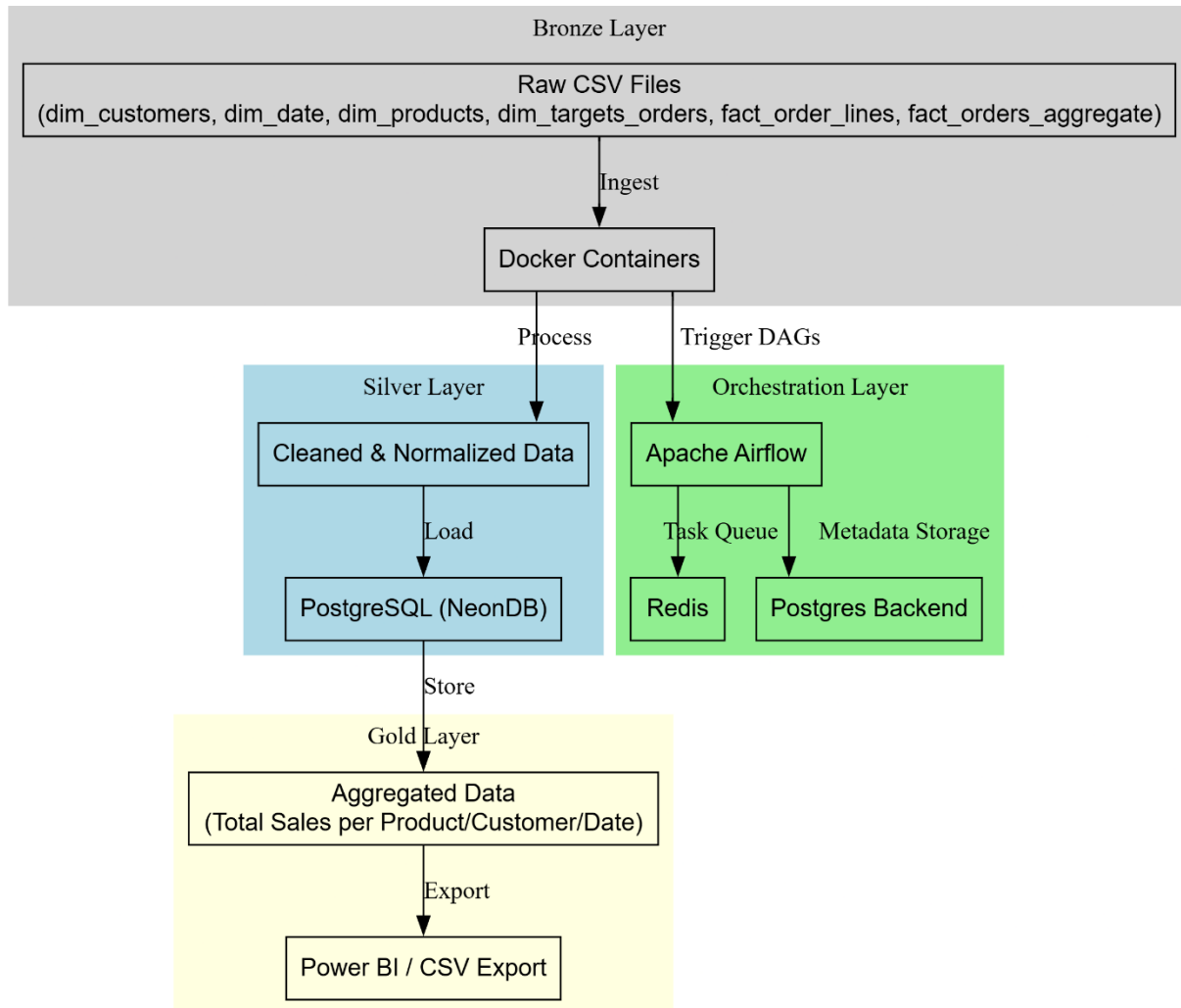| Component	Technology
Orchestration	Apache Airflow
Storage	PostgreSQL
Broker	Redis (Celery Executor)
Versioning	Docker + Docker Compose
BI Tool	Power BI / CSV export
UI Access	Airflow UI, PGAdmin

Datasets Used

File Name	Description
dim_customers.csv	Customer master data
dim_date.csv	Calendar date dimension
dim_products.csv	Product master with category, brand, etc.
dim_targets_orders.csv	Sales targets per customer-product
fact_order_lines.csv	Line-level order facts

Airflow DAG Structure

> Airflow DAG: medallion_pipeline_dag.py

1. **bronze_load_dag**
   - o   Task 1: Load raw CSVs to PostgreSQL raw tables
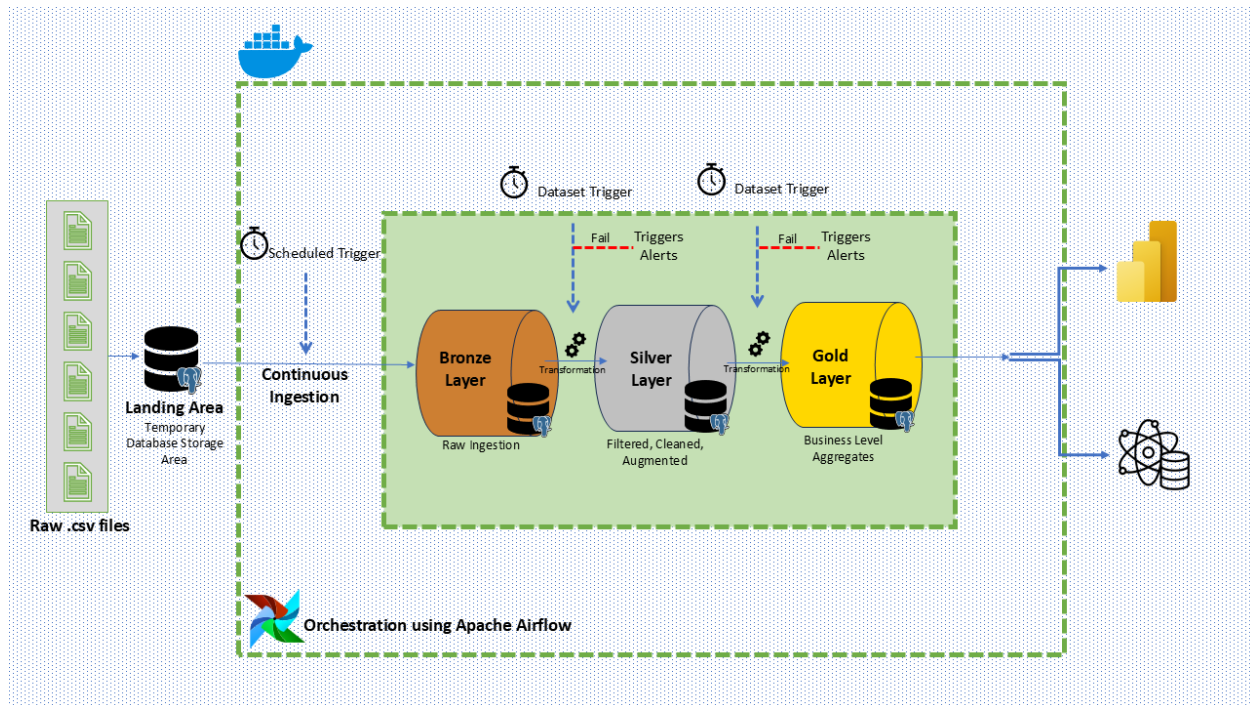   - o   Task 2: Validate row count and log
2. **silver_transform_dag**
   - o   Task 1: Clean nulls, standardize types
   - o   Task 2: Join dimension tables to facts
   - o   Task 3: Save to staging tables
3. **gold_aggregation_dag**
   - o   Task 1: Calculate business KPIs
   - o   Task 2: Create aggregation tables (e.g. `agg_daily_sales`)

          o   Task 3: Export gold tables to Power BI / CSV
4. **monitoring_dag**
          o   Monitors task failures and sends alerts

Pipeline Tasks and Business Logic

Bronze Layer Tasks

- **Load Raw Files** into tables like:
- No data manipulation, only ingestion.
- Validated for schema conformity.
- Schedule: Daily at 01:00 AM

Silver Layer Tasks

- **Data Cleansing**:
    - Replace missing/null values.
    - Normalize data types.
- **Schema Normalization**:
    - Create surrogate keys.
    - Join with `dim_date` for temporal logic.

Gold Layer Tasks

- **KPI Aggregations**:
  - Total daily orders
  - Target vs achieved orders
  - Revenue by product and region
- **Export** to Power BI folder or `.csv`

Deployment & Infrastructure

- Docker Compose for local development:

  Services: postgres, airflow-webserver, airflow-scheduler

- Data Warehouse: postgres NeoDB Local Server

Monitoring and Logging

- All DAGs log to `/logs` volume.
- PGAdmin used for DB monitoring.

Deployment Instructions

1. Clone the repo: `git clone <repo_url>`
2. Navigate: `cd FMCG_Supply_Chain`
3. Set environment variables in `.env`
4. Build services: `docker compose build`
5. Start pipeline: `docker compose up airflow-init && docker compose up`
6. Access Airflow: http://localhost:8080