Predicting Cryptocurrencies Exchange Rates Based on Cryptocurrencies News Sentiment Analysis

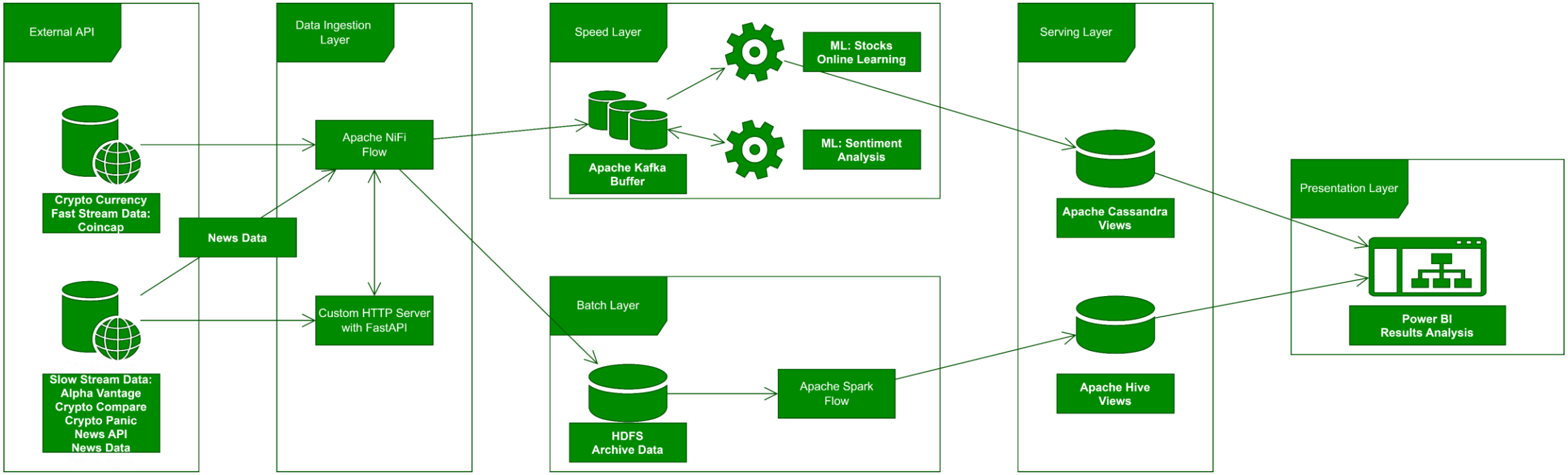
Maciej Pawlikowski, Hubert Ruczyński,
Bartosz Siński, Adrian Stańdo

Business goals

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.

Architecture reminder

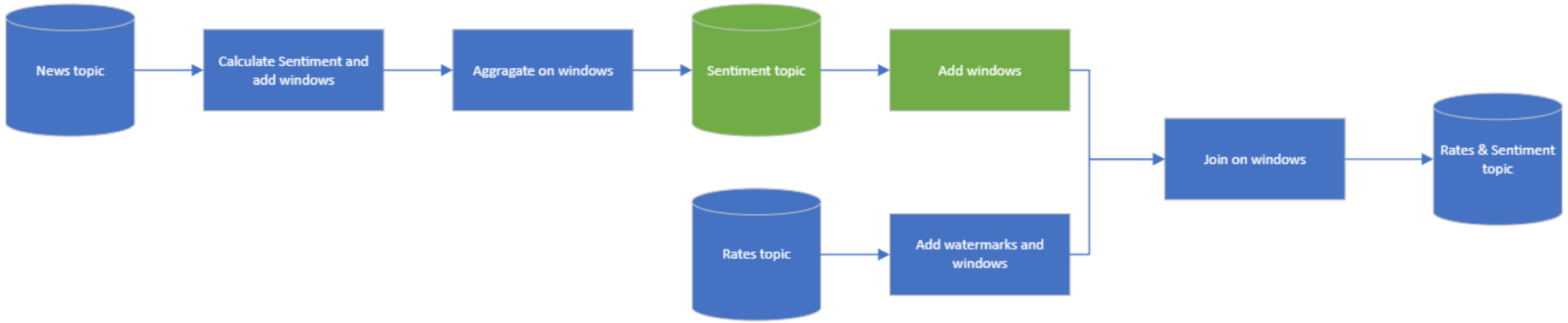


Batch views

article_id	content	time_stamp	final_bitcoin	final_ethereum	final_dogecoin
01b9662d24d32df3e...	The crypto market...	2024-01-12 18:55:30	true	null	null
873172e57834a7ee6...	Loading... Loadin...	2024-01-12 18:47:18	true	null	null
5154a634a904b0dd2...	Loading... Loadin...	2024-01-12 18:41:22	true	null	null
94f5022ad8700d20c...	The Bitcoin Ordin...	2024-01-12 18:39:29	true	null	null
5b4dd20959dd5374d...	Loading... Loadin...	2024-01-12 18:33:33	true	null	null
dd83eaf552dd83a45...	’s former CEO Do ...	2024-01-12 18:28:43	true	null	null
f9064bb56cf2cb77e...	CNBC presenter Ji...	2024-01-12 16:40:06	true	true	null
ce4eb9070d7dc650b...	This week, the Un...	2024-01-12 16:30:00	true	null	null
c7eb9aba7761820df...	The City of Peter...	2024-01-12 16:26:41	true	null	null
9b39d6e22b665d149...	Wall Street analy...	2024-01-12 16:26:06	true	null	null
7be882d47c1463c1b...	Hello! This is Ma...	2024-01-12 16:26:00	true	null	null
bcac4f98c9821bfbb...	Bitcoin struggled...	2024-01-12 16:15:03	true	null	null
72d10be2f72af8297...	Bitcoin struggled...	2024-01-12 20:15:02	true	null	null
eba0fc34d8effd410...	Gold has historic...	2024-01-12 20:14:00	true	null	null

id	symbol	rateusd	time_stamp
bitcoin	BTC	42453.607391795864	2023-12-18 21:31:...
ethereum	ETH	2211.009232299348	2023-12-18 21:31:...
dogecoin	DOGE	0.0916126651483873	2023-12-18 21:31:...
bitcoin	BTC	42453.607391795864	2023-12-18 21:31:...
ethereum	ETH	2211.009232299348	2023-12-18 21:31:...
dogecoin	DOGE	0.0916126651483873	2023-12-18 21:31:...
bitcoin	BTC	42453.607391795864	2023-12-18 21:31:...
ethereum	ETH	2211.009232299348	2023-12-18 21:31:...
dogecoin	DOGE	0.0916126651483873	2023-12-18 21:31:...
bitcoin	BTC	42454.45432141532	2023-12-18 21:31:...
ethereum	ETH	2211.034691149461	2023-12-18 21:31:...
dogecoin	DOGE	0.0916446774237382	2023-12-18 21:31:...
bitcoin	BTC	42454.45432141532	2023-12-18 21:31:...
ethereum	ETH	2211.034691149461	2023-12-18 21:31:...
dogecoin	DOGE	0.0916446774237382	2023-12-18 21:31:...
bitcoin	BTC	42451.98697537083	2023-12-18 21:32:...
ethereum	ETH	2211.2652637755523	2023-12-18 21:32:...
dogecoin	DOGE	0.0916362385217326	2023-12-18 21:32:...
bitcoin	BTC	42451.98697537083	2023-12-18 21:32:...
ethereum	ETH	2211.2652637755523	2023-12-18 21:32:...

Joins work!




```
+------------------------------------------+-------+------+------------------+-----------------+-----------------------+--------------------+
|window                                    |id     |symbol|changePercent24Hr |priceUsd         |currency_timestamp     |avg_sentiment       |
+------------------------------------------+-------+------+------------------+-----------------+-----------------------+--------------------+
|{2024-01-12 17:22:00, 2024-01-12 17:24:00}|bitcoin|BTC   |-5.6425508208902625|43498.94234103732|2024-01-12 17:22:13.549|5.673158458583333E11|
+------------------------------------------+-------+------+------------------+-----------------+-----------------------+--------------------+

-------------------------------------------
Batch: 35
-------------------------------------------
+------------------------------------------+-------+------+------------------+-----------------+-----------------------+--------------------+
|window                                    |id     |symbol|changePercent24Hr |priceUsd         |currency_timestamp     |avg_sentiment       |
+------------------------------------------+-------+------+------------------+-----------------+-----------------------+--------------------+
|{2024-01-12 17:22:00, 2024-01-12 17:24:00}|bitcoin|BTC   |-5.6455058911320470|43447.90908099249|2024-01-12 17:22:23.156|5.673158458583333E11|
+------------------------------------------+-------+------+------------------+-----------------+-----------------------+--------------------+
```

ML improvement

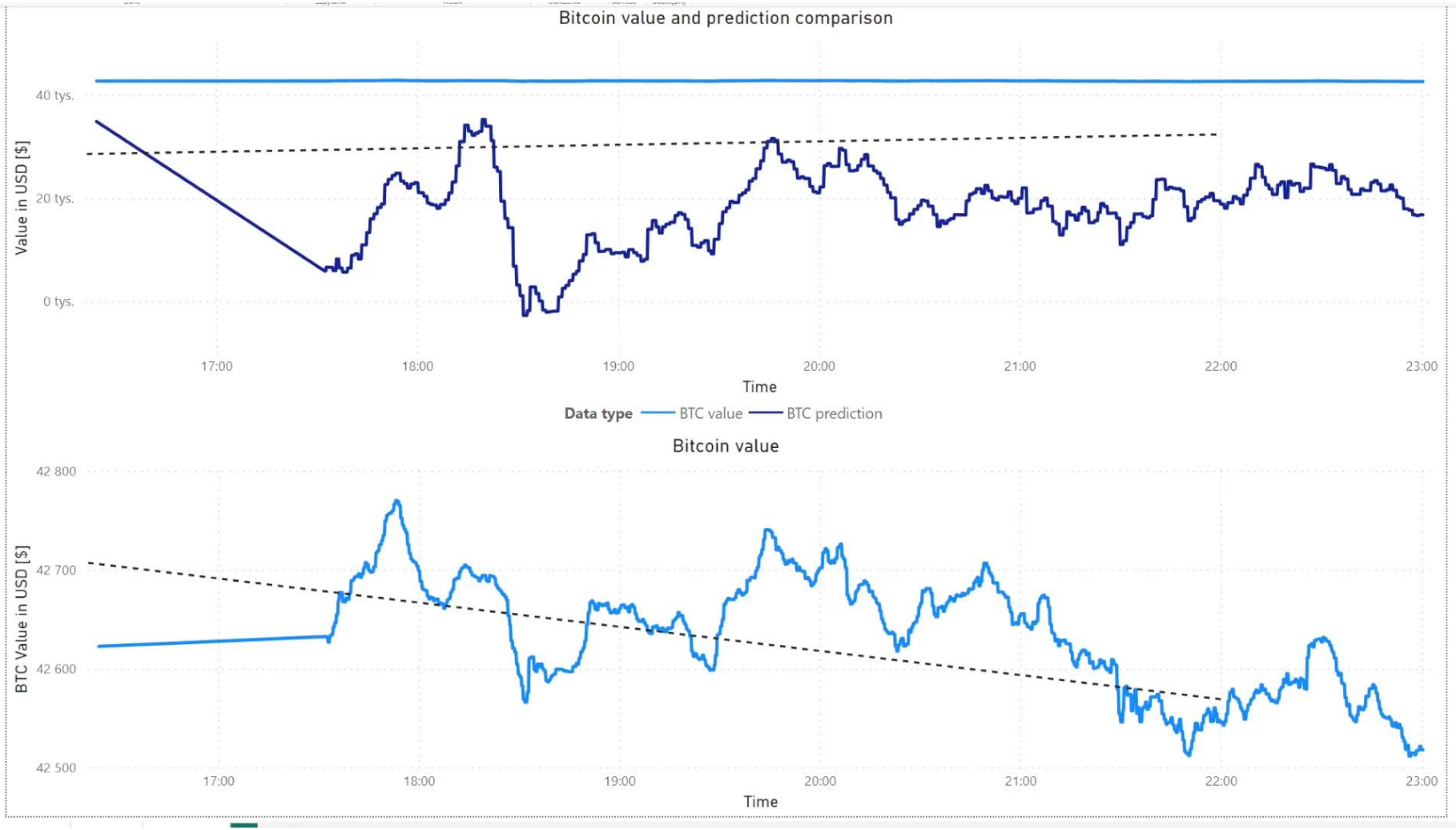
1. The oldest dummy and very baseline approach
   - *y(t) = y(t-1) * c(t-1)*,
     where *c(t-1)* is the direction and value change in the last 24 hours (data from API).

2. Curve-fitting approach
   - For 10 lagged values, fit a linear function and the prediction is the next point for the curve.
   - Model is retrained each time an instance comes to the system.
   - Naturally deals with missing data: less data to fit a curve.

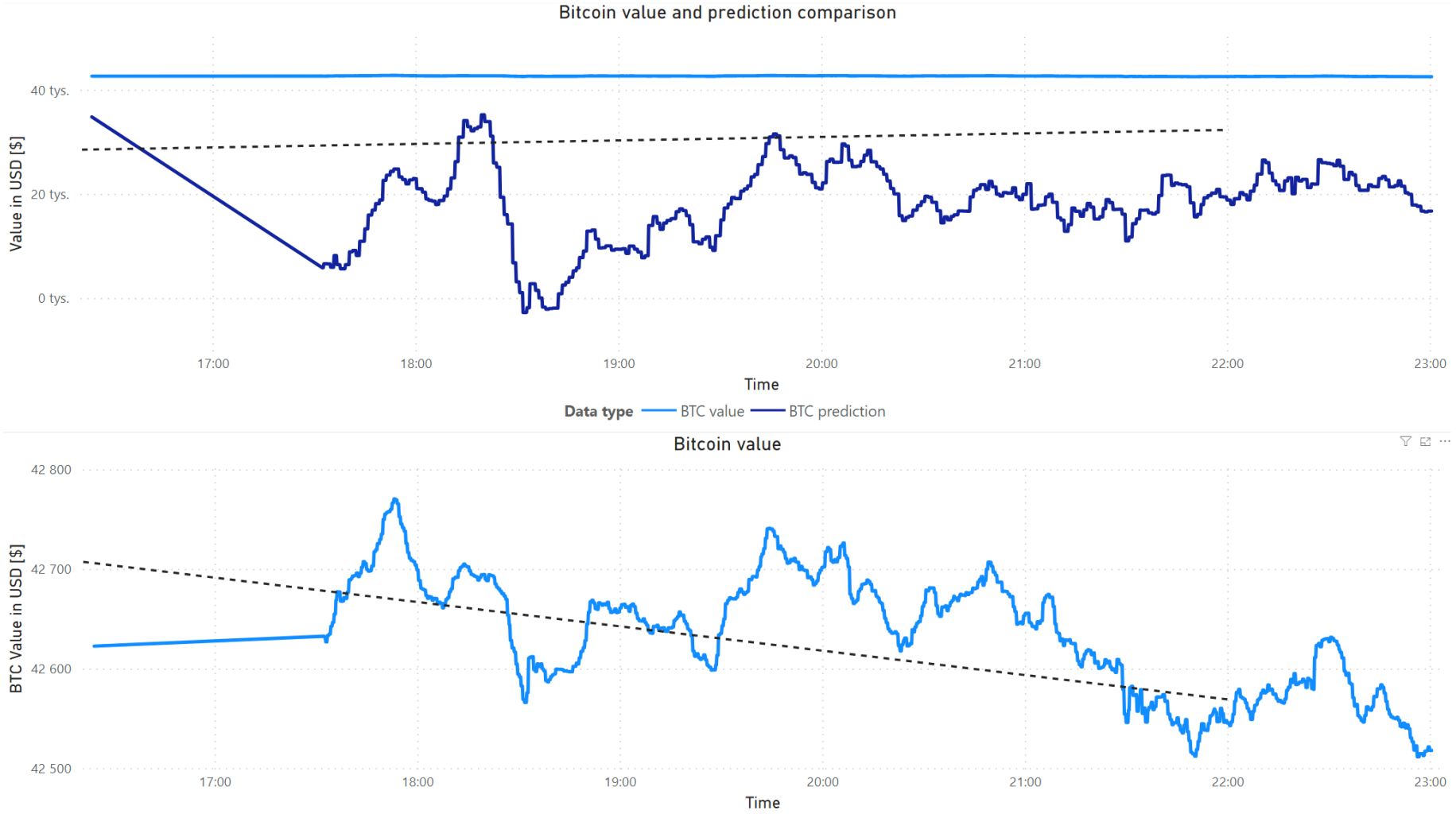
3. Adaptive Random Forest Regressor
   - Online learning algorithm
   - Adapts to the changes in data (ADWIN concept drift detector is included)
   - One model for each time series
   - Input features: 10 lagged values
   - Missing data: they are added by drawing values from the normal distribution with parameters estimated using the rest of the sliding window.

Performance estimation: **rolling RMSE for the last 10 values.**

Dashboard

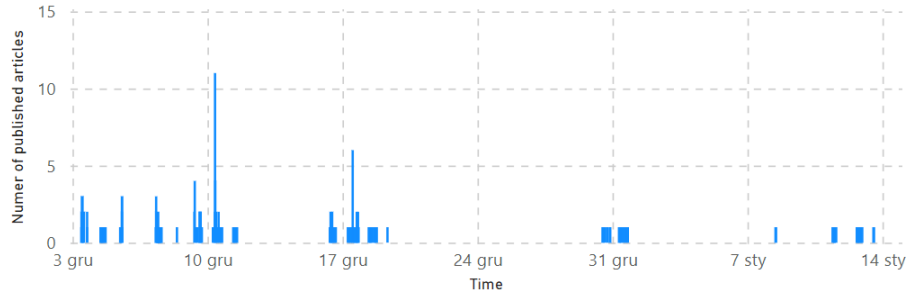

Bitcoin value and prediction comparison
40 tys.
20 tys.
0 tys.
Value in USD [$]
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time
Data type
BTC value
BTC prediction
Bitcoin value
42 800
42 700
42 600
42 500
BTC Value in USD [$]
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

Dashboard - baseline

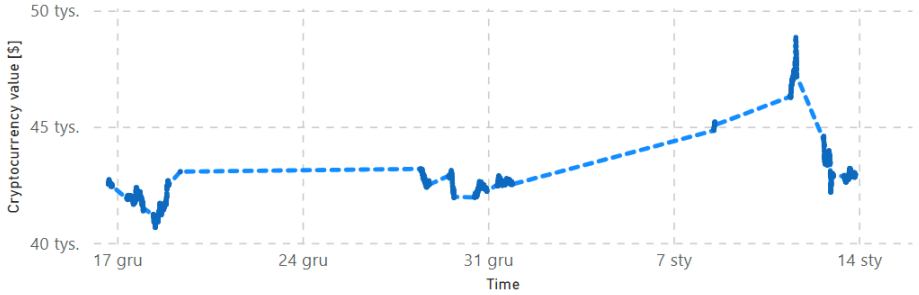


Dashboard - batch

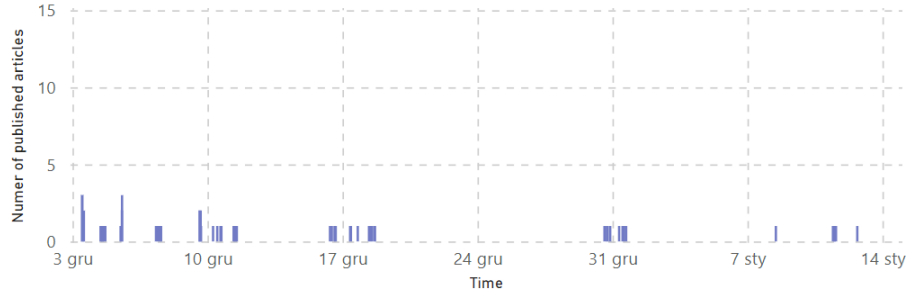
Number of Bitcoin (BTC) articles that occured at particular time



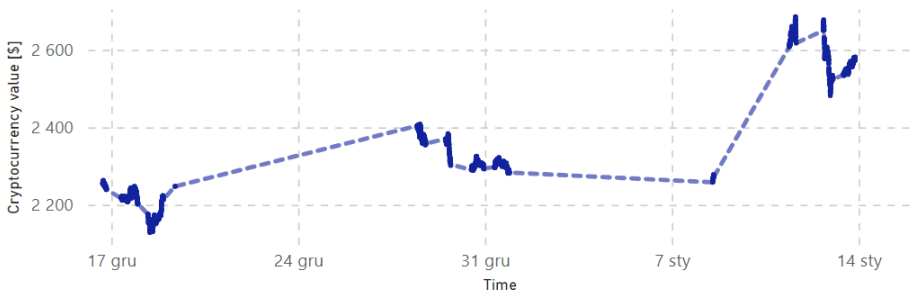
Bitcoin (BTC) value throughout project



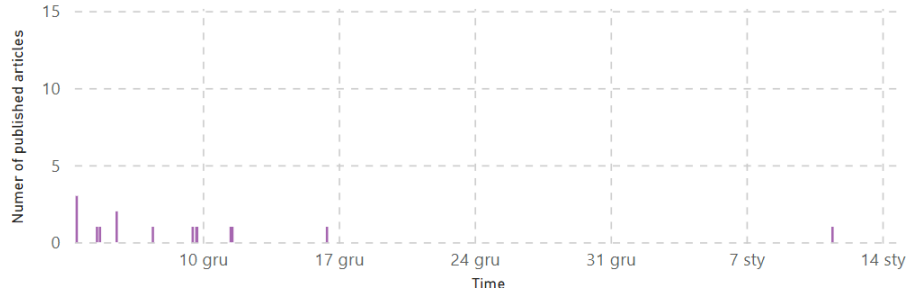
Number of Ethereum (ETH) articles that occured at particular time



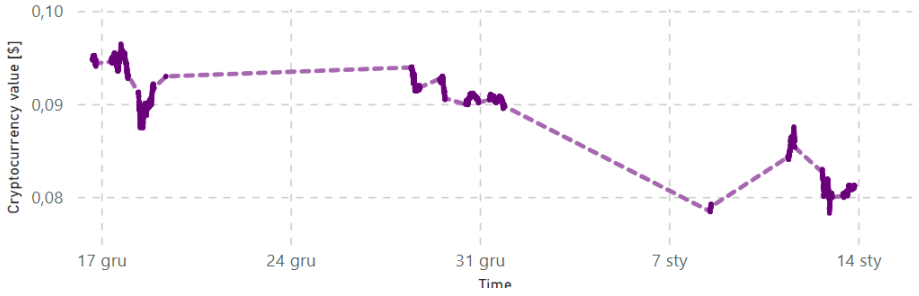
Ethereum (ETH) value throughout project



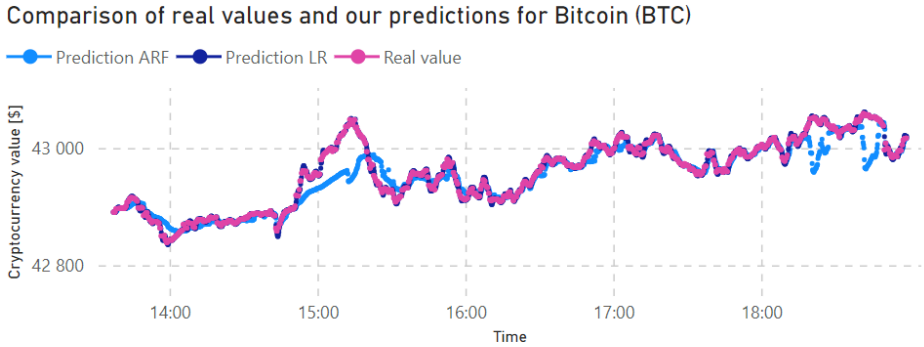
Number of Doge Coin (DOGE) articles that occured at particular time

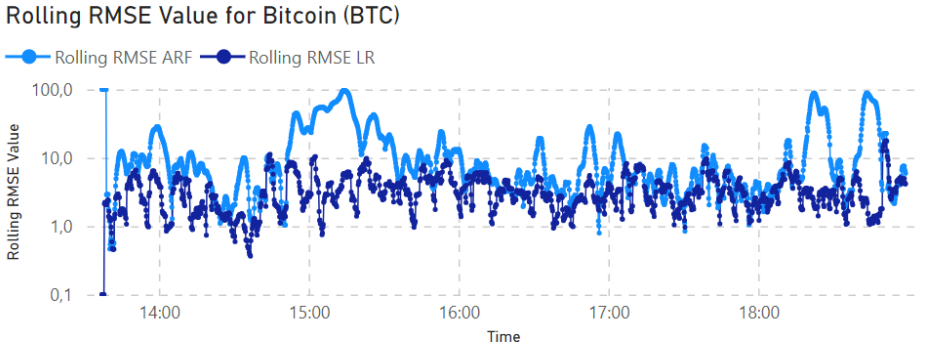


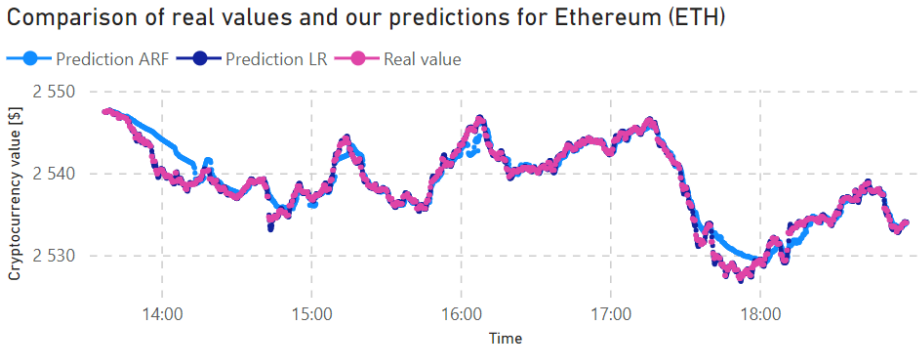
Doge Coin (DOGE) value throughout project

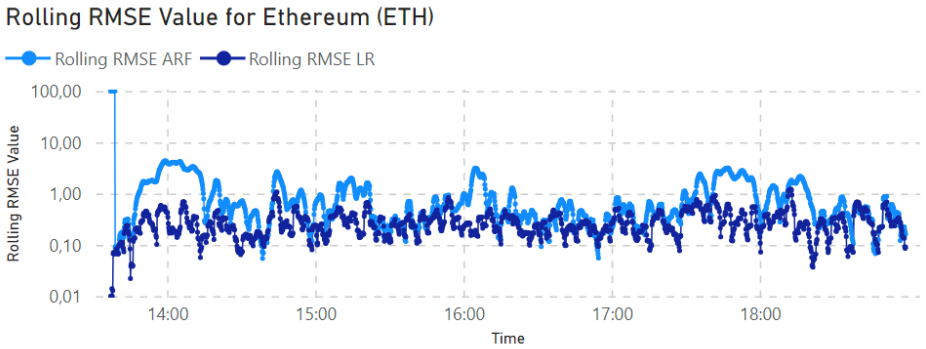


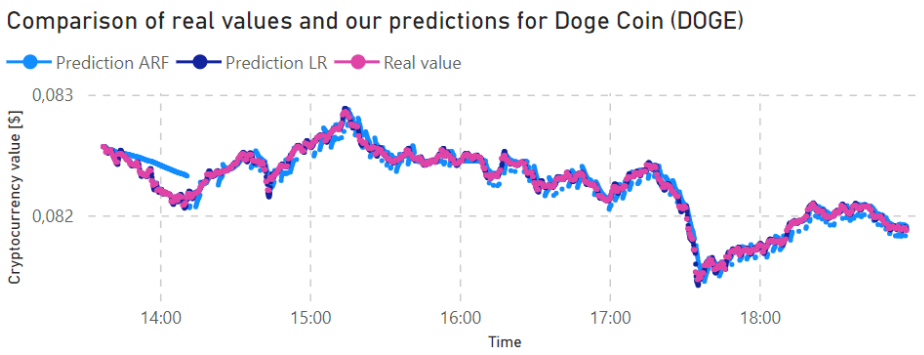
Dashboard – real-time

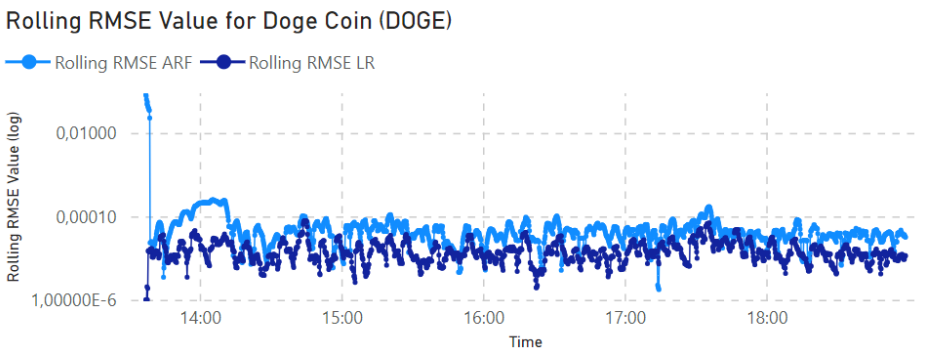












Quality and business benefits

1. By analyzing past trends in the data and reevaluating user intuition against model predictions users can gain an advantage when investing in the cryptocurrency market.
2. User can assess the accuracy of the model by analyzing past predictions.
3. Not including sentiment information – considerable drawback.
4. The effectiveness of our solution in forecasting future trends is significantly limited, as we are making predictions for the next window only.
5. Implemented system is relatively slow.

Non Techincal Aspects

1. If the user is blindly basing his business decisions on the results of our analysis, it can have a huge impact on the user's economic welfare.
2. If many people would invest according to our predictions then it could lead to permanent changes in market and cryptocurrency rates on the global scale.
3. Errors in data collected from the API cause faulty analysis, which could lead to the wrong investments.

Summary

1. We implemented a fully **dockerized**, in-house Big Data system, accessible online from scratch. Unfortunately, we were not able to make it a **distributed** solution.
2. We implemented the **Data Ingestion Layer** with the usage of **NiFi**, and wrote an online scrapper that gathers the texts with the usage of **FastAPI**.
3. We implemented the **Batch Layer** storage with the usage of **HDFS**.
4. We implemented the **Batch Views** in **Hive**, taking data from HDFS.
5. We implemented the **data buffering** in a Speed Layer with **Kafka**.
6. We implemented the **Sentiment Analysis** module in PySpark that puts the results in a separate **Kafka** topic. Unfortunately, we did **not use** it in our final solution.
7. We managed to merge separate data streams (cryptocurrency, and sentiment **Kafka** topic) into one in a separate **Kafka** topic. Unfortunately, we **did not use** it in our final solution.
8. We implemented **online learning models (ARF and LR)** that use the Speed Layer data to train and predict cryptocurrency rates. The results obtained by both models are **satisfactory**. Unfortunately, we did not manage to follow the initial ideas of (1) **using the sentiment** data for model training, and (2) we did not combine the **batch-training, and speed-tuning** concepts.
9. We implemented the **real-time views** in **Cassandra**, which takes the results from the Speed Layer.
10. We implemented the **Presentation Layer** in **Power BI**, as a Dashboard with 3 views: Dummy model results, Batch data, and Real-time data.


Thank you for
attention

Encountered issues

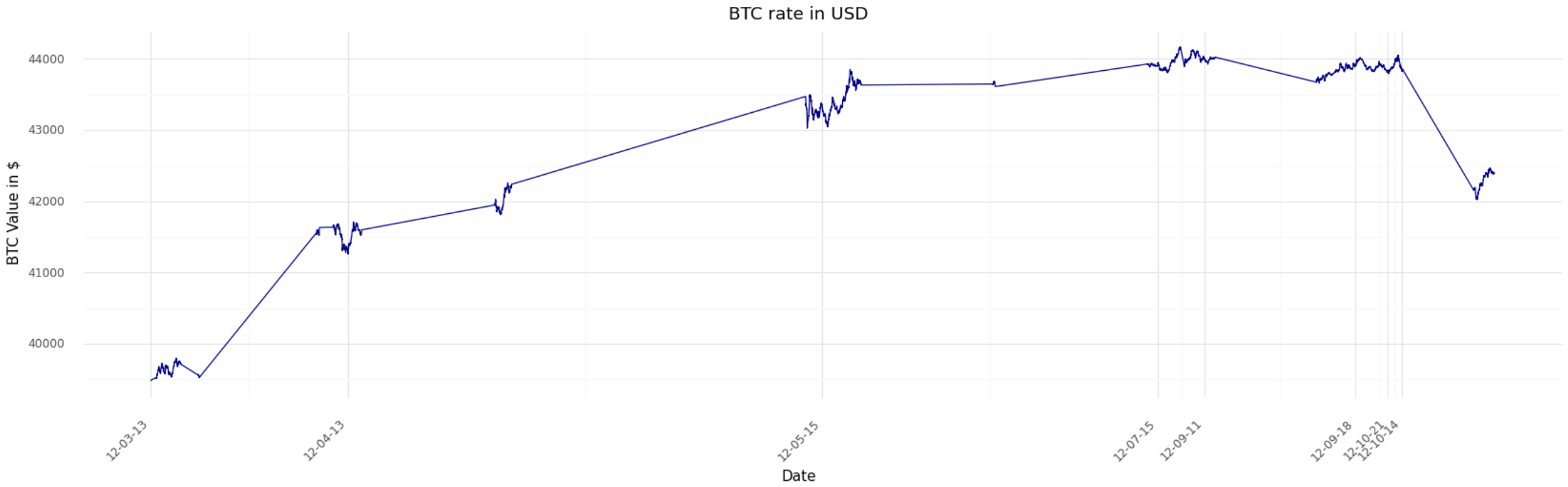
1. Docker images immediatly working alone, but not when combined.
2. Lots of guides for small parts, not a whole system.
3. Creating route tables for all containers, as the conainers don’t see each other.
4. The need of writing configuration scripts for all containers, as their operating systems are increadibly raw (lack of ping, or sudo commands).
5. Mounting the volumes, sometimes requires copying configuration files which are somewhere in the container.
6. Extremely complex docker-compose configuration files.

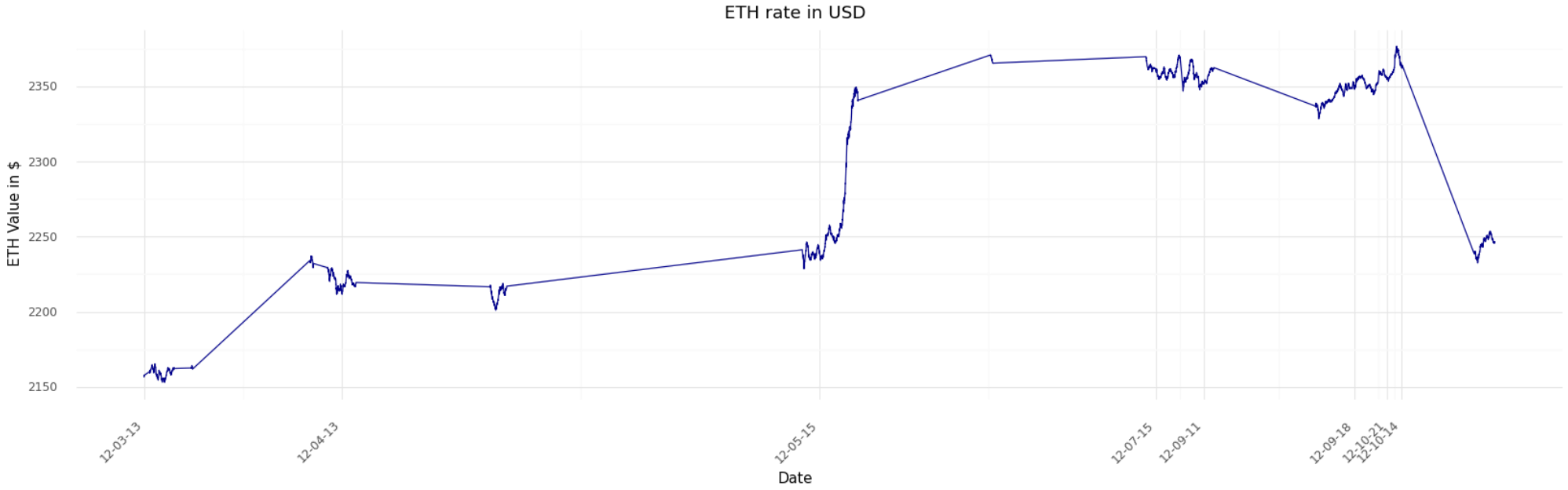
EDA – cryptocurrencies rates

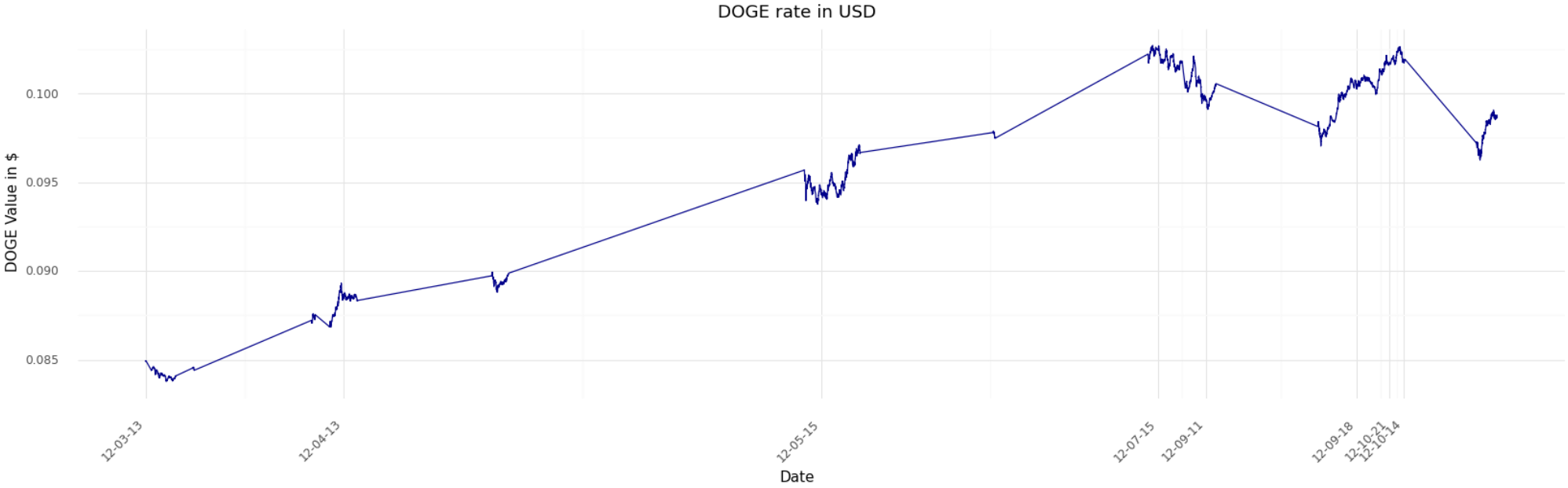
1. 17 unique cryptocurrenices: ('DAI', 'ZEC', 'LTC', 'XPRT', 'DASH', 'CRO', 'QTUM', 'ETH', 'DVPN', 'BTC', 'USDT', 'BNB', 'WAVES', 'EOS', 'DOGE', 'RUNE', 'BCH'),
2. 14452 records for each of them (different timestamps),
3. No missing values in meaningful columns.

	id	symbol	currencySymbol	type	rateUsd	timestamp	date
714	binance-coin	BNB	NaN	crypto	227.936381	1701606435166	12-03-13
715	thorchain	RUNE	NaN	crypto	6.976767	1701606435166	12-03-13
716	dash	DASH	NaN	crypto	31.650866	1701606435166	12-03-13
717	eos	EOS	NaN	crypto	0.699670	1701606435166	12-03-13
718	persistence	XPRT	NaN	crypto	0.244109	1701606435166	12-03-13

EDA – cryptocurrencies rates




ETH rate in USD
ETH Value in $
2350
2300
2250
2200
2150
12-03-13
12-04-13
12-05-15
12-07-15
12-09-11
12-09-18
12-10-21
12-10-14
Date

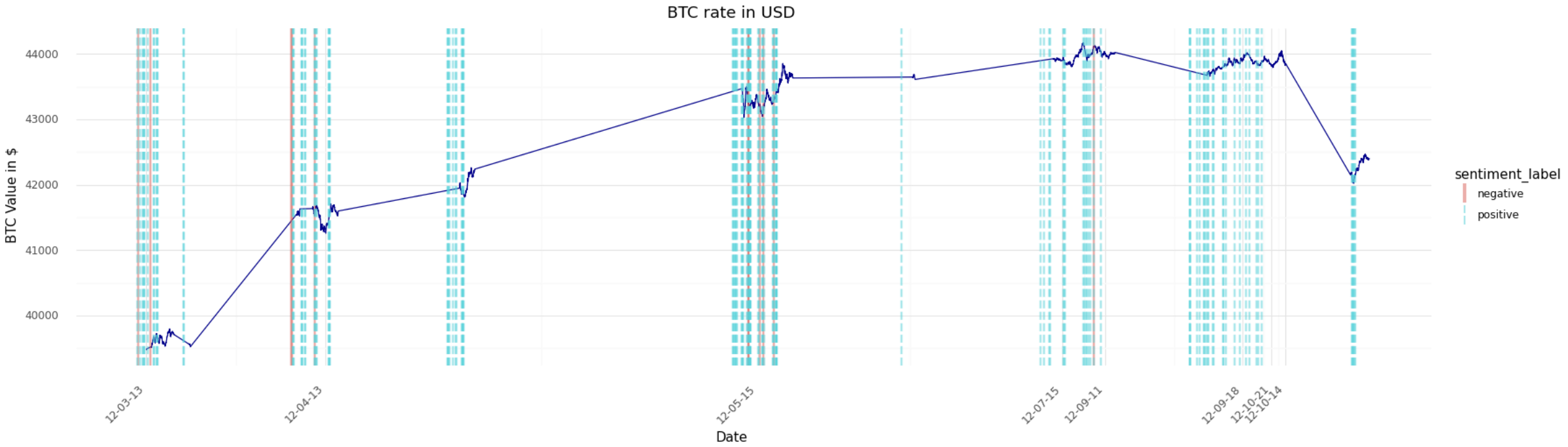

DOGE rate in USD
DOGE Value in $
0.100
0.095
0.090
0.085
12-03-13
12-04-13
12-05-15
12-07-15
12-09-11
12-09-18
12-10-21
12-10-14
Date

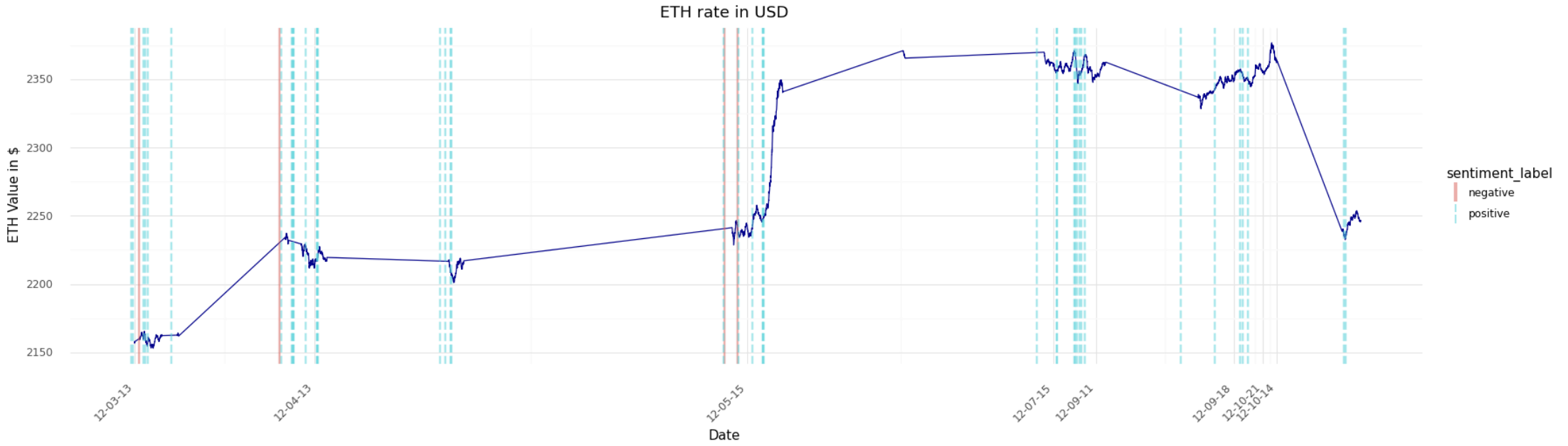
EDA – News data

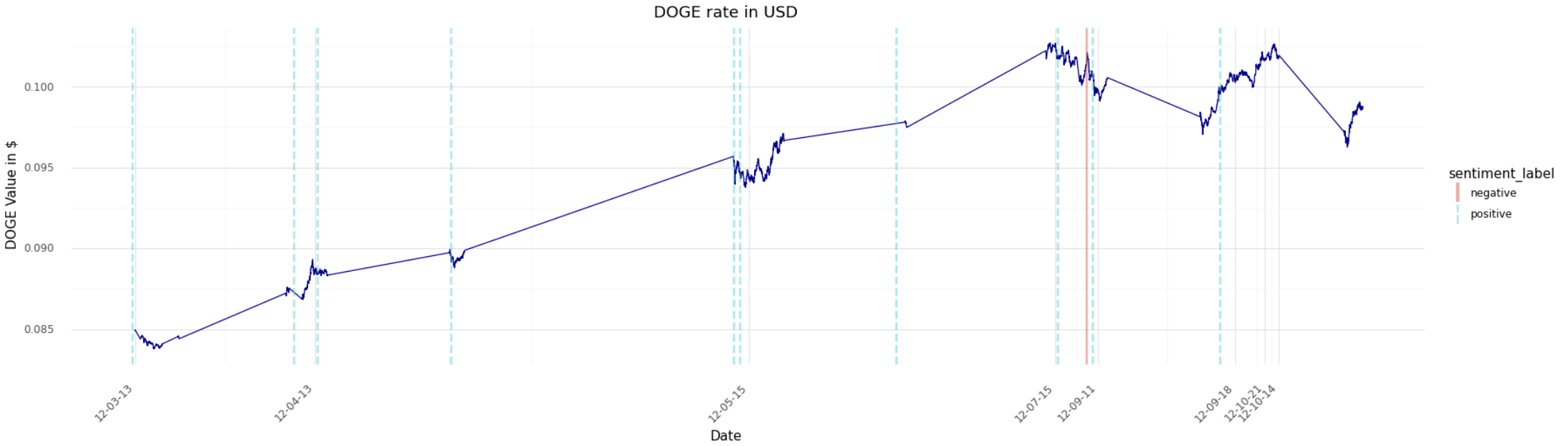
1. Datasources which used scrapper have low quality, and we have problems to make them usable.
2. Nevertheless, most of scrapped data comes from news.io which delivers high quality.
3. We have 508 English articles with full texts.
4. E.g. 192 of them mention Bitcoin, 63 Ethereum, and 16 Doge Coin

	article_id	title	link	keywords	creator	video_url	description	content	timestamp	image_url	source_id	source_priority	country	category	language
0	aff9633c7b572ed6a39f6e12f9f989d5	Exchange Analysts Reveal Four Sources of Bitco...	https://en.bitcoinsistemi.com/exchange-analyst...	[Analysis, Bitcoin, News]	[Mete Demiralp]	""	Analysts from cryptocurrency exchange Bitfinex...	Analysts from cryptocurrency exchange Bitfinex...	1701936425000	https://en.bitcoinsistemi.com/wp-content/uploa...	bitcoinsistemi	7453936	[turkey]	[top]	english
1	b40084ed2dadc26f1dc4aa35c4ee3bcd	Cold Supply Faces Challenges In Cost, Temperat...	https://www.businessworld.in/article/Cold-Supp...	""	""	""	Industry experts discuss the current state of ...	"Transporting or supplying food is not that di...	1701936317000	https://static.businessworld.in/article/articl...	businessworld	150658	[india]	[top]	english
2	fbfd495a6cea7b42a7e59d89dfd4e623	Fidel Castro's Sister Who Worked For The CIA D...	https://www.ibtimes.co.uk/fidel-castros-sister...	[World]	""	""	Juanita Castro lived in exile in Miami for dec...	"INTERNATIONAL BUSINESS TIMES uk NOTICEBOARD M...	1701935869000	""	ibtimes	458722	[united kingdom]	[top]	english
		Following UK					Robinhood's	Robinhood's long-							

EDA – News data




ETH rate in USD
ETH Value in $
2350
2300
2250
2200
2150
12-03-13
12-04-13
12-05-15
12-07-15
12-09-11
12-09-18
12-10-21
12-10-14
Date
sentiment_label
negative
positive


DOGE rate in USD
DOGE Value in $
0.100
0.095
0.090
0.085
12-03-13
12-04-13
12-05-15
12-07-15
12-09-11
12-09-18
12-10-21
12-10-14
Date
sentiment_label
negative
positive

Documentation

Big-Data-System-Cryptocurrencies

This repository contains results of the project during Big Data Analytics course at 2nd semester of Master's Degree Studies in the field of Data Science at Warsaw University of Technology (WUT). Our developer team consists of 4 students: Maciej Pawlikowski, Hubert Ruczyński, Bartosz Siński, and Adrian Stańdo.

The aim of the project is to deploy an end-to-end solution based on the Big Data analytics platforms.

Technological stack

In order to enable truely dsitributed computing our solution is heavily based on deploying each service in separate docker containers. With the usage of Tailscale, we designed a network that where the containers can communicate with each other. We aggregated various containers into separate subgroups, which share the same subnet. Such groups are described by the `COMPOSE_*`, where exists a `docker-compose.yaml` file which starts all containers described in the directory.

For now, our solution involves the following components:

- Docker,
- Tailscale,
- Portainer,
- Apache Hadoop 3.2.1 (namenode 2.0.0, java 8),
- Apache NiFi 1.23.2,
- Apache Kafka 3.4,
- Apache Spark 3.0.0,
- Apache HBase 2.2.6 (previously 1.2.6),
- Apache Hive 2.3.2 (metastore-postgresql 2.3.0),
- Apache Cassandra 4.0.11

How to run the project?

Prerequisites

You need Docker (Linux) or Docker-Desktop (Windows) installed.

You need WSL on Windows.

You need at least 10GB of unused RAM memory for the sytem.

You need Tailscale (Linux) or Tailscale add-in (Windows) installed.

You need proper tokens to the APIs mentioned somewhere in the main folders READMEs.

First-time set-up

Each folder that start with `COMPOSE_*` contains `docker-compose.yaml` file to run different parts of the system. Study `README.md` files in each directory to see whether additional environment variables have to be set.

Additionally you have to configure hive, during the first runs.

```
docker cp hive-server:/opt/hive/conf/hive-site.xml .
docker cp ./hive-site.xml spark-master:/spark/conf/
rm ./hive-site.xml
```

Starting containers

If the variables are set, you can run the following script to start all components on a singular machine:

```
./start.sh
```

Stopping containers

In order to stop all containers running on a singular machine execute:

```
./stop.sh
```

Big-Data-System-Cryptocurrencies

Public

Watch 2 Fork 0 Starred 1

hubert 4 branches 0 tags Go to file Add file Code

This branch is 10 commits ahead of main. Contribute

HubertR21 HBase -> Cassandra bd77127 yesterday 42 commits

COMPOSE_cassandra	HBase -> Cassandra	yesterday
COMPOSE_hbase-hive	HBase -> Cassandra	yesterday
COMPOSE_kafka-cluster	HBase -> Cassandra	yesterday
COMPOSE_nifi-hdfs	HBase -> Cassandra	yesterday
COMPOSE_spark	HBase -> Cassandra	yesterday
templates	add kafka	last week
.gitignore	update documentation	3 days ago
Encountered issues.md	HBase -> Cassandra	yesterday
README.md	HBase -> Cassandra	yesterday
network.sh	READMEs update	last week
start.sh	HBase -> Cassandra	yesterday
start_network.sh	READMEs update	last week
stop.sh	HBase -> Cassandra	yesterday

README.md

Big-Data-System-Cryptocurrencies

This repository contains results of the project during Big Data Analytics course at 2nd semester of Master's Degree Studies in the field of Data Science at Warsaw University of Technology (WUT). Our developer team consists of 4 students: Maciej Pawlikowski, Hubert Ruczyński, Bartosz Siński, and Adrian Stańdo.

About

This repository contains results of the project during Big Data Analytics course at WUT

Readme
Activity
1 star
2 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 3

adrianstando Adrian Stańdo
HubertR21 Hubert Ruczyński
bsinski Bartosz Siński

Languages

Python 85.3% Shell 7.8% Dockerfile 6.9%

Files

hubert

Go to file t

COMPOSE_cassandra
COMPOSE_hbase-hive
COMPOSE_kafka-cluster
COMPOSE_nifi-hdfs
COMPOSE_spark
templates
.gitignore
Encountered issues.md
README.md
network.sh
start.sh
start_network.sh
stop.sh

Enocuntered issues log

11.11.23

- Used Tailscale for distributed computing which generated lots of issues (tens of hours spent by the whole team).

18.11.23

- We designed the first version of docker-compose for NiFi and HDFS - only one of us knew how it worked, and it worked properly only on 1 PC.

25.11.23

- We had issues with adding a volume to HDFS and NiFi, so we don't lose the data from both. We had to find a workaround by copying files from NiFi to local, and only then mounting the volume. A few hours lost.

01.12.23

- We had major issues with connecting Kafka to anything, after 5 hours we finally found out that it has to be in the same network as HDFS and Nifi.

02.12.23

- Only then did we realize that for the containers to see objects in different subnets we have to attach them to static IPs, so we had to design a whole network structure from scratch, write a bash script which starts it up, and so on.. another few hours.

03.12.23

- During the development an update to NiFi occured, they changed the requirements, and we had to mount the logs file to the volume additionally. Another hour lost.
- We had a problem with running Jupyter lab to develop the code, as it didn't see anything. It turned out that we had to configure route tables for all containers so they see themselves in various subnets.