# Data Engineering Requirements Document

## Project Name: Financial Snapshot for AIM 2.0

\*\*Prepared By\*\*: [Your Name]  
\*\*Date\*\*: [Insert Date]

## 1. Project Overview

The Financial Snapshot project aims to generate key financial metrics (KPIs) for \*\*Salesforce team\*\* consumption. The KPIs provide insights into account balances, loan utilization, transaction volumes, and product penetration. The final deliverables are targeted for delivery by \*\*Friday, Jan 17\*\*, requiring efficient ETL pipeline development, validation, and delivery.

## 2. Project Objectives

* Automate the extraction, transformation, and loading (ETL) of financial data.
* Compute the following KPIs:
* 1. Account & Balance Overview
* 2. Loan & Credit Utilization
* 3. Transaction Volume
* 4. Product Penetration
* Deliver KPIs to the Salesforce team in the required format (SQL table or CSV file).

## 3. Data Requirements

### 3.1 Input Data Sources

|  |  |  |  |
| --- | --- | --- | --- |
| Data Source | Table/File Name | Key Fields | Owner/Team |
| Account Balances | accounts | account\_id, account\_type, balance | Jess's Team |
| Loan Utilization | loans | customer\_id, credit\_limit, credit\_used | Jess's Team |
| Transactions | transactions | transaction\_id, customer\_id, amount, date | Jess's Team |
| Product Information | products | customer\_id, product\_id, product\_type | Jess's Team |

### 3.2 Output Requirements

\*\*Format\*\*: CSV files or SQL tables.  
\*\*Destination\*\*: Salesforce consumption layer.

|  |  |  |
| --- | --- | --- |
| KPI | Output Fields | Delivery Type |
| Account & Balance Overview | Total Balance, Average Balance, Contribution by Type | Table/File |
| Loan & Credit Utilization | Total Credit Used, Utilization %, Avg Utilization | Table/File |
| Transaction Volume | Total Volume, Total Count, Average Amount | Table/File |
| Product Penetration | Penetration %, Multi-Product Rate | Table/File |

## 4. Data Pipeline Requirements

### 4.1 ETL Pipeline Stages

1. Stage 1: Data Extraction - Extract data from the source tables and validate integrity.
2. Stage 2: Data Transformation - Compute KPIs using optimized queries and logic.
3. Stage 3: Data Loading - Deliver outputs to the target format.

### 4.2 Tools & Technologies

|  |  |
| --- | --- |
| Task | Tool/Technology |
| Data Extraction | SQL, PySpark, Hive |
| Data Transformation | PySpark, SQL |
| Data Validation | SQL, Pandas (Python) |
| Data Storage | HDFS, Cloud Storage |
| Automation/Orchestration | Apache Airflow / Cron Job |

## 5. Validation and Risks

### 5.1 Validation Steps

* Schema Validation: Verify required fields exist.
* Data Quality Checks: NULL values, duplicates.
* KPI Validation: Cross-check against mock data.
* End-to-End Testing: Ensure pipeline produces consistent outputs.

### 5.2 Risks and Mitigation

|  |  |  |
| --- | --- | --- |
| Risk | Impact | Mitigation |
| Source data delays or quality issues | Delivery delays | Early validation and stakeholder sync |
| Large data volumes causing performance lag | Pipeline failure | Optimize joins and transformations |
| Ambiguity in KPI definitions | Incorrect results | Confirm requirements with stakeholders |

## 6. Delivery Timeline

|  |  |  |
| --- | --- | --- |
| Task | Owner | Timeline |
| Requirement Gathering | Data Engineer | Day 1–2 |
| Data Extraction & Exploration | Data Engineer | Day 3–5 |
| Transformation Logic Development | Data Engineer | Day 6–10 |
| Pipeline Implementation | Data Engineer | Day 11–14 |
| Validation & Testing | Data Engineer | Day 15–18 |
| Final Delivery | Data Engineer | Day 19–21 |