Payment Security – Smart Fraud Detection & Analysis

Milestone 1: Data Collection & Preprocessing

In this milestone, we focus on collecting and preprocessing data that will serve as the foundation for our database, data warehouse, and ML models for fraud detection.

1️⃣ Dataset Selection

We decided to use the Kaggle dataset "Bank Transaction Fraud Detection" to build our database, data warehouse, and big data pipeline.

- This dataset identifies and helps prevent fraudulent activities in financial transactions.
- It contains key attributes relevant to fraud detection, including:
    - Customer_ID
    - State
    - City
    - Transaction_Date
    - Transaction_Amount
    - Transaction_Type
    - Merchant_ID
    - Device_Type
    - Customer_Email

These attributes are expected to influence the probability of fraudulent transactions.

2️⃣ Preprocessing Steps

- Handling missing or inconsistent data
- Transforming and formatting dates and transaction types
- Ensuring realistic distribution of fraudulent vs. non-fraudulent transactions