Financial Services Capstone Project – Project Document

Dataset Link :[financial services datasets](https://datamasterconsultingin-my.sharepoint.com/:f:/g/personal/naval_thedatamaster_in/Emkb_q8oMSdFnRTSz-BzM-cBdWjmOVz5y4rX4pl7vYbHAQ?e=qqmgf4)

1. Project Introduction

Overview

This capstone project simulates a real-world Digital Banking Lakehouse implementation on Databricks. As a data engineer at "FinanceCorpDB Bank," you will modernize legacy infrastructure using Databricks’ unified analytics platform. You’ll tackle realistic financial datasets, build medallion architecture pipelines (Bronze → Silver → Gold), perform PII masking, SCD implementation, and deliver business KPIs via dashboards—preparing you for advanced data engineering roles.

Business Context

* Unify customer and product data across multiple banking channels
* Build real-time, regulatory-compliant analytics foundations
* Enable powerful self-service dashboards, KPI computation, and modern governance

Learning Objectives

* Medallion Architecture design and development
* Delta Live Tables (DLT) implementation
* PII masking and data privacy techniques
* SCD Type 2 dimension modeling
* Data quality monitoring and management
* Advanced Delta Lake optimization (OPTIMIZE, VACUUM, ZORDER)
* Unity Catalog for data governance
* Business metrics and executive dashboard creation
* Collaboration, source control, and operational best practices

2. Dataset Details

Provided Datasets

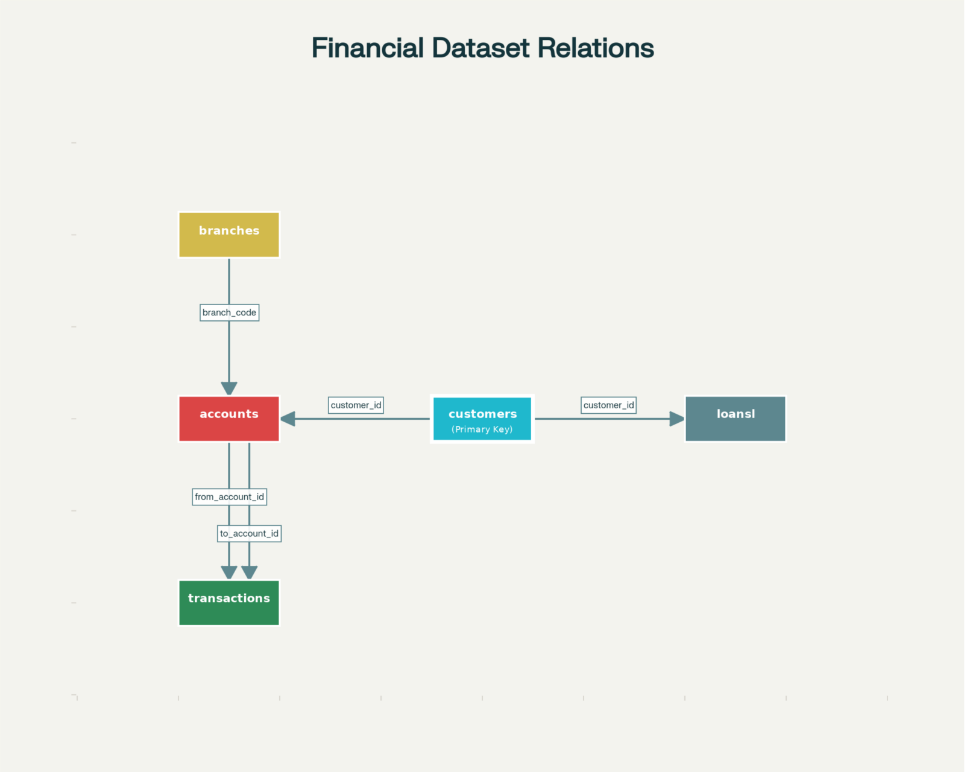
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Format | Rows | Description | Key Data Issues |
| customers.csv | CSV | 5,300 | Customer master, PII, KYC | PII, duplicates, nulls |
| accounts.json | JSON | 8,000 | Accounts w/ nested JSON | Nested, orphaned records |
| transactions.csv | CSV | 25,500 | Transactions by accounts | Duplicates, nulls, orphaned |
| loans.jsonl | JSONL | 2,000 | Loan portfolio, nested | Nulls, nested, orphaned |
| branches.json | JSON | 23 | Branch master, operations | Missing critical data |

Data Quality Issues Embedded

* Duplicates across all datasets (3-5%)
* Nulls in ~10-15% of selected columns
* Orphaned records (accounts, transactions, loans w/ invalid references)
* PII fields (emails, phones, PAN, Aadhar)
* Nested and multi-format JSON
* Missing/inconsistent categorical values
* Negative/invalid balances

Dataset Relationship Diagram

Financial Datasets Relationship Diagram - Shows how customer, account, transaction, loan, and branch data interconnect through primary and foreign keys



3. Solution Architecture

* Raw Layer: Databricks Volume (raw), holding all 5 data files
* Medallion Architecture:
* Bronze Layer: Raw ingestion into Delta tables, minimal transformation, schema evolution
* Silver Layer: Data cleansing, deduplication, masking, referential integrity, SCD implementation
* Gold Layer: KPI/business marts, customer 360, dashboard tables
* Governance: Unity Catalog for all schema/tables, row-level security, lineage, audit trail
* Orchestration: Databricks Jobs and Pipelines (scheduling, dependencies, notifications)
* Analytics: Databricks SQL Dashboards, Genie natural-language Q&A, Alerts

Technology Stack

Databricks, Delta Lake, Apache Spark, Python/SQL/Scala, Unity Catalog, DLT, GitHub, BI Dashboards

4. Project Execution Steps

Step 1: Setup & Data Ingestion

* Create Unity Catalog
* Catalog: capstone\_project
* Schemas: bronze, silver, gold
* Create Volume
* Volume: raw under capstone\_project.bronze
* Upload Datasets
* Download 5 data files from provided drive (or local copy)
* Upload to raw volume in workspace

Step 2: Bronze Layer (Raw Data Ingestion)

* Use Auto Loader/structured streaming for CSV/JSON/JSONL files
* Minimal transformation: schema inference, add file metadata
* Quarantine malformed records

Step 3: Silver Layer (Transformation & Quality)

* Data Cleansing
* Deduplicate rows (exact, fuzzy)
* Null handling: drop/impute/business rule
* Mask PII: hash or obfuscate sensitive fields
* Standardize categorical fields and dates
* Referential Integrity
* Valid customer references in accounts, loans, transactions
* SCD Type 2 Implementation
* Track historical changes in silver\_dim\_customers, possibly branches
* Use effective\_start, effective\_end, is\_current columns

Step 4: Gold Layer (Business KPIs)

* Aggregate for Analytics
* Customer 360: accounts, balances, loans, activity segment, risk category
* Transaction KPIs: volumes, failure rate, channel mix, avg/median stats
* Loan KPIs: default rates, outstanding amounts, interest rate analytics
* Business KPI Examples
* Active customers/accounts by month
* Monthly transaction volume and failure rate
* Wealth segment assignment
* Credit score distributions, risk evaluation
* Compliance and operational metrics
* Dashboard Tables: Pre-aggregate for fast BI query response

Step 5: Dashboard & Genie Analytics

* Databricks SQL Dashboards
* Executive Overview: KPIs, channel mix, failure trends
* Customer Analytics: Segments, geography, activity
* Loan Portfolio: Default rates, outstanding/principal stats
* Operations: Volume, branch, transaction status
* Genie Q&A: Natural-language query publication over trusted views
* Alerts: KPI alerts for business thresholds, pipeline notifications

Step 6: Source Control & Automation

* GitHub Repository
* Commit all code, notebooks, configs, and documentation
* Include data dictionary, setup instructions, troubleshooting
* Jobs & Pipelines
* Orchestrate DLT and dashboard refresh jobs
* Set up data quality monitoring scripts, health alerts

Step 7: Maintenance & Optimization

* Delta Lake Maintenance
* OPTIMIZE gold and frequent silver tables by key dimensions
* VACUUM old versions, monitor retention
* ZORDER by customer/account/date/transactional keys as appropriate
* Schedule auto-maintenance jobs for reliability

5. Project Outcomes

Technical Achievements

* End-to-end lakehouse implementation: 15+ Delta tables, DLT pipeline, SCD, PII masking
* 40k+ records processed and analyzed
* Automated data pipeline orchestration and quality monitoring

Business Impact

* Customer 360 analytics with actionable KPIs
* Full loan portfolio and transaction intelligence
* Compliance and operational quality dashboards

Governance & Collaboration

* Unity Catalog-enabled security and lineage
* Version control in GitHub
* Live Dashboards and Genie Q&A for self-service analytics

6. Deliverables Checklist

* Unity Catalog setup and documentation
* 5 raw datasets uploaded to volume
* Bronze, Silver, Gold layers created via DLT/Jobs
* Transformation/scripting notebooks (with markdown explanations)
* KPIs and dashboard tables in gold layer
* 4 Databricks SQL dashboards + Genie Q&A space
* Source GitHub repository with full project, data dictionary, architecture, troubleshooting
* Automated data quality and health alerts
* Final summary presentation or executive readout

7. Evaluation & Support

Rubric

* Technical implementation and code quality
* Data governance and compliance (PII, SCD, security)
* KPI development, dashboard insights, business value
* Automation, monitoring, optimization
* Documentation, collaboration, and presentation skills

Support

* Troubleshooting guide included in repo
* Data dictionary
* Sample queries, dashboard layouts, and diagram
* Suggested timeline: 1-2 weeks, ~15-20 hours