**BDM Proposal: Realtime Fraud Detection System**

Pawel Urbanowicz, 108015016

Martin Ledl, 108012012

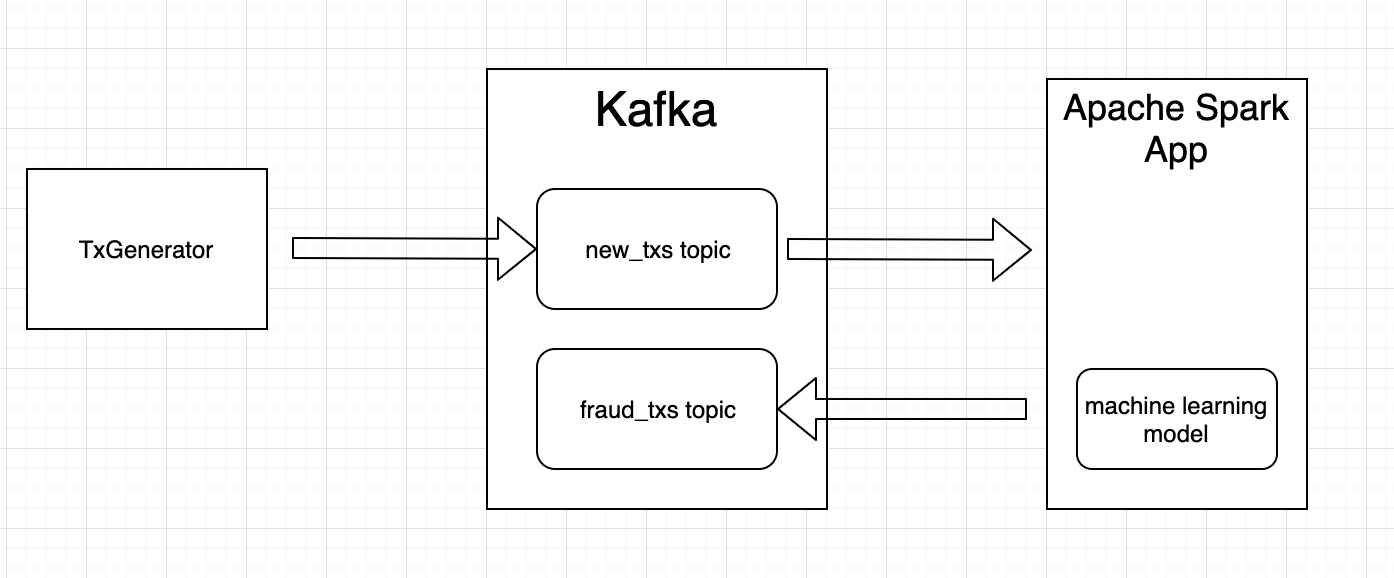
Tobias Kick, 108998413

Our aim is to create a real time fraud detection system. We're planning to use dataset from a recent Kaggle competition (available at: https://www.kaggle.com/c/ieee-fraud-detection). The technology stack will consist of Apache Spark, Apache Kafka, Docker and Python as programming language.

Our data consists of the following two tables which are joined together by a unique transaction identifier and divided into training and test datasets:

* Identity: consists of 41 features, which represent identity information such as the network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions.
* Transaction: consists of 394 features, interesting features are different payment timedeltas and payment amount as well as credit card information, the purchased product and information about the purchaser.

**Architecture for the real-time processing:**



* TxGenerator - imitating real time transactions & pushing into ‘new\_txs’ topic.
* Application/Script for training machine learning model
* Apache Spark App for detecting fraud transactions in real time & pushing back into ‘fraud\_txs’ topic.

**Team members responsibilities:**

* Set up the first flow between all elements of the above architecture. [PAWEL]
* Create the TxGenerator application [PAWEL]
* Conduct data preprocessing [PAWEL, MARTIN, TOBIAS]
* Create the machine learning model [MARTIN, PAWEL]
* Create the Apache Spark app [MARTIN, TOBIAS]
* Presentation and report [MARTIN, TOBIAS]

**Prefered time slot for presentation:**

We would prefer the first time slot for our presentation (on date: Dec. 27, 2019).