Real-Time Credit Cards Transactions Analysis

Bianchi Edoardo
Free University of Bolzano/Bozen

Abstract: Due to the number and variety of online paid services, the amount of transactions happening in real-time is huge. The goal of this project is build a system based on streams that updates a dashboard presented to the final user.

1. Application Domain

The chosen application domain concerns transactions that take place online via an electronic payment method, that is credit cards. The final aim of the project is to analyze some aspects of the transactions that are carried out by online users, allowing you to monitor what happens in real time.

2. Data Source

Given the difficulty in finding real time source of transaction data, we will use a *csv* file containing simulated transaction generated with a data generation tool. The dataset can be found at this page on Kaggle. For this project some of the columns present in the dataset will be considered, in particular: id, category, amount, gender, city, state, city population, job and timestamp.

Data will be accessed with a python script, *producer.py* that ”converts” the data into a stream-like source, sending json data to Kafka.

3. Technologies Architecture

The application consists of different components:

1. **Producer**: Used to generate the stream of data, simulating a real time source. Sends out json data to a Kafka topic.
2. **Kafka + ZooKeeper + Kafka UI**: To store topics (original topics and topics resulting from processing steps)
3. **Spark**: Used to compute new topics with query results.
4. **Kafka WebSocket proxy**: Used to consume topics in the browser.
5. **RxJS + ECharts**: The final dashboard presented to the user, it uses RxJS to get data from Kafka topics, Bootstrap (or other frameworks) and Apache Echarts for the pagination.
6. **Docker**: Used to manage some of the components.

4. Functionalities

The system is used to compute some queries on the data. The results are then showed to the final user using a dashboard. The final dashboard is made up of different types of graphs and plots, populated with the results of the queries performed.