ETL Specification Document

Project Name: Sales Intelligence Platform

Document Version: 2.0

Prepared By: BI Engineering Team

Date: May 13, 2025

# Medallion Architecture Overview

Bronze: Raw ingestion from source systems  
Silver: Cleaned and validated data  
Gold: Aggregated KPIs for analytics

# Source Systems

CRM\_DB - PostgreSQL - customer\_info  
SALES\_DB - MySQL - sales\_records  
WEB\_TRACKING - CSV on S3 - user\_clicks

## Table: bronze\_customer\_info

customer\_id - INT - Unique ID  
name - STRING - Full name  
email - STRING - Email address  
signup\_date - DATE - Date of account creation

## Table: bronze\_sales\_records

sale\_id - INT - Sale transaction ID  
customer\_id - INT - Foreign key  
product - STRING - Product name  
sale\_amount - DECIMAL - Amount  
sale\_date - DATE - Transaction date

## Table: silver\_customers

customer\_id - INT - From bronze  
first\_name - STRING - SPLIT(name)[0]  
last\_name - STRING - SPLIT(name)[1]  
email\_clean - STRING - Lowercase(email)  
signup\_year - INT - EXTRACT\_YEAR(signup\_date)

## Table: gold\_sales\_summary

customer\_id - INT  
num\_sales - INT - COUNT(sale\_id)  
total\_revenue - DECIMAL - SUM(sale\_amount)  
avg\_revenue - DECIMAL - AVG(sale\_amount)

# Transformation Mapping Matrix

bronze\_customer\_info.name -> silver\_customers.first\_name (SPLIT[0])  
bronze\_customer\_info.name -> silver\_customers.last\_name (SPLIT[1])  
bronze\_sales\_records.sale\_amount -> gold\_sales\_summary.total\_revenue (SUM)

# Schedule and Frequency

Bronze: Hourly via Spark  
Silver: Hourly via Spark  
Gold: Daily at 2 AM via Airflow

# Data Quality Checks

|  |  |  |
| --- | --- | --- |
| Check Type | Target Table | Rule |
| Null check | silver\_customers | customer\_id IS NOT NULL |
| Email format | silver\_customers | email\_clean LIKE '%@%.%' |
| Duplicate detection | gold\_sales\_summary | customer\_id NOT IN (SELECT ...) |
| Positive amount | gold\_sales\_summary | total\_revenue > 0 |