WARSAW UNIVERSITY OF TECHNOLOGY

FACULTY OF MATHEMATICS AND INFORMATION SCIENCE

Real-time fraudulent transactions detection

Big Data Analytics

Salveen Singh Dutt (317298)
Karina Tiurina (335943)
Mikołaj Malec (298828)
Patryk Prusak (305794)

Supervisor

mgr inz. Jakub Abelski

Warsaw 2024

Introduction

The goal of this project is to plan and implement financial transactions processing system which identifies suspicious and fraudulent activity in real-time. Given the large volume of incoming data, the project will utilize big data technologies as well as advanced machine learning algorithms for anomaly detection.

High level description

The main idea is to implement an automatic transactions processing so that anytime a fraudulent activity occurs, the transaction is blocked for further manual review. The aim is to reduce financial losses of the end-users and to enhance the security of online payment, ensuring a safer experience for all customers.

There are two main end-users of the project: financial institutions (we will call them 'Managers') and their customers executing the payments. Although both categories can benefit from the solution, in our implementation we will mainly focus on Managers to limit additional data in storage.

The list below contains main features that we expect to implement for Managers:

1. Fraudulent transactions are automatically highlighted so that it is easier to identify suspicious activity;

2. The history of transactions is stored and available for later review;

3. A dashboard with statistics of fraudulent activity is available and customizable for better localisation of issues (e.g. too large amount, unusual location);

4. Anomaly-detection model is continuously updated so that fraud detection utilizes new historical data and is more accurate on future transactions;

5. Data streaming processing and batch jobs are customizable so that the testing of model's performance is simplified.

Data sources

Due to strict security regulations on personal and financial data, it is quite challenging to find open source real transactions data both for model training and streaming. Therefore, available synthetic and anonymized datasets will be used. The table below contains description of the preliminary data sources.

Data Source	Content	Volume	Link
Fraudulent Transactions Data	Dataset for predicting fraudulent transactions for a financial company. Available feautres: step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDests, oldbalanceDest, newbalanceDest, isFraud, isFlaggedFraud	6,362,620 rows and 10 columns (493.53 MB)	Kaggle
Credit Card Fraud	Contains features with transactional context like geographical location, transaction medium, and spending behavior relative to the user's history. Available feautres: distance_from_home, distance_from_last_transaction, ratio_to_median_purchase_price, repeat_retailer, used_chip, used_pin_number, online_order, fraud	1,000,000 transactions (58.9 MB)	OpenML
Credit Card Transactions Synthetic Data Generation	A collection of synthetic credit card transaction data. Each transaction record includes a variety of features commonly associated with credit card transactions, including transaction amount, merchant category code, time of transaction, and more	1,785,308 transactions; 5,000 customers; (153.66 MB)	Kaggle
Credit Card Fraud Detection	Transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.	284,807 transactions (150.83 MB)	Kaggle

Table 1: Data sources

Data-streaming API will be implemented from scratch. The assumption is that it will use the above datasets; with a specified time-frame, it will choose a random transaction which was not used for training and push it for further processing. For the testing purposes, the probability of a fraudulent transaction will be set manually to some high enough constant value.

Data processing

Generally, financial transactions are expected from a different sources, from which we highlight two: Web Transfer and Credit Card. Data from streaming api will be collected and saved as master data in storage. Further processing of it will be divided into three layers:

1. Speed Layer (streaming)
    - Data preprocessing including transformation to a specific format;
    - Real-time fraud detection on all of the incoming transactions.
2. Batch Layer
    - Data processing and filtering for the model training
    - ML model training with a fixed schedule (e.g. every 10 minutes)
3. Serving Layer
    - Client interface highlighting fraud transactions, accepting/blocking transactions;
    - Data visualization with customizable filters

Project architecture

Figure 1 shows an outline of the project architecture.

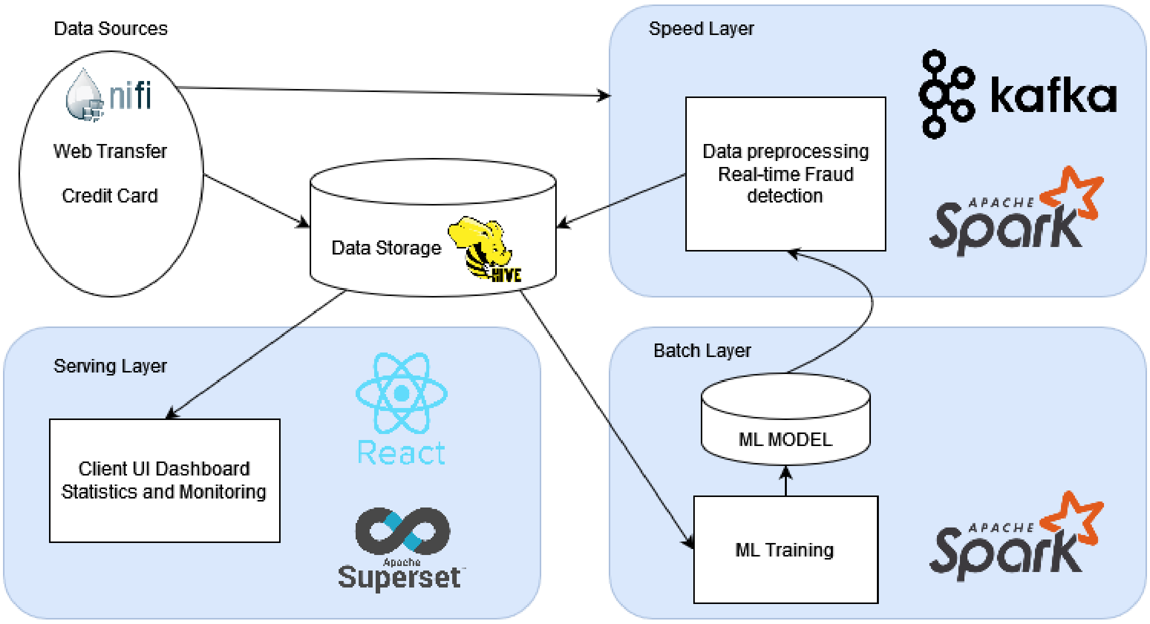

Figure 1: Project architecture

The following Big Data platforms will be used:

- Apache NiFi: to collect and distribute the data from different sources;

- Apache Hive as a data storage;
- Apache Kafka: to work with the streaming data;
- Apache Spark: to make the batch processing and model training;
- Apache Superset (to be agreed with the supervisor): for data analysis on the user-interface. If the service will not be approved, UI will be implemented from scratch using JS framework, e.g. React.

Preliminary tasks assignment

The table below contains the list of team members and preliminary allocation of tasks to team members,

Team member	**Tasks**	**Supporter**
Salveen Singh Dutt	Batch processing of the historical data for up-to-date model training (Batch Layer).	Karina Tiurina
Karina Tiurina	Fraud detection model training and fine-tuning; Data stream processing (Speed Layer).	Salveen Singh Dutt
Mikołaj Malec	Data ingestion, collection and pre-processing.	Patryk Prusak
Patryk Prusak	Data visualisation and configuration on the UI (Serving Layer).	Mikołaj Malec

Table 2: Tasks assignment