E-Commerce Analytics Pipeline

Transforming raw e-commerce data into actionable insights using a robust ELT pipeline. This project leverages Databricks, Spark, Delta Lake, and AWS S3 to deliver key business metrics.

Made with GAMMA

Abstract: From Raw Data to Insights

Data Ingestion

Orders, order_items, and products ingested into a Bronze layer.

Data Cleaning

Cleaned and standardized in a Silver layer for reliability.

KPI Aggregation

Aggregated into business KPIs in a Gold layer.

Automated Workflows

Scheduled jobs with Databricks Workflows and error handling.

Made with GAMMA

Introduction: Scalable Analytics for E-commerce

Modern e-commerce platforms generate vast transactional data. To effectively analyze performance and customer behavior, a scalable data pipeline is essential.

This project utilizes Databricks and Spark for distributed computation, and Delta Lake for ACID-compliant storage, creating a multi-layer architecture (Bronze → Silver → Gold) for retail analytics.

Made with GAMMA

Problem Statement: Unreliable Data to Business-Ready Insights

1. ### Raw Data Challenges

   Null values, inconsistent formats, and duplicate records.

2. ### Automated ELT Pipeline

   Ingest, standardize, clean, and transform transactional data.

3. ### Dimensional Modeling

   Create fact and dimension tables for structured analysis.

4. ### Business KPIs

   Generate key performance indicators using Gold aggregations.

Made with GAMMA

Dataset Description: Core E-commerce Data

Orders Dataset

Customer purchase, delivery, and freight information.

- order_id
- customer_id
- order_status

Order Items Dataset

Line-item transactions for each order.

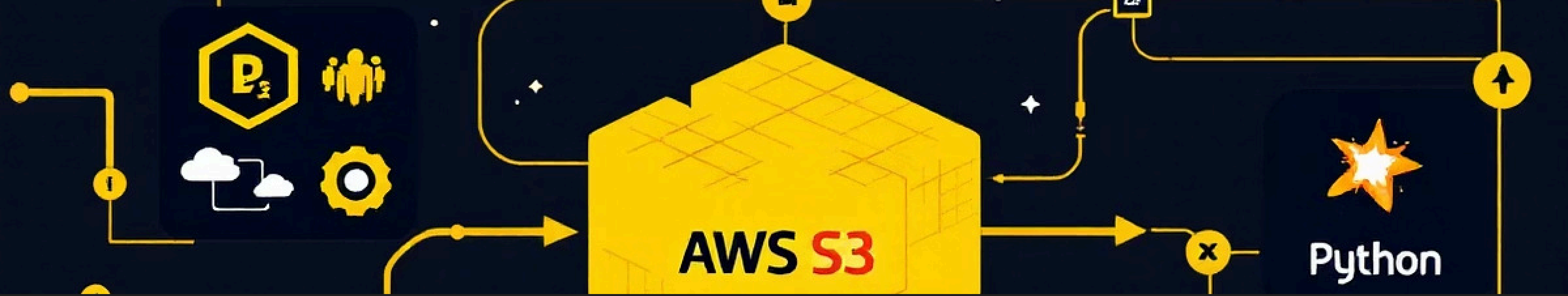
- order_id
- product_id
- price

Products Dataset

Product metadata including category and dimensions.

- product_id
- product_category_name
- product_weight

Made with GAMMA



Project Objectives & Technologies

Objectives

- Ingest raw CSV from AWS S3
- Clean and standardize data
- Create dimension and fact tables
- Compute business KPIs
- Implement error handling and logging
- Schedule daily job execution

Tools & Technologies

- Databricks (PySpark + Delta Lake)
- AWS S3
- Unity Catalog
- Python / Spark DataFrames
- Databricks Workflows

Made with GAMMA

Data Transformation Summary

01

Data Joining

Orders, order_items, and products are joined.

02

Table Creation

fact_sales, dim_product, and dim_customer tables are created.

03

Gold Computations

Item revenue, monthly revenue, customer LTV, category revenue, and delivery time metrics.

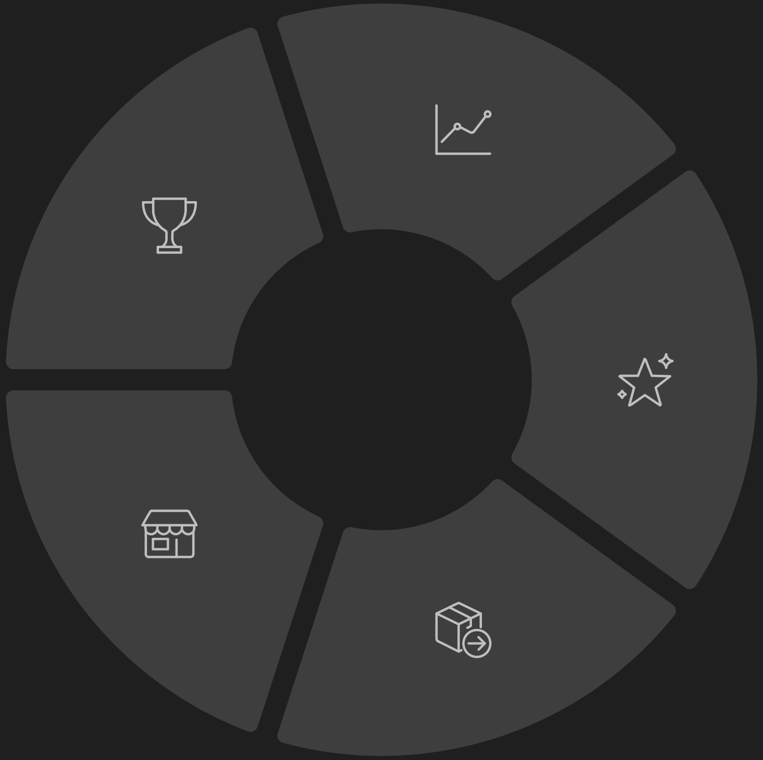
04

Analysis-Ready Data

Gold data prepared for dashboards and KPIs.

Made with GAMMA

Key Insights & Interpretation



Best Selling Category

Identified highest grossing product categories.

Order Count

Distinct order-level KPI to measure volume.

Monthly Revenue Trend

Revealed seasonality and peak periods in sales.

Top Customers (LTV)

Ranked customers by revenue to identify high-value segments.

Average Delivery Time

Assessed logistics efficiency and areas for improvement.

Made with GAMMA

Error Handling & Scheduling

Error Handling

Gold transformation wrapped in a try/except block for fault tolerance and traceability. This ensures pipeline robustness.

Scheduling

Pipeline notebooks orchestrated using Databricks Workflows, scheduled daily at 3 AM UTC. Email alerts configured for failures.

- 01 Bronze → 02 Silver → 03 Gold
- Daily at 3 AM UTC
- Email alerts on failure

Made with GAMMA

Conclusion & Future Scope

Scalable ELT Pipeline

Successfully built using Databricks, Spark, and Delta architecture.

Production-Grade Maturity

Automated scheduling and logging ensure reliability and traceability.

Future Extensions

Real-time ingestion, BI dashboards, customer segmentation, and demand forecasting.

Prepared by: MAHASWETHA AS - DATA ENGINEERING – DATABRICKS

Made with GAMMA