FraudSight: End-to-End Fraud Detection & Risk Analytics Platform

Author: Kamran Habib

Tools: Azure Databricks (PySpark + Delta Lake) · Azure Data Factory · SQL · Power BI (DAX + Interactive Dashboards)

Focus Areas: Data Engineering · Fraud Analytics · Risk Modeling · ETL Pipeline Orchestration · Business Intelligence

---

1. Executive Summary

This project presents the design and implementation of a comprehensive **Financial Fraud Intelligence Platform** using Azure Databricks, Delta Lake Medallion Architecture, and Power BI. The objective of this platform is to detect fraudulent activities across transactions, merchants, accounts, and device behavior by engineering an end-to-end data pipeline and developing interactive intelligence dashboards.

The solution integrates multiple datasets, performs structured cleansing and transformation, and delivers insightful fraud metrics that support risk analysts in identifying fraud patterns, merchant anomalies, and account-level behavioral threats.

---

2. Business Problem

Financial fraud continues to escalate due to increasing digital transactions, diverse merchant ecosystems, and rising account takeover attempts. Organizations require reliable, scalable systems that consolidate raw operational data, transform it into an analytics-ready format, and provide risk teams with actionable insights.

This project aims to solve the following challenges:

- Detect fraudulent transactions efficiently
- Identify high-risk merchants and MCC categories
- Analyze abnormal account behavior
- Highlight device/IP-based anomalies
- Deliver interactive dashboards for investigative use

---

3. Data Architecture & Pipeline Design

3.1 Medallion Architecture (Bronze → Silver → Gold)

The data pipeline followed the Databricks Medallion Model:

Bronze Layer – Raw Data Ingestion

- Raw datasets ingested as-is into Delta tables
- No transformations applied
- Ensured schema preservation for auditability
- Stored in Azure Data Lake

Silver Layer – Cleansing & Standardization

Transformations included:

- Standardizing column names
- Fixing datatypes (dates, numbers, booleans)
- Handling null values
- Joining FactTransactions with FraudLabels
- Deriving new fields: transaction date, month, weekday, timestamps

- Preparing dimension tables (Merchant, Account, MCC, Date)

Gold Layer – Analytics-Ready Modeling

Created Gold-layer analytical models:

- **FactTransactionsGold** – enriched transaction table with fraud indicators
- **FactDailyAccount** – daily account-level transaction summaries
- **MerchantRiskGold** – merchant fraud metrics, MCC distributions
- **FactLoginRisk** – device, IP, and login behavior indicators
- **DimAccount, DimMerchant, DateDimension** – cleaned dimensions

These tables powered Power BI dashboards.

---

4. Data Engineering in Databricks

4.1 SQL Transformations

SQL was used for:

- Aggregations (daily spend, daily transactions)
- Fraud scoring rules
- Merchant category fraud calculations
- Chargeback and fraud breakdowns
- Joining facts and dimensions
- Creating analytical summaries used in dashboards

4.2 PySpark Transformations

PySpark performed:

- Large-scale joins
- Window functions for moving averages
- Feature engineering (velocity, high-amount flags, time-of-day analysis)
- Conversion into optimized Delta Lake tables
- Performance tuning with partitioning and ZORDER

---

5. Power BI Semantic Model

Gold-layer tables were loaded into Power BI to build a clean semantic model following Star Schema principles.

Facts:

- FactTransactions
- FactLoginRisk
- DailyAccountAggregates

Dimensions:

- DimAccount
- DimMerchant
- DimMCC
- Date

Power Query Transformations:

- Ensured correct data types
- Replaced null values
- Renamed columns consistently
- Created a dedicated Date dimension
- Organized the model into folders

Measure Layer (DAX):

Over 40 DAX measures created, including:

- Total Transactions
- Fraud Transactions
- Fraud Rate %
- Chargeback Transactions
- MCC Fraud Rate
- Fraud 7-Day Moving Average
- Failed Logins, MFA Failures, Device/IP counts
- High Amount / High Velocity / Night-time fraud indicators

All measures organized into clean subfolders.

---

6. Dashboard Summaries

6.1 Fraud Overview Dashboard

Purpose: Provide executives with a real-time summary of fraud performance.

Key Insights Delivered:

- Total Transactions, Total Amount, Fraud Transactions
- Fraud Rate trend over time
- Fraud per 1,000 transactions
- Fraud vs Transaction timeline comparison
- Countrywise fraud distribution
- Risk indicators: International Fraud, High Amount Fraud, High Velocity, Night Fraud, New Device Fraud

(See Appendix A.1 for dashboard screenshot)

6.2 Merchant Risk Analysis Dashboard

Purpose: Identify merchant-related fraud anomalies.

Visuals Included:

- Merchant Fraud Trend
- MCC Category Fraud Rate
- Total Merchant Transactions & Fraud Count
- Merchant Chargeback Rate
- Slicers for MCC, Merchant, Country

Business Impact:
Reveals high-risk merchants and MCC segments with recurring fraud patterns.

(See Appendix A.2 for dashboard screenshot)

6.3 Account Behaviour Analysis Dashboard

Purpose: Understand user behavior patterns that may indicate compromise.

Metrics Displayed:

- Daily Spend vs Daily Transaction
- Fraud 7-Day Moving Average
- Account summary table
- Slicers for account exploration

(See Appendix A.3 for dashboard screenshot)

6.4 Account Risk Summary Dashboard

Purpose: Surface device and login anomalies.

KPIs Tracked:

- Unique Devices
- Unique IPs
- Failed Logins
- Failed MFA
- Rooted Device Attempts
- IP Reputation

(See Appendix A.4 for dashboard screenshot)

---

7. Insights & Business Value

This platform enables:

Fraud Detection

- Identifies unusual spikes in fraud activity
- Detects countries, merchants, and accounts with abnormal behavior

Merchant Risk Monitoring

- Highlights MCC fraud concentration
- Flags repeatedly fraudulent merchant segments

Account Compromise Detection

- Tracks abnormal login frequency
- Detects risky IP behavior
- Unusual device usage patterns

Operational Efficiency

- Replaces manual fraud analysis with automated dashboards
- Provides consistent and repeatable fraud metrics

---

8. Conclusion

The Fraud Intelligence Platform demonstrates how modern data engineering techniques, combined with Power BI analytics, can transform raw financial data into actionable fraud insights. The system follows industry best practices and showcases capability across:

- Data Engineering
- BI Dashboarding
- Fraud Analytics
- Azure Databricks Pipeline Design
- DAX Measure Development
- Dimensional Modeling

This project is production-ready and reflects the type of fraud analysis ecosystems used in financial institutions today.

---

9. Future Enhancements

- Real-time streaming fraud detection
- ML-based fraud prediction (logistic regression, gradient boosting)
- Graph-based fraud ring detection
- Automated alerting integration
- Integration with case management systems

---

10. Appendix A — Power BI Dashboard Screenshots

A.1 Fraud Overview Dashboard

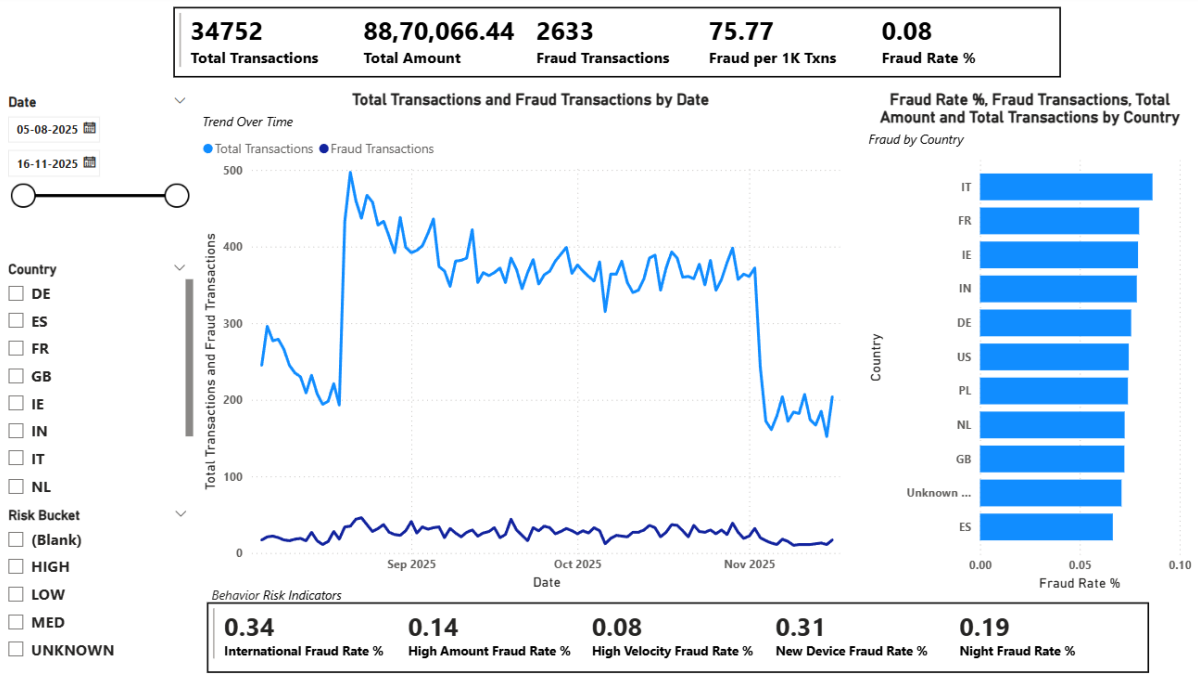


Figure A.1 — Fraud Overview dashboard summarizing core fraud KPIs.

A.2 Merchant Risk Analysis Dashboard

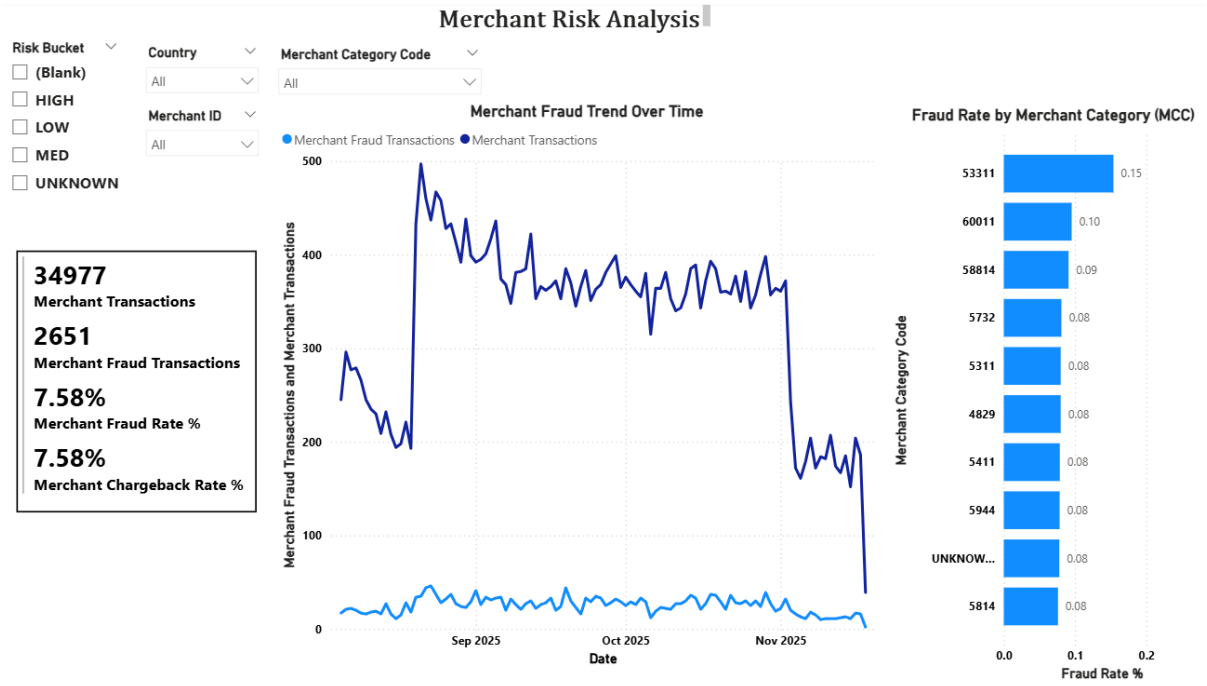


Figure A.2 — Merchant-level fraud risk and MCC category analysis.

A.3 Account Behaviour Analysis Dashboard

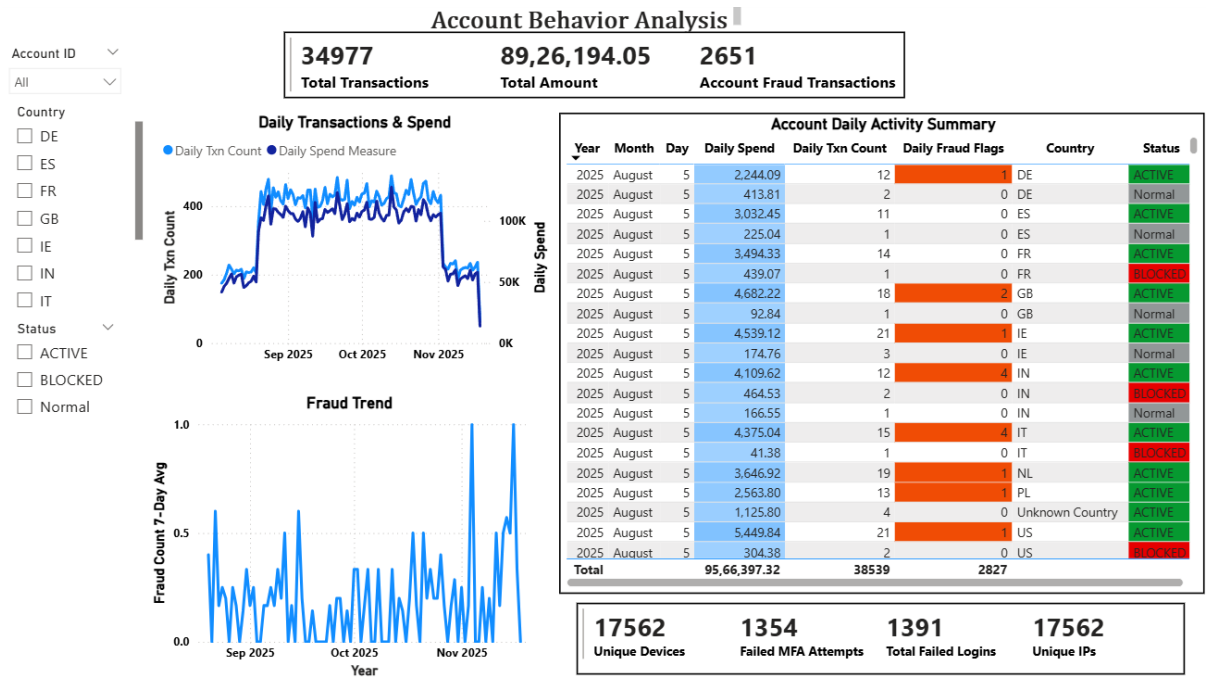
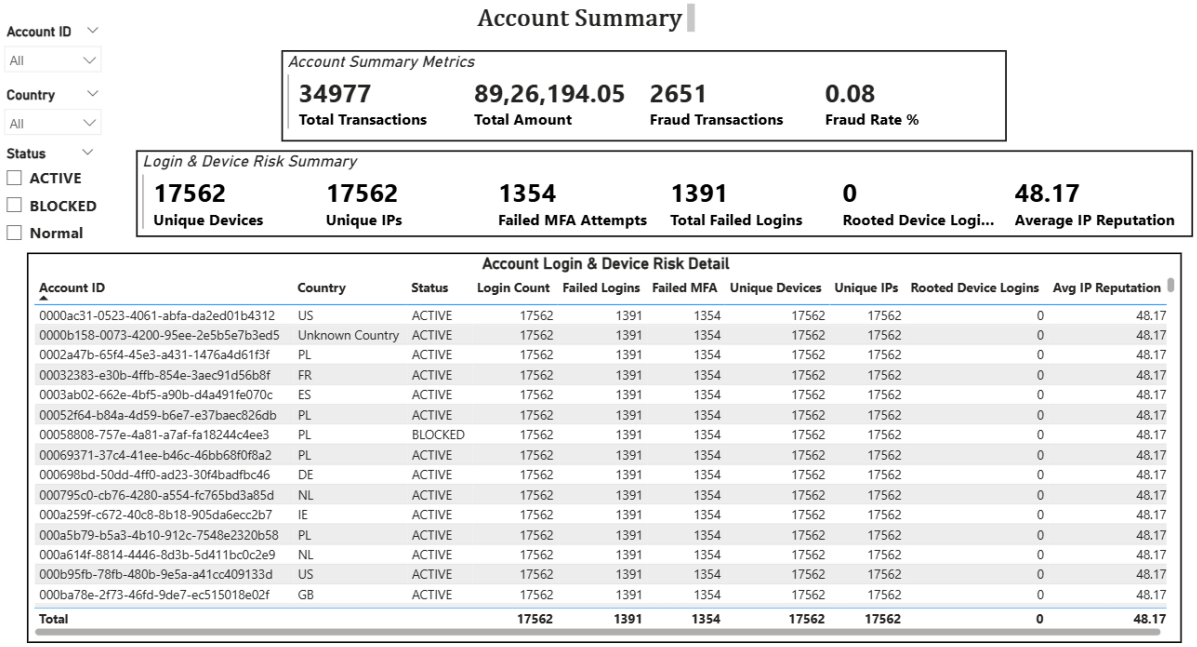


Figure A.3 — Account spending patterns, daily activity, and fraud moving average.

A.4 Account Risk Summary Dashboard



Account Login & Device Risk Detail

Account ID	Country	Status	Login Count	Failed Logins	Failed MFA	Unique Devices	Unique IPs	Rooted Device Logins	Avg IP Reputation
0000ac31-0523-4061-abfa-da2ed01b4312	US	ACTIVE	17562	1391	1354	17562	17562	0	48.17
0000b158-0073-4200-95ee-2e5b5e7b3ed5	Unknown Country	ACTIVE	17562	1391	1354	17562	17562	0	48.17
0002a47b-65f4-45e3-a431-1476a4d61f3f	PL	ACTIVE	17562	1391	1354	17562	17562	0	48.17
00032383-e30b-4ffb-854e-3aec91d56b8f	FR	ACTIVE	17562	1391	1354	17562	17562	0	48.17
0003ab02-662e-4bf5-a90b-d4a491fe070c	ES	ACTIVE	17562	1391	1354	17562	17562	0	48.17
00052f64-b84a-4d59-b6e7-e37baec826db	PL	ACTIVE	17562	1391	1354	17562	17562	0	48.17
00058808-757e-4a81-a7af-fa18244c4ee3	PL	BLOCKED	17562	1391	1354	17562	17562	0	48.17
00069371-37c4-41ee-b46c-46bb68f0f8a2	PL	ACTIVE	17562	1391	1354	17562	17562	0	48.17
000698bd-50dd-4ff0-ad23-30f4badfbc46	DE	ACTIVE	17562	1391	1354	17562	17562	0	48.17
000795c0-cb76-4280-a554-fc765bd3a85d	NL	ACTIVE	17562	1391	1354	17562	17562	0	48.17
000a259f-c672-40c8-8b18-905da6ecc2b7	IE	ACTIVE	17562	1391	1354	17562	17562	0	48.17
000a5b79-b5a3-4b10-912c-7548e2320b58	PL	ACTIVE	17562	1391	1354	17562	17562	0	48.17
000a614f-8814-4446-8d3b-5d411bc0c2e9	NL	ACTIVE	17562	1391	1354	17562	17562	0	48.17
000b95fb-78fb-480b-9e5a-a41cc409133d	US	ACTIVE	17562	1391	1354	17562	17562	0	48.17
000ba78e-2f73-46fd-9de7-ec515018e02f	GB	ACTIVE	17562	1391	1354	17562	17562	0	48.17
Total			**17562**	**1391**	**1354**	**17562**	**17562**	**0**	**48.17**

Figure A.4 — Device/IP risk indicators and login behaviour analysis.