Software Requirements Specification (SRS)

1. Introduction

1.1 Purpose

This SRS defines the functional and architectural requirements for an event-driven ETL pipeline. The pipeline processes raw data from storage into a structured data warehouse through Bronze, Silver, and Gold layers using technologies such as Spark, Hive, Kafka, Pub/Sub, Oracle, and MySQL.

1.2 Scope

The ETL pipeline covers ingestion, transformation, enrichment, and aggregation of raw data. It enables a streamlined data flow for downstream analytical and business intelligence use cases.

2. Architecture Overview

Storage (RAW Data)

|

v

[ETL Job 1: Spark --> Hive Stage Table] ---> Bronze Layer

|

v

[ETL Job 2: Spark --> DW Tables] ---> Silver Layer (Filtering, Cleaning, Dedup, Explode)

|

v

[Publish Event to Pub/Sub]

|

v

[Consumer Microservice: Fetch from Oracle/MySQL, Push to Pub/Sub]

|

v

[ETL Job 3: Spark --> Intermediate DW Table]

|

v

[ETL Job 4: Aggregation Spark Job] ---> Gold Layer

3. Data Layers

3.1 Bronze Layer

- Source: Raw data from storage systems (e.g., cloud storage).

- Destination: Hive stage tables.

- Purpose: Preserve raw format for traceability.

3.2 Silver Layer

- Processes:

- Data filtering

- Null handling and type casting

- Deduplication

- Exploding nested structures

- Output: Cleaned, flat DW tables (with unique IDs)

3.3 Gold Layer

- Input: Silver layer tables and intermediate enriched data

- Process: Joins and aggregations

- Output: Analytical tables/KPIs for consumption

4. Functional Requirements

4.1 Ingestion Job (Bronze Layer)

- Input: Files from storage

- Process: Spark reads and writes to Hive staging

- Output: Raw data in Hive stage tables

4.2 Transformation Job (Silver Layer)

- Input: Hive stage tables

- Process: Cleaning, deduplication, and flattening

- Output: Structured DW tables with cleaned data

4.3 Event Trigger and Publication

- Trigger: New entries in Silver Layer

- Action: Publish events to Kafka or Pub/Sub

4.4 Consumer Microservice

- Trigger: Message from Pub/Sub

- Process: Fetch enriched data from Oracle/MySQL

- Output: Publish enriched data back to Pub/Sub

4.5 Intermediate Processing Job

- Input: Pub/Sub enriched messages

- Process: Spark job to parse and store in DW intermediate table

- Output: DW table with enriched information

4.6 Aggregation Job (Gold Layer)

- Input: Silver and Intermediate DW tables

- Process: Spark performs joins and aggregations

- Output: Gold Layer tables ready for reporting

5. Technologies

- Processing Engine: Apache Spark (Java, Maven)

- Storage & DW: Hive, Oracle, MySQL

- Messaging: Kafka, Google Pub/Sub

- Languages: Java for microservices and Spark jobs

End of Document