Project Proposal for a
Credit Card Fraud detection system using Apache Spark

Instructor:
Dr. Essam Mansour

GROUP 14 TeamMembers:
Apoorv Semwal – 40083939
Ekjot Singh – 40082643
Hartaj Singh Waraich– 40115775
Piyush Kumar – 40119786

- **Problem Identification and its importance:** In this modern era, people are rapidly shifting towards digital payment methods, leading to a skyrocketing growth in the number of digital transactions being carried out. According to the World Payments Report 2018 published by Capgemini and BNP Paribas, global non-cash transactions are estimated to reach 726 billion by 2020 **[1] [2]**. This surge in the use of digital payment methods also resulted in the rapid growth of digital frauds. To put things into perspective, the fraud valuation by 2020 exceeds the combined profits posted by Coca-Cola ($2 billions), Warren Buffet's Berkshire Hathaway ($24 billions) and JP Morgan Chase ($23.5 billions) in 2017 **[3]**. So it becomes imperative to detect these fraudulent transactions if we do not want the card users to bear the consequences, viz., money loss or poor credit rating. We are addressing this concern by building a Machine Learning model on a fiscal dataset, containing regular and fraud transactions, and predict whether a new transaction can be labelled as fraudulent or genuine.

- **Methodology:** The present architecture we have in our mind heavily relies on **distributed processing framework Apache Spark** and **Cassandra as our distributed storage layer**. We have primarily split the overall solution into **3 core Apache Spark Jobs**, namely:
  a. **First SparkSQL** based job for transforming and then saving our datasets into **Cassandra**. (Basically **handling the E(Extract)-T(Transform)-L(Load) part**). To start with we have obtained a **144MB licensed dataset from Kaggle with a usability rating of 7.1**.**[4]**
  b. **Second SparkML** based job that loads the transformed dataset from Cassandra and **trains a model** on it. As for the Model, we might be comparing the accuracy of multiple models but to start with we would consider training a **Random Forest or a Support Vector Machine** for classifying a transaction as genuine or fraud.
  c. **Third SparkStreaming** job that would use the trained model to make predictions for any new incoming transactions. New incoming transactions would either be **fed from a new data set or** if time permits, we would like to consume new transactions from a pre-set **Apache Kafka Topic.**
  d. We are planning to use **Eclipse** as our development environment with **Java** as our primary language used during development. **Maven** would be used **for managing our builds and dependencies.**

- **Evaluation:** In order to have precise evaluation metrics we preferred defining some measurable outcomes that are exactly aligned with our project design described above. They are as follows:
  a. After successfully pre-processing the dataset available from Kaggle and loading it to Cassandra we should be able to observe roughly 280,000 transactions with some 20 features for each transaction.
  b. After successfully training our model for the prepared data set we should be able to achieve an accuracy of around 85% in our predictions for the new transactions.

- **Timeline:** Based on the project plan, to exactly quantify the figures, we are proposing a timeline of **7 to 8 weeks for developing an end to end working solution**. However we fully understand that, based on what the course and the instructor expects, our proposed plan might get refined over the implementation phase but aiming to be an agile team we should be prepared for accommodating any such changes.

A proper breakup of our milestones would be as follows:

- **Milestones:**
  1. Setting up the required distributed platforms and ensuring we are seamlessly able to make them communicate. It would involve setting up **Apache Spark**, ensuring we are able to use the various libraries it has to offer **(Spark SQL, Spark ML and Spark Streaming) [5]**, and are able to communicate with **Cassandra as our storage layer**. **Achievable by 10th-Oct-2019 (1.5 Weeks).**

  2. Preparing our **first Spark Job based on Spark SQL** **Achievable by 20th-Oct-2019 (1.5 Weeks).**

  3. Preparing our **second Spark Job based on Spark ML** **Achievable by 7th-Nov-2019 (2.5 Weeks).**

  4. Preparing our **third Spark Job based on Spark Streaming** **Achievable by 17th-Nov-2019 (1.5 Weeks).**

  5. Preparing Final paper and Presentation. **Achievable by 24th-Nov-2019 (1 Week).**

- **Risks:** Following are the some of the risks we have identified which we might encounter during the course of this project:
  1. **Operating system and IDE compatibility issues** while setting up the various platforms. Although our plan is to use Windows with Eclipse as our IDE but in order to handle this risk we **would also be setting up a virtual machine with Ubuntu and IntelliJ on it**.
  2. **Learning curve** involved in getting familiar with new technologies getting **too steep**. Being eager newcomers to the domain, on account of getting stuck with some issue we would **immediately prefer seeking assistance from the instructor or the lab assistants** rather than losing time on it.
  3. Managing time and coordination in efforts with different individuals having their own **commitments to other courses and jobs would be a challenge**. We would be handling this by deciding the days in a week where all four of us are available and can work together. In order to synchronize our work we would be using **Git-Hub** as the version control system.

- **Division of Work:** Although we are planning our tasks in a way that all four of us can contribute equally and but at the same time in order to respect the timelines it makes sense to divide the work so that tasks can be developed in parallel and get integrated afterwards.
  A rough distribution would be like:
  - All four of us getting involved in setting up the development environment on our respective machines. **(20% of the overall efforts)**
  - Piyush and Hartaj working on developing the Spark Jobs for Initial Data Preparation. **(30% of the overall efforts)**
  - Apoorv and Ekjot working on developing the Spark Jobs for Model Training and Prediction. **(35% of the overall efforts)**
  - All four of us getting involved in integrating the various jobs for an end to end solution. **(15% of the overall efforts)**

    This breakup in any way does not really fix the responsibilities and we as a team would all be sharing our resources and learning at every step of the development. We hope that this project ends up being a great learning experience for our team and for the entire class as well.

- **References:**

[1]: **https://worldpaymentsreport.com/wp-content/uploads/sites/5/2018/10/World-Payments-Report-2018.pdf**
[2]: **https://www.cnbc.com/2017/10/09/digital-payments-expected-to-hit-726-billion-by-2020-study-finds.html**
[3]: **https://nilsonreport.com/upload/content_promo/The_Nilson_Report_10-17-2016.pdf**
[4]: **https://www.kaggle.com/mlg-ulb/creditcardfraud**
[5]: **https://spark.apache.org/docs/latest/quick-start.html**