Payment Security – Smart Fraud Detection & Analysis

Final Project Documentation

Project Overview

This project implements a complete end-to-end fraud detection system for banking transactions, combining database design, ETL, data warehousing, cloud data engineering, machine learning, automated reporting, and a web application for visualization and fraud prediction.

The system is designed to:

- **Detect fraudulent transactions.**
- **Provide insights into customer and merchant risk.**
- **Automate data processing pipelines using cloud technologies.**
- **Deliver automated daily fraud intelligence to executive leadership.**

Milestone 1 – Data Collection & Preprocessing

Dataset:

- **Source: Kaggle "Bank Transaction Fraud Detection" dataset.**
- **Key Attributes: Customer_ID, State, City, Transaction_Date, Transaction_Amount, Transaction_Type, Merchant_ID, Device_Type, Customer_Email.**

Preprocessing Steps:

- **Handle missing or inconsistent data.**
- **Transform and format dates and transaction types.**
- **Remove duplicates.**
- **Ensure realistic distribution of fraudulent vs. non-fraudulent transactions.**

Outcome:

Cleaned dataset ready for ingestion into the database and ETL pipelines.

Milestone 2 – Database Design & Integration

Database: Payment_Fraud_DB on SQL Server.

Design Highlights:

- Fully normalized to 3NF.
- Enforces referential integrity and constraints.

Tables:

- Customers, Accounts, Transactions, Merchants, Merchant_Categories, Devices, Locations, Cities, States.

Triggers:

- trg_UpdateFraudStatus: Automatically updates Fraud_Trans and Is_Banned based on transactions.

Mock Data Generator:

- Python script generating realistic synthetic transactions.
- Uses account, merchant, device, and location history to produce realistic risk scores.
- Outputs CSV for ingestion into database.

Outcome:

Structured OLTP database for fraud detection with realistic transaction data.

Milestone 3 – Data Warehouse & ETL

Data Warehouse Design:

- Star Schema architecture with dimensions:
  Dim_States, Dim_Cities, Dim_Locations, Dim_Categories, Dim_Merchants, Dim_Customers, Dim_Accounts, Dim_Devices, Dim_Transaction, Dim_Date.

- Fact Table: Fact_Transactions including surrogate keys and measures like Account_Balance and Transaction_Amount.

ETL Development (SSIS):

- Extract, Transform, Load pipelines built in SSDT.
- Dimension tables loaded in parallel, followed by fact table.
- Data validation and incremental loading implemented.

Outcome:

A clean and structured DWH supporting fraud analytics queries and reports.

Milestone 4 – Automation with Airflow

Airflow DAG Implemented:

- ETL from OLTP Database → Data Warehouse
    - Runs daily.
    - Automates data movement from OLTP to DWH.
    - Tasks: Extract → Transform → Load.
    - Monitoring through Airflow UI with logs and success checks.

Outcome:

Daily automated ETL with monitoring and error logging.

Milestone 5 – Machine Learning

ML Model:

- Random Forest classifier trained locally on Kaggle dataset.
- Preprocessing included advanced feature engineering (time features, customer statistics, ratios, transaction flags).
- Class imbalance handled via weighted classes.

Deployment:

- Model deployed on Databricks.
- Predictions integrated into ETL pipeline in Silver layer of Medallion Architecture.
- Managed using MLflow for experiment tracking, model versioning, and DBFS/Azure Blob Storage deployment.

Evaluation:

- ROC AUC calculated for model performance.
- Top 10 important features identified.

Outcome:

Fraud detection model operational in cloud, feeding real-time predictions into ETL.

Milestone 6 – Cloud Data Engineering

Azure Resources:

- Resource Group for project management.
- Azure SQL Server hosting OLTP and DWH.
- Storage account with containers: bronze/, silver/, gold/.
- Databricks Workspace for ETL, ML, and Medallion workflow.

Databricks Pipeline:

- Bronze Layer: Raw transactions from Mock Generator.
- Silver Layer: Preprocessing + ML predictions.
- Gold Layer: Merges Bronze and Silver into enriched dataset.

Azure Data Factory (ADF):

- ETL from OLTP → DWH implemented.
- ETL from Gold → DWH partially planned (not completed due to subscription limitations).

Outcome:

Full cloud pipeline delivering curated, fraud-annotated data.

Milestone 7 – Data Analysis

Purpose:

- Generate actionable insights from historical and predicted fraud transactions.

Analysis Activities:

- Combine ML predictions with transaction and customer features.
- Profile customers and merchants based on transaction behaviors and risk scores.
- Identify anomalies, high-risk transactions, and trends over time.

Tools & Techniques:

- Power BI for interactive dashboards and data visualization.
- Aggregations, descriptive statistics, and reporting using Power BI's built-in features.

- **Reports include:**
    - **Fraud rate by day/week/month.**
    - **Customer & merchant risk ranking.**
    - **Transaction anomalies based on amount and frequency.**

Outcome:

Admins can explore historical data, monitor fraud trends, and assess risks effectively using Power BI dashboards and reports.

Milestone 8 – Web Application

Purpose:

- **Provide bank admins with interactive dashboards and on-demand fraud prediction capabilities.**

Features:

- **Dashboard & Visualization: Interactive charts for fraud trends, customer & merchant risk, transaction anomalies.**
- **Fraud Prediction Tool: Admin uploads new transaction(s) → receives predicted fraud labels and risk factors.**
- **Security: Role-based access for authorized admins only.**

Outcome:

Admins can visualize historical data, detect anomalies, and run real-time fraud predictions on new transactions.

Milestone 9 – Fraud Intelligence Automation

This milestone introduces an automated reporting module that generates a daily fraud briefing for executive stakeholders.

Overview

A fully automated process that refreshes fraud data every morning, processes key fraud indicators, and delivers a standardized PDF report without any manual intervention.

Key Features

- Daily automated fraud summary at 05:15 AM.
- Standardized PDF report using a banking-style template.
- Automated email delivery to Fraud & Risk units.
- Zero manual effort required.

Business Value

- Eliminates manual reporting tasks.
- Ensures consistent and reliable daily insights for decision-makers.
- Strengthens governance and improves visibility into suspicious activity.

System Architecture Summary

1. Data Collection & Preprocessing (On-Prem)
    - Local database (SQL Server) stores historical transactions, customers, accounts, merchants, devices, and locations.
    - Preprocessing on-prem handled missing data, duplicates, and formatting for dates and transaction types.
2. Database Design & Integration (On-Prem)
    - Local SQL Server database: normalized tables enforcing referential integrity.
    - Mock Data Generator runs on-prem to create synthetic transactions for testing and enrichment.
3. Data Warehouse & ETL (On-Prem)
    - ETL pipelines built in SSIS: extract, transform, and load dimension tables (in parallel) and fact tables.
    - Data validation and incremental loading ensured integrity.
4. Automation with Airflow (On-Prem)
    - Airflow DAGs automate ETL from local DB → DWH.
    - Provides scheduling, monitoring, and logging of workflow tasks.

5. Machine Learning (Cloud)

- **Fraud detection model trained locally on Kaggle dataset.**
- **Deployed to Databricks for cloud inference.**
- **Integrated into ETL pipeline to produce predicted labels in Silver layer of Medallion Architecture.**
- **Managed with MLflow for experiment tracking and model versioning.**

6. Cloud Data Engineering (Cloud)

- **Azure SQL Server: hosts OLTP + DWH.**
- **Databricks Pipeline:**
  - **Bronze Layer: raw transactions (Mock Generator output).**
  - **Silver Layer: preprocessing + ML predictions.**
  - **Gold Layer: enriched dataset (merge of Bronze & Silver).**
- **Azure Data Factory (ADF): ETL from OLTP → DWH; partial plan for Gold → DWH load.**

7. Data Analysis

- **Power BI dashboards provide historical and predicted fraud insights.**
- **Profiles customers and merchants based on transaction behaviors and risk scores.**
- **Detects anomalies, high-risk transactions, and trends over time.**

8. Web Application (Admin Dashboard)

- **Interactive visualization of fraud trends, customer/merchant risk, and transaction anomalies.**
- **Allows on-demand fraud prediction for new transactions.**
- **Connected to Gold dataset in Databricks.**
- **Role-based access for authorized admins.**

Completed Components

- Local database setup, backup, and preprocessing.
- Database design and integration with normalized tables, constraints, and Mock Data Generator.
- ETL pipelines via SSIS for dimension and fact tables, with validation and incremental loading.
- Airflow on-prem to automate ETL from local DB → DWH.
- Machine learning model training locally, deployment on Databricks, and MLflow integration.
- Full Databricks pipeline (Bronze → Silver → Gold) for preprocessing and ML predictions.
- ADF pipelines for cloud ETL from OLTP → DWH.
- Data analysis dashboards and reports in Power BI.
- Admin web application for visualization and on-demand fraud prediction.
- Job scheduling for mock data generation and automated pipeline execution.
- Daily Fraud Intelligence Automation pipeline (Power BI refresh → DAX processing → PDF → Email delivery).

Conclusion

The project demonstrates a complete end-to-end cloud data engineering and fraud detection ecosystem integrating:

- Local and cloud databases.
- ETL and data warehouse pipelines.
- Machine learning models for fraud prediction.
- Automated fraud intelligence reporting.
- Data analysis dashboards and a web-based admin platform.

Despite minor limitations (partial Gold → DWH load in ADF), the system successfully ingests, transforms, predicts, analyzes, and delivers high-quality, fraud-annotated insights ready for banking operations.