

Lab

Course Code: CSE 4460

Course Name: Big Data Analytics

# Problem Statement

The main goal is to identify fraudulent transactions between bank customers using only data visualization.

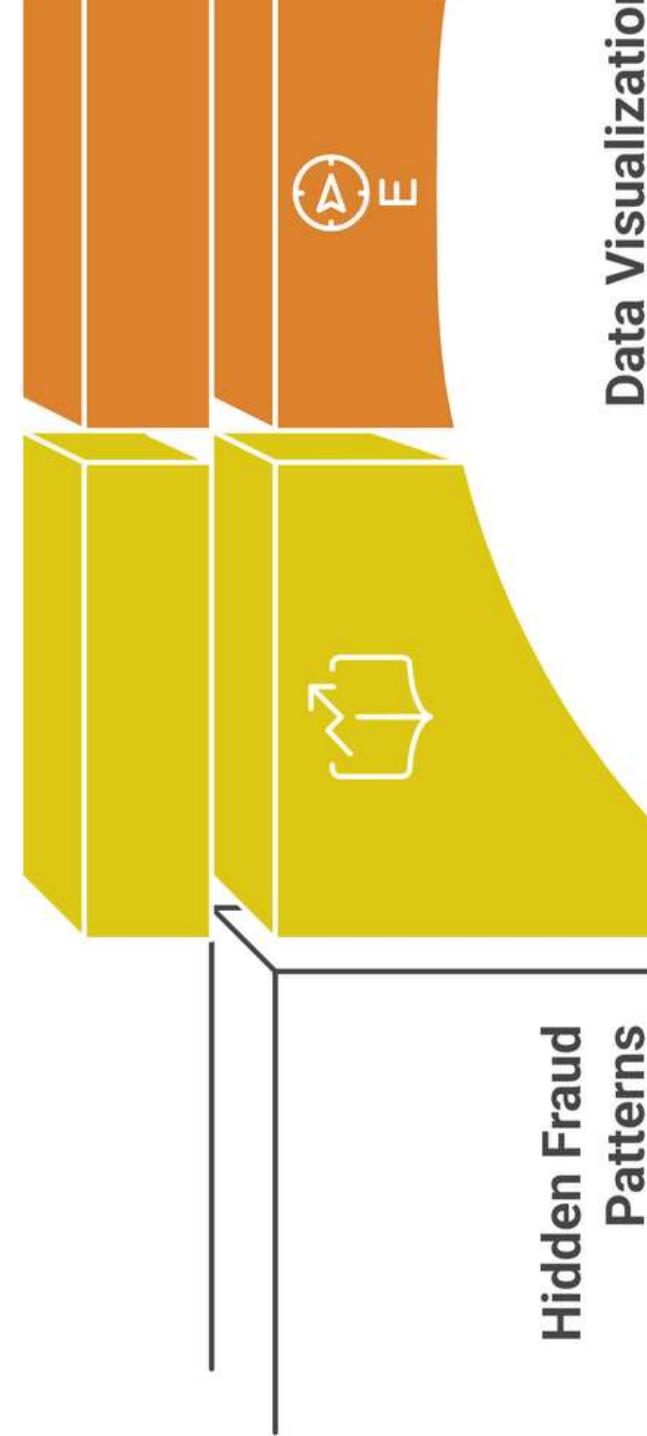
# Using Visualization to See, Understand Transaction

## Explainability

Visualize fraud for clear understanding

## Exploration

Discover new fraud techniques



Data Visualization

# Conceptual Idea to Detect I



## Identify Small Transactions

Transactions smaller than average are flagged

Analyze Transaction Patterns

# The "Under the Radar" Approach



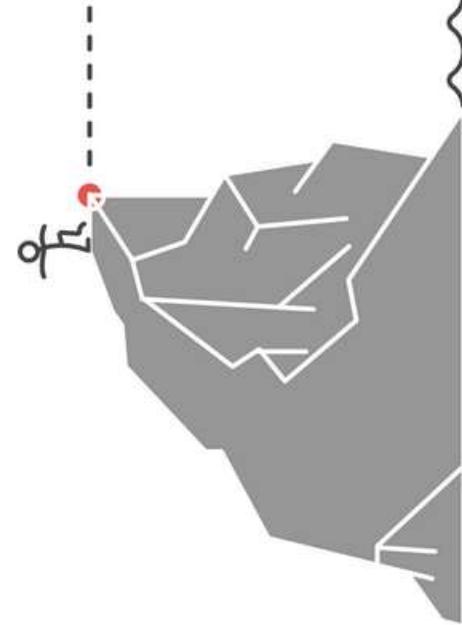
The core idea is that criminals often try to stay hidden by making their fraudulent transactions look insignificant. Large, unusual transactions (like buying a car in another country) are easy for banks' automated systems to flag as suspicious. To avoid this, criminals use a simple tactic: they make many small

# The Camouflage

A "money mule" is a person who, knowingly or unknowingly, lets criminals use their legitimate bank account to move stolen money. The goal is to make the money harder to trace back to the original crime.

To look innocent, a mule account will often maintain a history of perfectly normal, everyday transactions. This creates a "behavioral baseline". The account might receive a regular salary, pay monthly bills, and have small, predictable expenses for groceries or coffee. It looks just like anyone else's account.

Normal Account Activity	Unusual Activity
Predictable, consistent financial behavior	Large, unpredictable, or unusual financial activity



The anomaly, or the red flag for fraud, is a

# The "Cash-Out"

This rule focuses on a specific behavior: how criminals extract stolen money from the financial system and turn it into untraceable cash. After collecting money into a mule account, they need to withdraw it. To avoid suspicion, they don't make one large withdrawal. Instead, they make numerous small withdrawals or individual transfers over a short period.



# Real Stories

একটি ব্যাংক  
ব্যক্তি তার  
চিল স্বাভাৱিক  
টাকার এক  
ঢেউ গড় মাত্ৰ  
লেনদেন ক  
সমেহজনক

The amount of fraudulent transactions is smaller than the average amount of transaction by all users.

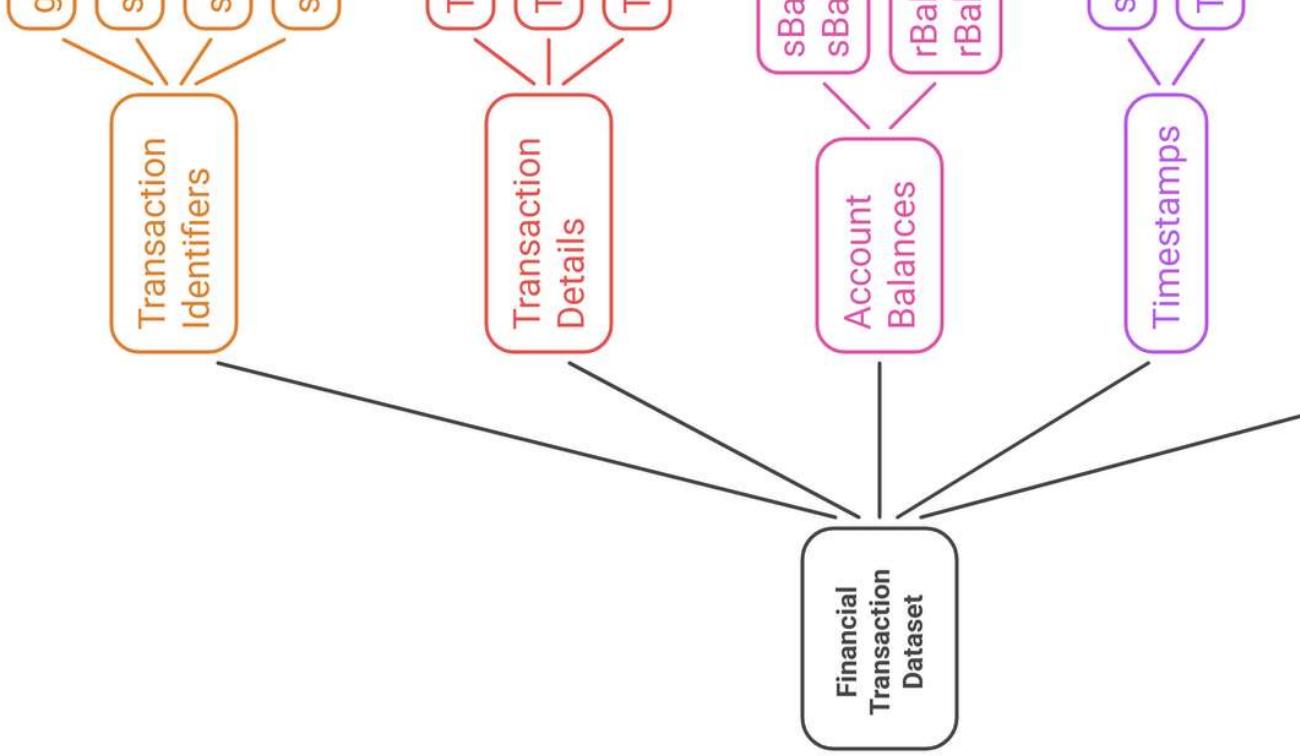
'অসিফ'-এর  
সন্ধানে বাজ  
ব্যালেন্স সব  
একদিন, তাৰ  
এৰপৰ, এই  
নম্বৰে ২০টি  
আচৰণের চ  
যাওয়া, এটাৰ

The mules can perform legitimate transactions but a sudden change in transferred money amounts corresponds to an anomaly.

একটি ব্যাংকে  
যাক, যেখানে  
কৰলে ব্যাংক  
কৌশল অবলু  
তিল বিকাশ ব  
পরিমাণ ৮,০০  
'ক্যাপ-আউট'

An account holder that did several transactions for individual or withdrawal purpose with an amount lower than average amount of transaction can be considered as a fraudster and

# Dataset



# Basic Characteristics

This is a critical characteristic for data cleaning and imputation:

- 54222 non-null : The majority of your columns are **complete**, meaning they have no missing values for any of the 54,222 records. This is great for analysis as you won't need to impute or drop rows based on these columns. These columns include: `gT`, `sID`, `rID`, `sAcc`, `rAcc`, `TranAmount`, `TranType`, `TransStatus`, `sBalBefore`, `sBalAfter`, `rBalBefore`, `rBalAfter`, `sf1`, `sf2`, `STD`, `RTD`, `NoDescription`, `TransID`, `sType`, `rType`.
- 0 non-null : Four columns, specifically `sf3`, `sf4`, `ef1`, and `ef2`, have 0 non-null entries. This means all 54,222 entries in these four columns are **null** or **empty**. This is a very strong indicator that these columns are entirely useless for your analysis and should be dropped, as you noted in your previous context.

It is observed that service fields 3,4, and empty fields 1 and 2 have no values.

# Data Preparation

```
#So drop the empty columns and columns with duplicate data.  
columns = ['sf3', 'sf4', 'ef1', 'ef2', 'sAccID', 'NoDescription']  
df.drop(columns, inplace=True, axis=1)  
data = df
```

# Data Analysis

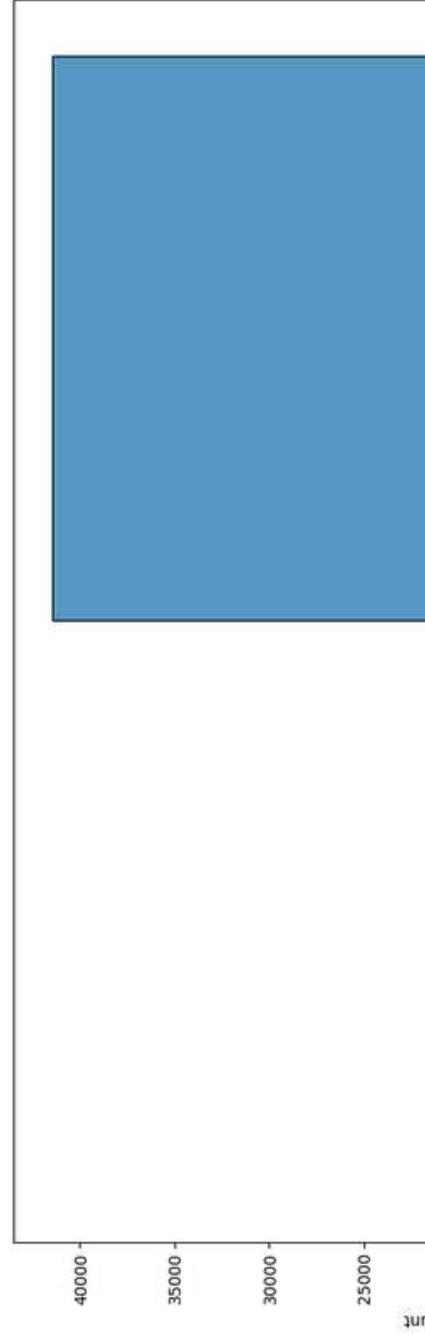
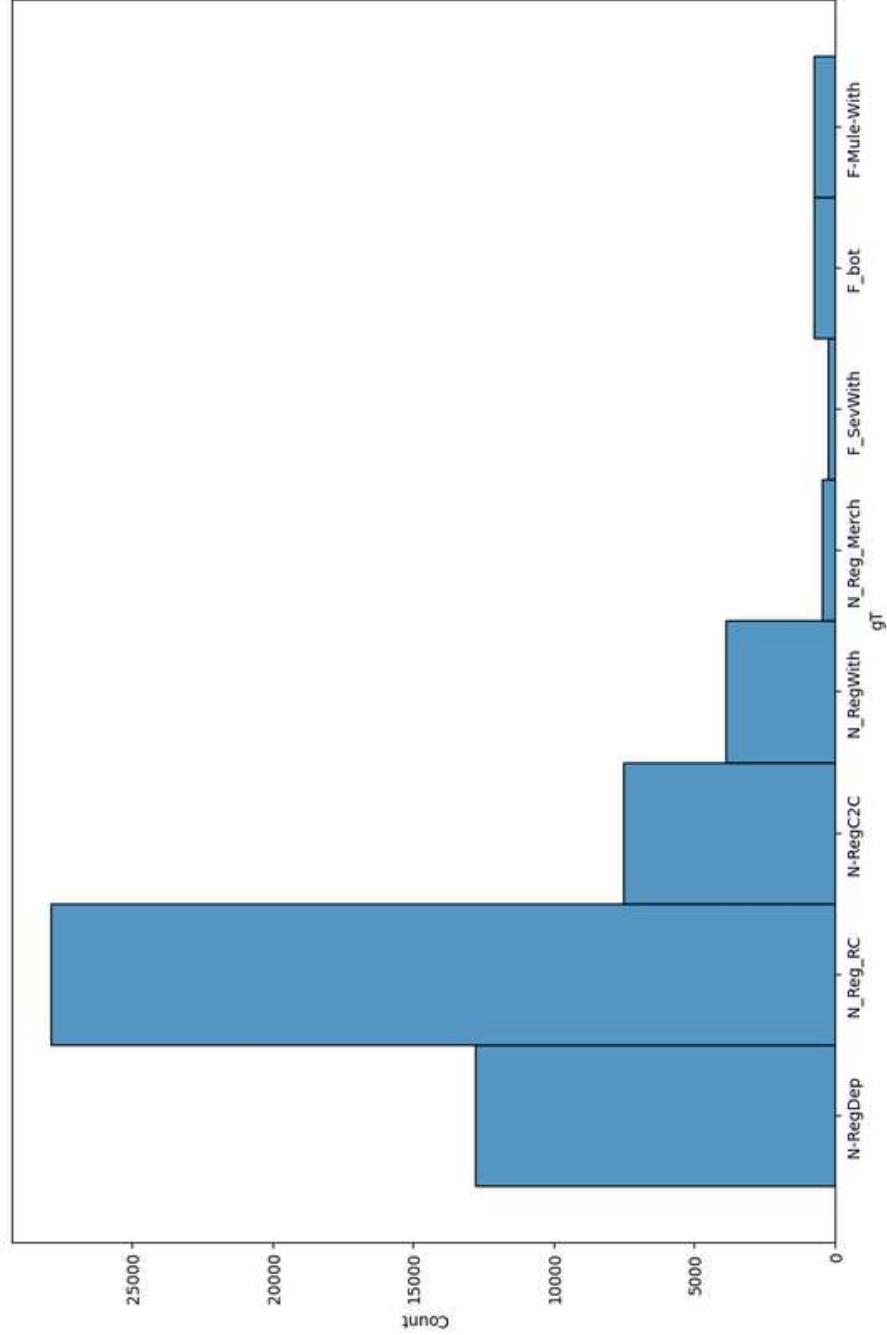
0

<b>gT</b>	N_RegDep	N_Reg_Ro
<b>sId</b>	PN_Ret2	PN_EU_1_50
<b>rld</b>	PN_EU_0_261	operator
<b>sAcc</b>	RAcc2	EUAcc1_50
<b>rAcc</b>	EUAcc0_261	Ac
<b>TranAmount</b>	131926.49	2054.4
<b>TranType</b>	Dt	ArRe
<b>TranStatus</b>	SU	SU
<b>sBalbefore</b>	1000000000.0	100000000.0
<b>sBalAfter</b>	999868073.51	99997945.5
<b>rBalBefore</b>	100131926.49	99180036.5
<b>rBalAfter</b>	100000000.0	99177982.7
<b>sf1</b>	True	True
<b>sf2</b>	True	True
<b>sTD</b>	1/6/2011 0:11:22	1/6/2011 0:16:3
<b>rTD</b>	1/6/2011 0:11:22	1/6/2011 0:16:3

df2=np.transpose(data)

df2

# Data Analysis



# Data Analysis

## 1. gT (Grouped Transaction Type)

mathematica

N_Reg_RC	27,981
N_RegDep	12,784
N_RegC2C	7,504
N_RegWith	3,899
F_Bot	731
F_Mule-With	729
N_Reg_Merch	442
F_SewWith	232



Insight:

- The majority of transactions are normal ( N\_Reg\_RC , N\_RegDep , N\_RegC2C )
- Fraudulent transactions ( F\_Bot , F\_Mule-With , F\_SewWith ) are relatively few ( 54,000+ ).

# Data Analysis

## 2. TranStatus

nginx

SU 54,222



Insight:

- Every transaction is marked as successful (SU).
- The dataset does not contain failed transactions (filtered out).

# Data Analysis

## 3. S Type (Sender Type)

riginx

EU	41,438
RET	12,784



Insight:

- Most senders are End Users (EU).
- A smaller portion are Retail accounts (RET).

# Data Analysis

## 4. rType (Receiver Type)

Category	Count
operator	27,981
EU	21,819
RET	4,860
MER	442



Insight:

- Most receivers are operators or other end users.
- Very few are retailers or merchants ( MER ).
- This suggests the system mainly facilitates person-to-person and merchant-based payments.

# Data Analysis, Cloud

- Dataset size: 54,222 successful transactions.
- Fraudulent cases: ~1,700 (3%) – very imbalanced class distribution
- Senders: mostly end users.
- Receivers: mostly operators and end users.
- Merchant transactions are rare, which may be unusual

# Building the Fraud Detection System

Part	What the Code Does	Purpose
1	Convert dataframe columns ( <code>sAcc</code> , <code>rAcc</code> , <code>TranAmount</code> ) into Python lists. Count total and unique senders/receivers.	Prepare size & unique
2	Create a directed transaction graph ( <code>DiGraph</code> ) with 1000 edges. Each edge = one transaction (sender → receiver). Edge thickness = transaction amount.	Visualize
3	Create an undirected graph ( <code>Graph</code> ) with all transactions. Edge weights = transaction amounts. Add hover tooltips.	Broaden
4	Build another graph ( <code>G2</code> ) with all transactions again.	Main workflow
5	Calculate node degrees (number of connections). Select accounts with $\geq 10$ transactions as "repeated nodes."	Filters out fraud suspects

# Building the Fraud Detection Pipeline

- 1 Import libraries
- 2 Load data
- 3 Check for missing values
- 4 Create a new column 'amount' for total transaction amount
- 5 Create a new column 'type' for transaction type
- 6 Create a new column 'category' for transaction category
- 7 Create a new dataframe df2 containing only transactions from possible fraud accounts.
- 8 From df2, drop transactions that are not of type Ind (Individual) or W1 (Withdrawal). Keep only these two types.
- 9 Calculate the average transaction amount. Mark transactions with amount ≥ average for removal.
- 10 Drop transactions above average → keep only smaller-than-average transactions in df4.
- 11 Print number of transactions left in df4.
- 12 Visualize graph again with PyVis:
  - Normal transactions (first 1000) in gray.
  - Fraud transactions (last 100) in red.

# How the Graph Connects

Concept	Meaning	Where in Code
1. Fraudulent transactions are smaller than the average transaction amount	Fraud often hides in small below-average transfers.	Part 9 & Part 10
2. Mules can perform legitimate transactions but sudden change in amounts = anomaly	Mule accounts look normal, but unusual patterns (e.g., sudden shifts, many transactions) are suspicious.	Part 5-7
3. Fraudsters = accounts doing multiple Individual/Withdrawal transactions below average amount	Accounts repeatedly doing small "Ind" or "W"	Part 8 & Part 9

# How the Graph Connects

Concept	Meaning	Where in Code
1. Fraudulent transactions are smaller than the average transaction amount	Fraud often hides in small below-average transfers.	Part 9 & Part 10
2. Mules can perform legitimate transactions but sudden change in amounts = anomaly	Mule accounts look normal, but unusual patterns (e.g., sudden shifts, many transactions) are suspicious.	Part 5-7
3. Fraudsters = accounts doing multiple Individual/Withdrawal transactions below average amount	Accounts repeatedly doing small "Ind" or "W"	Part 8 & Part 9

# Why No ML/DL Way

1. Exploratory Phase
  - The notebook's goal was to **explore** data and understand fraud.
  - Instead of training a classifier, you applied domain-driven rule (transaction amount).
2. Severe Class Imbalance
  - Fraudulent cases are < 3% of total transactions.
  - Training ML/DL directly on such skewed data without handling learning) would lead to a biased model that just predicts "no".
3. Graph Structure
  - Transactions were modeled as a network (graph).
  - ML/DL would need **Graph Neural Networks (GNNs)** or embedding (DeepWalk) to handle graph structure. That's more complex.
4. Interpretability
  - The rule-based method (small amount, frequent transaction

# References

- [1] Evgenia Novikova, Igor Kotenko and Evgenii Fedotov. *Interaction Mobile MoneyTransfer Services*, International Journal of Mobile 6(4), 73-97, October-December 2013.
- [2] Rieke, R., Zhdanova, M., Repp, J., Giot, R., & Gaber, C. (2011) Process Behavior Analysis. The 2nd International Workshop on Process Management (RaSIEM 2013) (pp. 66-75).

中華人民共和國  
全國人民代表大會常務委員會  
關於修改《中華人民共和國憲法》的決議