Bank 360: Intelligent Banking & Risk Assessment System

Project Documentation

Date: November 30, 2025 **Tech Stack:** Azure Data Factory, Azure Databricks, Azure SQL, Power BI, Azure Machine Learning **Team Members:**

- Abdallah Ramadan Abdallah Ali
- Ahmed Yehya Saad Nafea
- Hagar Mohamed Mabrouk Mohamed
- Merehan Ibraheem elmotasem Hassan
- Rania Ossama Hassan Abd Elhaleem
- Yousef Wael Omar AboDaif

---

1. Executive Summary

Traditional banking risk models often rely on static demographic data—such as income brackets and age—which fails to capture the dynamic financial behavior of a customer. "Bank 360" is an end-to-end Data Engineering and Machine Learning solution designed to bridge this gap.

By simulating a complete banking ecosystem with over 4 million transactions, we have built a data pipeline that ingests raw operational data, transforms it into actionable financial intelligence, and calculates a dynamic "Behavioral Risk Score." This score feeds into Machine Learning models (Random Forest and LightGBM) to automate loan approval decisions, ensuring the bank minimizes default risk while maximizing lending profitability.

---

2. Problem Statement

The modern lending landscape faces three critical challenges:

1. **Static Decisioning:** Loan officers rely on "Snapshot" data (e.g., a pay slip from last month) rather than "Video" data (continuous spending habits).
2. **Data Fragmentation:** Customer data is often siloed. Checking accounts, credit card debt, and loan histories exist in disparate systems, preventing a 360-degree view of the customer.
3. **Invisible Risk:** High-risk behaviors—such as consistent overdrafts, gambling, or "credit card cycling" (paying debt with debt)—are often invisible until a default occurs.

Our Solution: Bank 360 creates a unified data ecosystem where every transaction, payment, and account balance is correlated. This allows for the calculation of an internal risk score that updates in near real-time based on actual user behavior.

---

3. Technical Architecture

The project is built on a cloud-native **Medallion Architecture** using the Microsoft Azure ecosystem.

3.1 Tech Stack

- **Orchestration:** Azure Data Factory (ADF) handles the movement of raw data from on-premise sources to the cloud.
- **Storage (Data Lake):** Azure Data Lake Storage Gen2 (ADLS) stores data in Bronze (Raw), Silver (Cleaned/Enriched), and Gold (Aggregated) layers using Delta/Parquet formats.
- **Compute & Processing:** Azure Databricks (PySpark) is the core calculation engine, handling distributed processing of 4M+ rows.
- **Database:** Azure SQL Database serves the final "Gold" tables for high-speed reporting.
- **Machine Learning:** Azure Machine Learning Studio (Designer) trains models using the engineered features.
- **Visualization:** Power BI connects to Azure SQL to provide executive dashboards.

3.2 Data Pipeline Flow

1. **Ingestion:** Raw CSV files are ingested into the **Bronze Layer**.
2. **Transformation (Silver):** PySpark notebooks clean data, handle nulls, and perform complex logic (Balance Reconciliation, Amortization).
3. **Enrichment:** The Risk Engine calculates the Behavioral Credit Score.
4. **Aggregation (Gold):** Data is joined into a `Customer_360` view and a `Loan_Training_Set`.
5. **Intelligence:** The ML model consumes the Gold data to predict approval/default.

---

4. Data Ecosystem & Simulation

To stress-test the engineering pipeline, we generated a massive synthetic dataset that simulates a realistic banking economy.

4.1 Data Volume

- **Transactions:** 4,000,000+ rows.
- **Accounts:** ~5,000 active accounts.
- **Loans:** ~3,000 approved loans with full repayment history.

4.2 Simulation Logic

The data is **not random**; it follows strict business rules:

- **Wealth Scaling:** Customers with high income generate higher transaction volumes and larger loan requests.
- **Unemployed Logic:** Unemployed entities do not receive "Salary" deposits but receive irregular "Cash Deposits" to simulate survival, allowing the model to detect income instability.
- **Credit Card Cycles:** Spending increases debt (negative balance), while payments reduce debt. Logic ensures users rarely exceed their assigned credit limits.

---

5. Engineering Logic (The Silver Layer)

The core intelligence of the system resides in the Silver Layer transformations performed in Azure Databricks.

5.1 Balance Reconciliation

We do not trust static balance fields. Instead, we mathematically derive the "True Balance" of every account by aggregating the entire transaction history:

> *Current Balance = Sum(All Deposits) - Sum(All Withdrawals)*

5.2 Loan Amortization Engine

We implemented a custom Python UDF (User Defined Function) to simulate the lifecycle of every loan.

- **Principal vs. Interest:** Every payment is split based on the loan's interest rate and the remaining balance.
- **Status Logic:** Loans transition through statuses: *Active -> Late -> Defaulted (stops paying) -> Paid Off*.

5.3 The Risk Engine (Behavioral Scoring)

We developed a proprietary algorithm to calculate a credit score (300-850) based on five internal factors:

Factor	Weight	Logic Description
Payment Reliability	35%	Ratio of On-Time payments to Total payments. Defaults trigger an immediate score drop.
Utilization Rate	30%	Ratio of Current Debt to Credit Limit. Usage > 80% is heavily penalized.
Credit History	15%	Rewards customers with accounts older than 5 years.
Credit Mix	10%	Bonus points for managing both revolving credit (Cards) and installment credit (Loans).
New Credit	10%	Penalty for excessive loan applications in a short timeframe.

---

6. Machine Learning Implementation

The goal of the ML component is to automate the decision-making process using the engineered data.

6.1 Model Architecture

- **Platform:** Azure Machine Learning Studio.
- **Algorithms Tested:** Random Forest Classifier and LightGBM.
- **Target Variable:** `Approved` (Binary: 1 for Approve, 0 for Reject).

6.2 Feature Engineering

The model was trained on the following engineered features from the Gold Layer:

- **DTI (Debt-to-Income):** Total monthly debt obligations divided by monthly income.
- **Liquidity Ratio:** Total liquid assets divided by the requested loan amount.
- **Behavioral Score:** The internal score calculated in the Silver layer.
- **Income Stability:** Employment status encoded (Full-time vs. Unemployed).

6.3 Results

The LightGBM model demonstrated superior performance in identifying high-risk applicants who appeared "safe" on paper but had underlying behavioral risks (e.g., high DTI disguised by high income).

---

7. Conclusion

Bank 360 successfully demonstrates how modern Data Engineering can transform banking operations. By moving from simple data storage to complex **behavioral processing**, we created a system that:

1. **Scales:** Handles millions of transactions via distributed computing (Spark).
2. **Predicts:** Identifies risk before it becomes a loss.
3. **Automates:** Reduces manual underwriting effort for clear-cut cases.

This project serves as a foundational architecture for a Digital Bank, with future scope to include real-time fraud detection using Kafka streaming.