Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

Predicting Cryptocurrencies Exchange Rates Based on Cryptocurrencies News Sentiment Analysis

Big Data Project

Solution Design

Authors:
Maciej Pawilkowski
Hubert Ruczyński
Bartosz Siński
Adrian Stańdo

Warsaw University of Technology
Faculty of Mathematics and Information Science
23 October, 2023

1 Business goal and potential benefits

The goal of this project is to create a system which enables its users to investigate the influence of the latest news articles on exchange rates of cryptocurrencies. Our tool will scrap current exchange rates, and recent news regarding cryptocurrencies in order to perform a sentiment analysis of those messages. Extracted features will be provided into the time-series predictive model that will present estimated exchange rates of selected cryptocurrencies, based on archival data, the most recent trends, and sentiment.

The end users will be able to track current exchange rates, our prediction, and the latest news concerning selected cryptocurrency. This information might help them make the best decision about when to exchange their money. As a result, they may save a lot of money or even find a way to influence main news sources to publish articles that would have positive impact on selected exchange rates.

2 Description of data sources

2.1 Coincap

Description: Coincap provides API with the real time cryptocurrencies data without API key.
Requests Limits: 200 requests per minute, exchange values update each 10 seconds.
Link: https://coincap.io/

2.2 Alpha Vantage

Description: Alpha Vantage website provides API with financial market data such as time series stock data, digital & crypto currency exchange rates and most importantly market news data which we will use in this project. Market news are both real-time and historical and come from global news outlets. They cover stocks, cryptocurrencies, forex and others. News data consists of title, url to the article, summary and the source. Additionally API returns sentiment analysis information. However we want to perform this analysis by ourselves and therefore we will not use this feature.

Requests Limits: 5 API requests per minute, 100 API requests per day. Possible academic access to premium plan with up to 1200 API requests per minute.

Link: https://www.alphavantage.co/

2.3 Crypto Compare

Description This API provides various crypto market data such as current price and trading information, block chain data and social stats for given coins. On top of that we can get information about latest news articles related to crypto currencies. We can choose categories and source feeds of requested articles. Similarly to the previous data source returned news data consists of title, url to the article, summary and the source.

Requests Limits: Lifetime - 250 000 calls, Day - 50 000 calls, Minute - 2 500 calls.

Link: https://min-api.cryptocompare.com/

2.4 NewsAPI

Description It is a simple HTTP REST API for searching and retrieving live articles from all over the web. API doesn't return entire articles only title, description, first 200 signs, author, and source. It allows us to request articles based on language, publication date and even based on keyword or phrase.

Requests Limits: We have access to articles with 24h delay, and can't pull articles older than one month. We can also make only 100 requests per day.

Link: https://newsapi.org/

2.5 CryptoPanic

Description JSON API that gives access to the most recent posts, indicators and sentiment data. It offers news articles and updates from various sources related to cryptocurrencies and block-chain technology. It also returns readers feedback in shape of number of likes, dislikes, comments, etc..

Requests Limits: We are limited to 200 last news posts, which we can request 5 times per second without max request limit.

Link: https://cryptopanic.com/developers/api/about

3 Architecture

Technologies which will be used in the project:

1. **Data flow:** `Apache NiFi`
   Raw data from APIs will be downloaded and preprocessed using `Apache NiFi` . Cryptocurrency data will be sent to both `HDFS Archive`, and `Apache Kafka Buffer`, as archival data will be used to cyclically train the model, and recent data will be used to fine-tune it. The crypto-news data will be also send to both places as as we tread the batch layer as a kind of back-up.

2. **Data storage:** `HDFS`, `Apache Hive` and `Apache HBase`
   As presented in the Figure 1, cryptocurrencies and crypto-news data will be saved in `HDFS Archive`. Additionally we will redirect the results of sentiment analysis here, whereas the stocks predictions will be accessible from the serving layer only. `Apache Hive` will be used as a serving layer for the Batch Layer data, whereas `Apache HBase` will be serving the current data from a Speed Layer.

3. **Data processing:** `Apache Spark`
   `Apache Spark` scripts will be widely used in our project at multiple steps, as presented in the Figure 1. First and foremost, it has components dedicated for NLP analysis such as calculating sentiment, thus it will be used for the ML part of the project. Moreover, we will use it to transform the data coming from `Apache Kafka Buffer`, and `HDFS Archive`, before they will be transported to the respective Serving Layer views.

4. **Speed Layer Buffer:** `Apache Kafka`
   `Apache Kafka` will serve as a buffer holding the recent cryptocurrency data used for the fine tuning of our model.

5. **Presentation Layer:** `Tableau` / `Jupyter Notebook` with `Apache Spark`
   We have to strategies concerning the presentation layer. The first one focuses on the usage of Business Intelligence (BI) tool, like `Tableau` to create a clean Dashboard with proper visualisations. The approach is however not certain, as we are not sure how BI systems can interact with our data sources, and if they are enabled in free versions. The second one is a backup, where we prepare visualisations in `Jupyter Notebook`, and transform data via `Apache Spark`. Both approaches will include a curve presenting recent exchange rates, and our prediction for the near future for selected cryptocurrencies. Moreover, on this plots we will mark the timestamps of latest news occurrences and print those news / their titles.

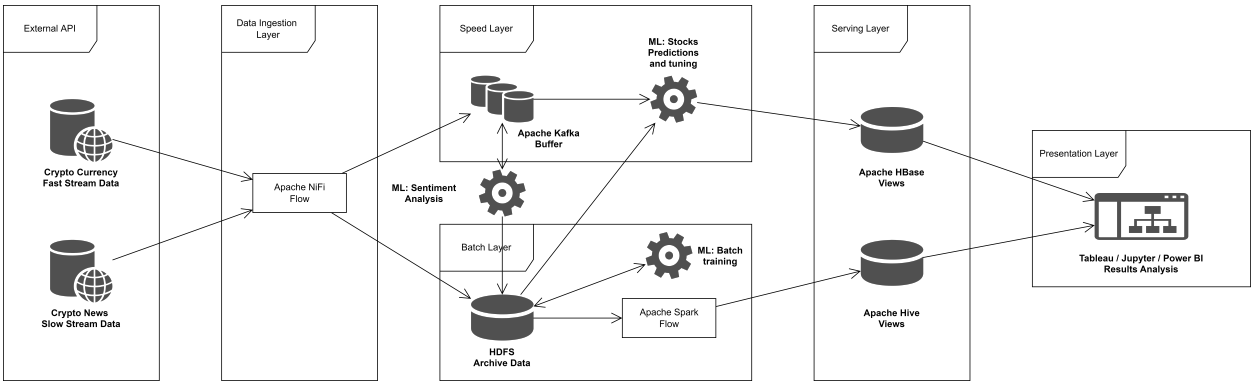

Figure 1: Architecture plan of the proposed solution.

4 Work division

Table 1: Work division between the team members.

Task Name	**Team Members**
Data Source Analysis	Maciej Pawlikowski, Bartosz Siński
Project Planning	Hubert Ruczyński, Adrian Stańdo
NiFi Flow Implementation	Adrian Stańdo
Speed Layer Implementation (incl. Spark)	Maciej Pawlikowski, Bartosz Siński
Batch Layer Implementation (incl. Spark)	Hubert Ruczyński
ML Implementation	Maciej Pawlikowski, Bartosz Siński
Serving Layer Implementation	Adrian Stańdo
Presentation Layer Implementation	Hubert Ruczyński

5 Technical solution

We plan to use multiple Docker containers to run all the services. If it is not possible to run them all on one machine, we will use a service like TailScale, which provides a zero-configuration VPN, and we will connect our local machines together.