

•Project Title:

“**End-to-End Credit Card Fraud Analytics: From Synthetic Data Engineering to Power BI Insights**”

•Subtitle: *A Data Engineering + BI Project*

• Name: **Tanmay Sharma**

•Role: Data/Business Analyst

•Date: 21/08/2025

Problem Statement

- Credit card fraud is a **multi-billion dollar problem** globally
- Fraudsters exploit **merchants, locations, peak hours**
- Businesses need **real-time fraud detection & insights**

: Project Workflow

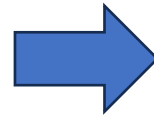
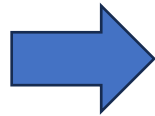
(→ Python → SQL → Power BI)

1.Synthetic Data Engineering (Python, Faker, Random)

2.Data Storage & Queries (MySQL / PostgreSQL)

3.ETL & Modelling (Cleaning, fraud labelling)

4.Visualization (Power BI dashboard)



Dataset Overview

- **200K+ transactions** (synthetic)
- **Fraud Rate: ~2%**
- Columns: customer_id, merchant, payment_method, txn_datetime, location, is_fraud
- Fraud Scenarios:
 - Rapid multiple txns
 - Impossible travel
 - High-risk merchants

Script.py

```
1 import pandas as pd
2 import random
3 from faker import Faker
4 from datetime import datetime, timedelta
5
6 fake = Faker()
7
8 # Config
9 ROWS = 200_000 # total transactions to generate
10 FRAUD_RATIO = 0.02 # 2% fraud
11 customers = [f"CUST{str(i).zfill(5)}" for i in range(1, 5001)]
12 merchants = ["Amazon", "Walmart", "Uber", "Netflix", "Gucci", "Apple Store", "Starbucks", "Shell Gas"]
13
14 data = []
15
16 for i in range(ROWS):
17     customer_id = random.choice(customers)
18     txn_time = fake.date_time_between(start_date="-1y", end_date="now")
19     merchant = random.choice(merchants)
20     location = fake.city()
21     payment_method = random.choice(["CreditCard", "DebitCard", "UPI", "Wallet"])
22
23     # Default (normal transaction)
24     amount = round(random.uniform(5, 500), 2)
25     is_fraud = 0
26
27     # Inject fraud patterns (2% transactions)
28     if random.random() < FRAUD_RATIO:
29         is_fraud = 1
30         fraud_type = random.choice(["high_amount", "rapid_fire", "impossible_travel", "odd_merchant", "time_based"])
31
32         if fraud_type == "high_amount":
33             amount = round(random.uniform(5000, 20000), 2)
34
35         elif fraud_type == "rapid_fire":
36             txn_time = datetime.now()
37             amount = round(random.uniform(100, 500), 2)
38
39         elif fraud_type == "impossible_travel":
40             location = random.choice(["New York", "London", "Tokyo", "Sydney"])
41             amount = round(random.uniform(50, 2000), 2)
42
43         elif fraud_type == "odd_merchant":
44             merchant = random.choice(["Luxury Watches", "Jewelry Store", "Private Jet Rentals"])
45             amount = round(random.uniform(2000, 15000), 2)
46
47         elif fraud_type == "time_based":
48             txn_time = txn_time.replace(hour=random.choice([1, 2, 3, 4]))
49             amount = round(random.uniform(500, 3000), 2)
50
51     data.append([i+1, customer_id, txn_time, amount, merchant, location, payment_method, is_fraud])
52
53 # Create DataFrame
54 df = pd.DataFrame(data, columns=[
55     "txn_id", "customer_id", "txn_datetime", "amount", "merchant", "location", "payment_method", "is_fraud"
56 ])
57
58 df.to_csv("synthetic_creditcard_fraud.csv", index=False)
59 print("✅ Synthetic fraud dataset generated successfully!")
60
```

transactions											
	A	B	C	D	E	F	G	H	I	J	K
1	transactions									FraudAnalysis 4h ago	
2	txn_id	customer...	txn_dateti...	amount	merchant	location	payment_...	is_fraud		Metric	Value
3	TXN001	CUST001	/2023 2:30 PM	1250	Amazon	New York	Credit Card	0		Total Transacti	10
4	TXN002	CUST002	/2023 3:45 PM	8750	Best Buy	Los Angeles	Debit Card	1		Fraud Transact	5
5	TXN003	CUST003	/2023 9:15 AM	2500	Walmart	Chicago	Credit Card	0		Fraud Percent	50.00%
6	TXN004	CUST004	/2023 6:20 PM	6500	Target	Houston	PayPal	1			
7	TXN005	CUST005	2023 11:30 AM	4200	Amazon	Phoenix	Credit Card	0		FraudDashboard 4h ago	
8	TXN006	CUST006	/2023 1:45 PM	9800	Best Buy	Philadelphia	Debit Card	1		Metric	Value
9	TXN007	CUST007	/2023 4:10 PM	3100	Walmart	San Antonio	PayPal	0		Total Transacti	10
0	TXN008	CUST008	/2023 8:30 PM	7250	Target	San Diego	Credit Card	1		Fraud Transact	5
1	TXN009	CUST009	/2023 8:45 AM	5300	Amazon	Dallas	Debit Card	0		Fraud Percent	50.00%
2	TXN010	CUST010	2023 12:15 PM	8900	Best Buy	San Jose	PayPal	1			
3											

Data in Quadratic AI For Better Understanding

Queries in MySQL:

-- Total Transactions, Fraud Transactions, Fraud %

```
SELECT
    COUNT(*) AS total_txns,
    SUM(is_fraud) AS fraud_txns,
    ROUND(SUM(is_fraud) * 100.0 / COUNT(*), 2) AS fraud_pct
FROM transactions;
```

-- Fraud by Payment Method

```
SELECT payment_method,
    SUM(is_fraud) AS fraud_txns,
    ROUND(SUM(is_fraud) * 100.0 / COUNT(*), 2) AS fraud_pct
FROM transactions
group by payment_method
Order by fraud_pct DESC;
```

-- Top Fraudulent Merchants

```
SELECT Merchant,
    SUM(is_fraud) AS fraud_txns,
    ROUND(SUM(is_fraud) * 100.0 / COUNT(*), 2) AS fraud_pct
FROM transactions
group by Merchant
Order by fraud_pct DESC;
```

-- Fraud by Hour of Day (Time-Based Fraud)

```
SELECT Hour(txn_datetime) as fraud_hour_of_Day,
    SUM(is_fraud) AS fraud_txns,
    ROUND(SUM(is_fraud) * 100.0 / COUNT(*), 2) AS fraud_pct
FROM transactions
group by Hour(txn_datetime)
Order by fraud_pct DESC;
```

-- High Amount Transactions (> \$5000) - Fraud Check

```
SELECT
    txn_id,
    customer_id,
    txn_datetime,
    amount,
    merchant,
    location,
    payment_method,
    is_fraud
FROM transactions
WHERE amount > 5000
ORDER BY amount DESC
LIMIT 10;
```



```
-- Month Over Month fraud_growth

WITH monthly_status AS (
    SELECT
        YEAR(txn_datetime) AS year_growth_fraud,
        MONTH(txn_datetime) AS month_growth_fraud,
        COUNT(*) AS total_txns,
        SUM(is_fraud) AS fraud_txns,
        ROUND(SUM(is_fraud) * 100.0 / COUNT(*), 2) AS fraud_pct
    FROM transactions
    GROUP BY YEAR(txn_datetime), MONTH(txn_datetime)
)

SELECT
    year_growth_fraud,
    month_growth_fraud,
    total_txns,
    fraud_txns,
    fraud_pct,
    LAG(fraud_txns) OVER (ORDER BY year_growth_fraud, month_growth_fraud) AS prev_month_fraud,
    ROUND(
        ( (fraud_txns - LAG(fraud_txns) OVER (ORDER BY year_growth_fraud, month_growth_fraud))
          / NULLIF(LAG(fraud_txns) OVER (ORDER BY year_growth_fraud, month_growth_fraud), 0) ) * 100,
        2
    ) AS mom_fraud_growth_pct
FROM monthly_status
ORDER BY year_growth_fraud, month_growth_fraud;
```



Key KPIs Tracked

- Total Transactions
- Fraud Transactions & Rate
- Fraud Amount
- Fraud by Merchant / Location / Hour
- Fraud by Payment Method

Power BI Dashboard

- 👉 Highlights:
- KPIs from MySQL integrated to Power BI
 - Peak fraud hours visualized
 - Top risky merchants
 - Hotspot locations



Insights & Findings

- Peak Hours:** Fraud spikes during **11:00–13:00 (morning to noon window)**
- Top Merchants:** Jewellery Stores & Luxury Watches
- Hotspot Locations:** New York, Tokyo, Sydney
- Payment Method Risk:** UPI & Wallet slightly higher
- SQL + Power BI synergy:** MySQL detected fraud patterns, Power BI visualized them

Conclusion

- Built an **end-to-end fraud detection pipeline**
- **MySQL**: handled data storage & fraud detection queries
- **Power BI**: delivered actionable insights with visuals
- Future Scope:
 - Real-time fraud alerts from SQL triggers
 - ML-based anomaly detection

Thank You

LinkedIn: <https://www.linkedin.com/in/tanmay-sharma-800599373/>

Git Hub: https://github.com/Tanu272004/Fintech_FraudDetection_Dataset.git

