Data Pipeline Assignment: E-Commerce Data Processing

# 1. Project Overview

This project demonstrates the implementation of a layered data pipeline using PySpark on Databricks Community Edition. The pipeline processes e-commerce sales data through the following layers: Raw, Processed, Enriched, and Aggregate. Each layer is designed for specific transformations, validation, and data quality checks.

# 2. Data Sources

|  |  |  |
| --- | --- | --- |
| File/Source | Format | Description |
| Customer.csv | CSV | Customer details |
| Products.csv | CSV | Product catalog |
| Orders.json | JSON | Order transactions |

# 3. Pipeline Architecture

The pipeline follows a layered architecture:

- Raw Layer: Ingests data from CSV/JSON files, cleans column names, and adds metadata.

- Processed Layer: Cleans and standardizes data fields (e.g., trimming strings, formatting dates, casting types).

- Enriched Layer: Joins customer, product, and order tables to create a business-ready dataset.

- Aggregate Layer: Computes profit summaries by year, category, sub-category, and customer.

# 4. Validation & Data Quality Checks

Key validation steps are implemented at each layer to ensure data integrity and consistency:

* • Column count and row count verification.
* • Null checks for critical fields (e.g., customer\_id, order\_id).
* • Duplicate detection for unique keys.
* • Date format checks for order and ship dates.
* • Profit rounding validation (2 decimal places).
* • Data type casting (e.g., profit, price).

# 5. Unit Testing Strategy

Unit tests are implemented using PySpark and Python's unittest framework. The tests cover:

* • Raw Data Layer: Schema validation, null checks, and duplicate detection.
* • Processed Data Layer: Data type consistency, string trimming, and format validation.
* • Enriched Layer: Join integrity, null checks in joins, and row count consistency.
* • Aggregate Layer: Profit summarization and partitioning checks.

# 6. Key Outputs

Final tables created by the pipeline:

|  |  |
| --- | --- |
| Table Name | Description |
| processed\_customer | Cleaned customer data |
| processed\_product | Cleaned product data |
| processed\_order | Cleaned order data |
| table\_enriched\_order | Joined customer, product, and order data (partitioned by country) |
| table\_customer\_profit | Customer-level profit summary (partitioned by country) |
| table\_agg\_profit | Profit aggregated by year, category, and customer (partitioned by year) |