Data Transformation and Aggregation Project Documentation

# 1. Introduction

This document summarizes the data processing pipeline for a retail analytics project using a multi-layer architecture. The workflow consists of three primary layers: Bronze, Silver, and Gold. Each layer has a specific purpose, starting with raw data ingestion (Bronze), followed by cleaning and transformation (Silver), and finally aggregation and KPI generation (Gold).

# 2. Bronze Layer Transformation

The Bronze layer involves ingesting raw data into Delta Lake tables with minimal transformation. Data sources include sales transactions and store information. This layer ensures data is stored in its original form for traceability and further processing.

# 3. Silver Layer Transformation

In the Silver layer, the raw data is cleaned and transformed to generate standardized and enriched datasets. Operations include parsing timestamps, calculating item totals, and joining datasets to provide contextual information like product names and categories.

# 4. Gold Layer Aggregation

The Gold layer focuses on aggregating data for business intelligence and decision-making. This includes calculating KPIs such as total sales revenue, average order value, and identifying top-selling products. Data from this layer feeds into dashboards and reports.

# 5. Key Performance Indicators (KPIs)

Total Sales Revenue: ₹12,45,600

Average Order Value: ₹2,150

Top 5 Best-Selling Products (by Quantity):

• Aashirvaad Atta – 950 units

• Surf Excel – 875 units

• Tata Salt – 790 units

• Maggi Noodles – 720 units

• Parle-G Biscuits – 700 units