Cryptocurrency Price Prediction Based on Reddit Sentiment Analysis

Milestone 4

Laura Bąkała

Dominika Umiastowska

Mateusz Wójcik

Patryk Wrona

Purpose of the project

- Research of dependency on public opinion and cryptocurrency prices
- Potential financial benefits by investing in cryptocurrencies
- Creating guidelines for subreddit moderation so that they can prevent price manipulation through public opinion

Recap: Cryptocurrencies

Bitcoin

Ethereum

Dogecoin

Cardano

XRP

Solana

Recap: Real-Time Data Sources

Reddit API

- 6 subreddits, 1 per cryptocurrency
- 60 requests per minute
- We use PRAW, which keeps us under the limit and simulates a stream

Coinbase API

- Cryptocurrency prices in relation to US dollars
- 8 requests per second
- We make between 6 and 300 requests per minute (configurable depending on our needs)

Recap: Historical Data Sources

Reddit Historical Submissions Archives

- By using subreddits related to crypto, one could shrink data from 5TB to 500MB
- We will use data only from 2022 due to high costs of storing master data set (each month od reddit submissions ~ 100 GB)

CoinGecko API

- API allows acquiring information's about cryptocurrencies prices from past 9 years with daily interval and past 90 days with hourly interval

Sentiment Data

- **Stock-Market Sentiment Data**
  - Tweets regarding Economics
  - 2,106 negative and 3,685 positive observations
- **Stock Sentiment Analysis Dataset**
  - Tweets regarding stocks
  - 500 observations
- **Crypto Sentiment Analysis Dataset**
  - Reddit comments regarding cryptocurrencies
  - 562 observations

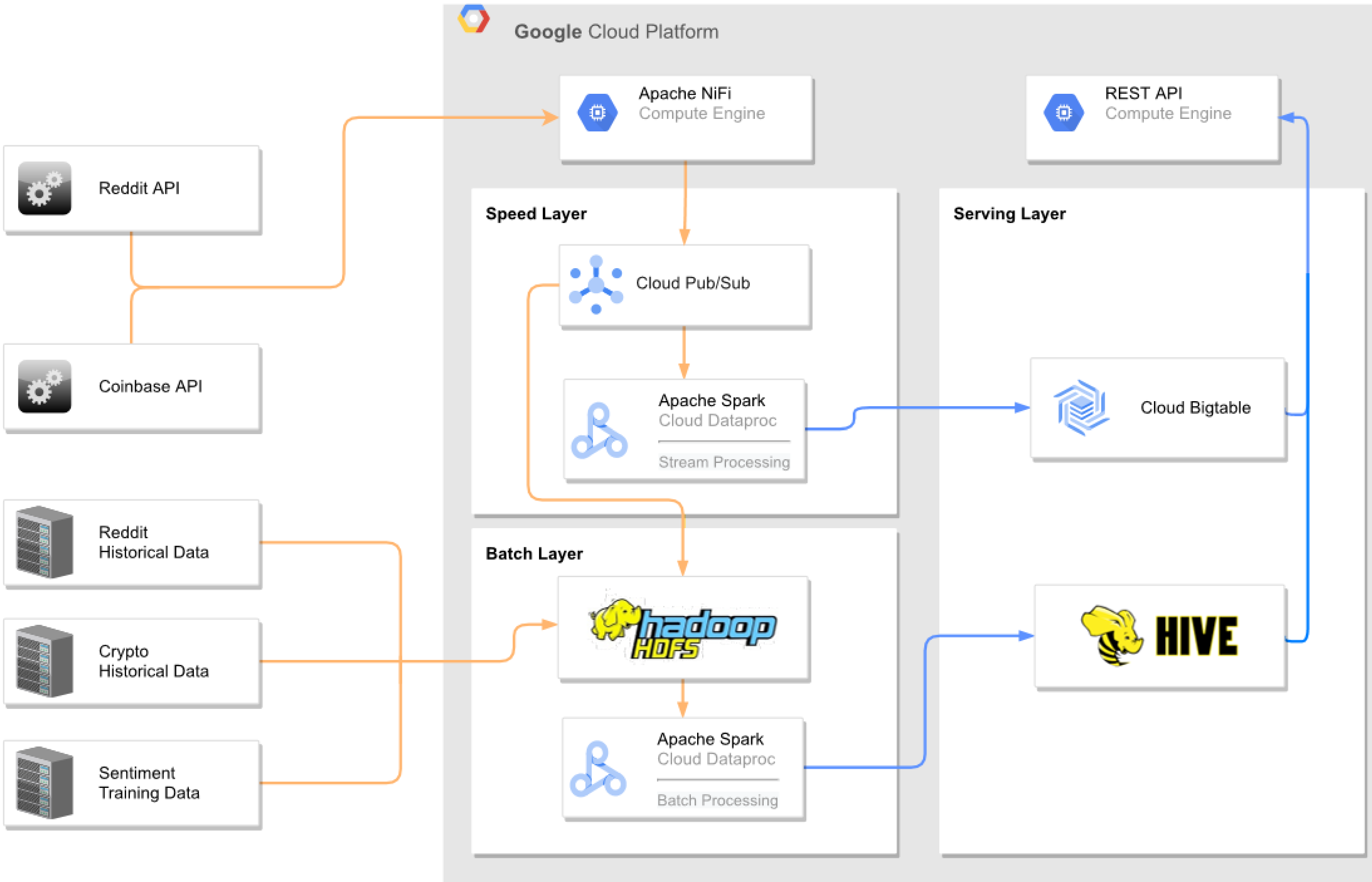
Processing historical data

Reddit data

- Text normalization
  - Tokenization
  - Lemmatization
  - Deleting stopwords and restricting to corpus
  - Lowering text
- Getting timestamp from date
- Limiting columns
- Unifying cryptocurrency ID
- Creating Hive tables

Crypto data

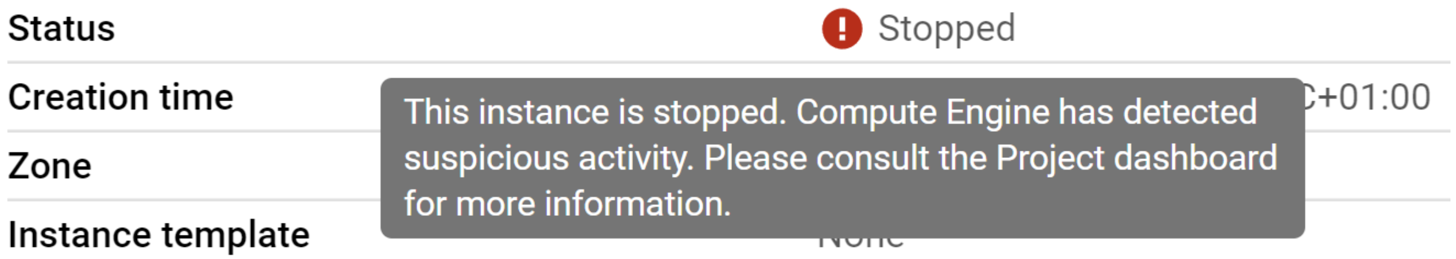
- Unifying cryptocurrency ID
- Getting timestamp from date
- Creating Hive tables



Google Cloud Platform
Apache NiFi
Compute Engine
REST API
Compute Engine
Reddit API
Coinbase API
Speed Layer
Cloud Pub/Sub
Apache Spark
Cloud Dataproc
Stream Processing
Serving Layer
Cloud Bigtable
Reddit
Historical Data
Crypto
Historical Data
Sentiment
Training Data
Batch Layer
hadoop
HDFS
HIVE
Apache Spark
Cloud Dataproc
Batch Processing

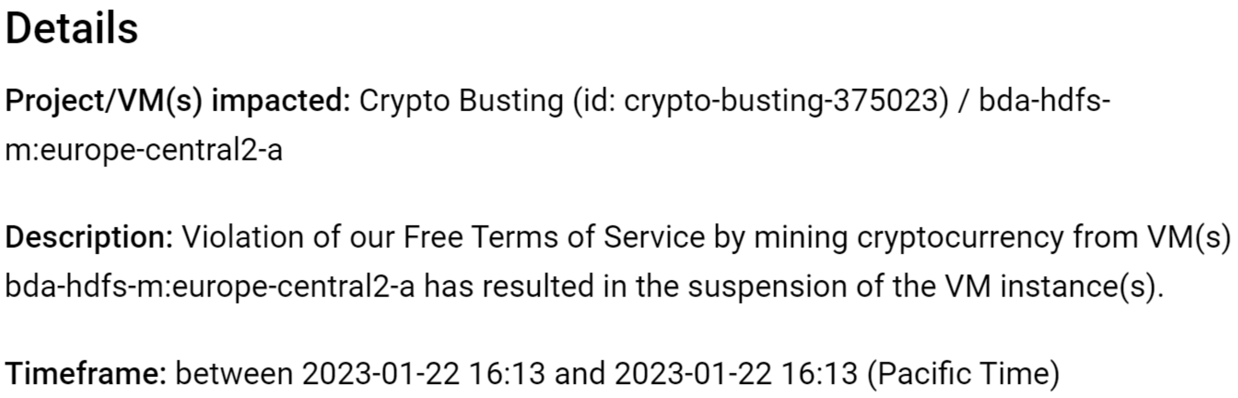
Google Cloud Problems

- Similar to previous stages, our cluster was stopped during the weekend due to the "suspicious activity" connected to cryptocurrency mining. This is why, some of the components and flows will be described in words.



Status	❗ Stopped
Creation time	C+01:00
Zone	
Instance template	None

This instance is stopped. Compute Engine has detected suspicious activity. Please consult the Project dashboard for more information.



Details

Project/VM(s) impacted: Crypto Busting (id: crypto-busting-375023) / bda-hdfs-m:europe-central2-a

Description: Violation of our Free Terms of Service by mining cryptocurrency from VM(s) bda-hdfs-m:europe-central2-a has resulted in the suspension of the VM instance(s).

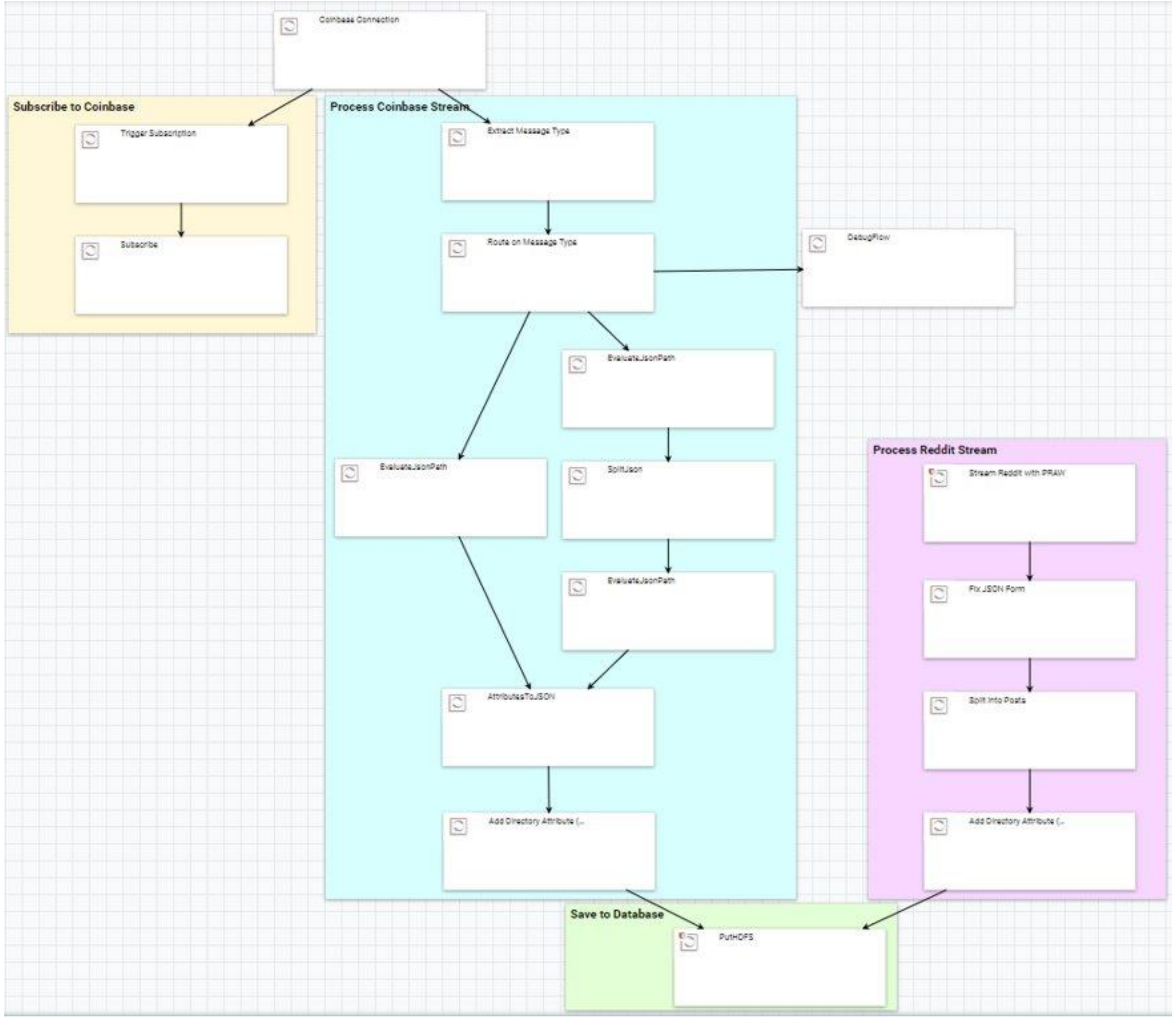
Timeframe: between 2023-01-22 16:13 and 2023-01-22 16:13 (Pacific Time)

NiFi Data Flow

Updates compared to the previous version (image unavailable):

- Error handling implemented for two key processors that route Flow Files based on their content type:
  If type is "error" (e.g., for missing product ID like "XRP-USD"), then:
  - error message is extracted from response,
  - error context is added,
  - the Flow File is sent to a processor that raises a warning,
  - and, subsequently, to a DebugFlow.
- *PublishToGCPPubSub* processor added to both Coinbase and Reddit flow to publish messages to the corresponding topics

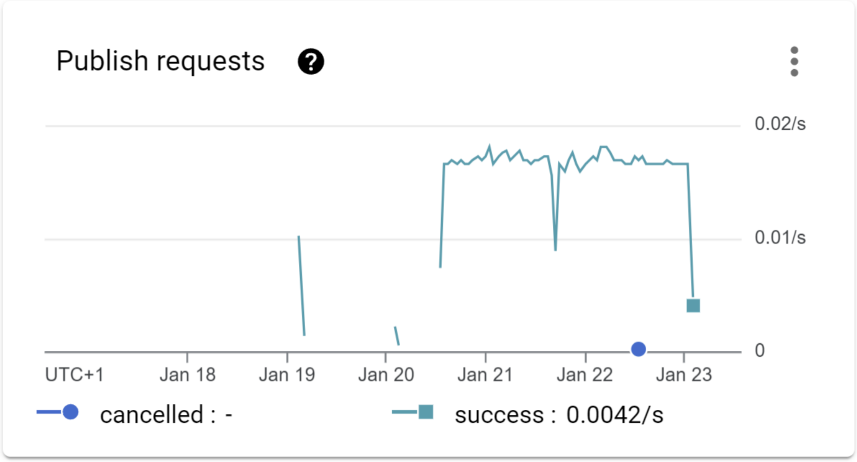
The old Flow

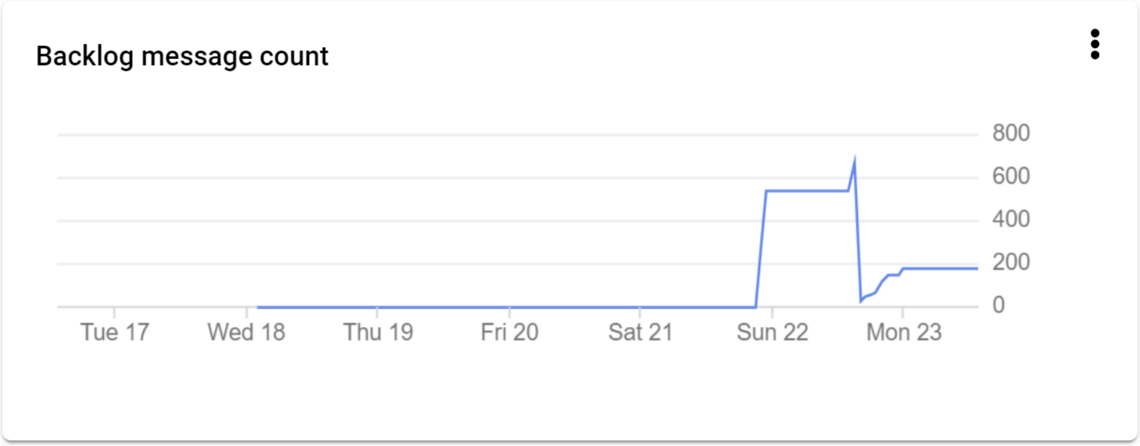

Coinbase Connection
Subscribe to Coinbase
Trigger Subscription
Subscribe
Process Coinbase Stream
Extract Message Type
Route on Message Type
DebugFlow
EvaluateJsonPath
EvaluateJsonPath
SplitJson
EvaluateJsonPath
AttributesToJSON
Add Directory Attribute (...
Process Reddit Stream
Stream Reddit with PRAW
Fix JSON Form
Split into Posts
Add Directory Attribute (...
Save to Database
PutHDFS

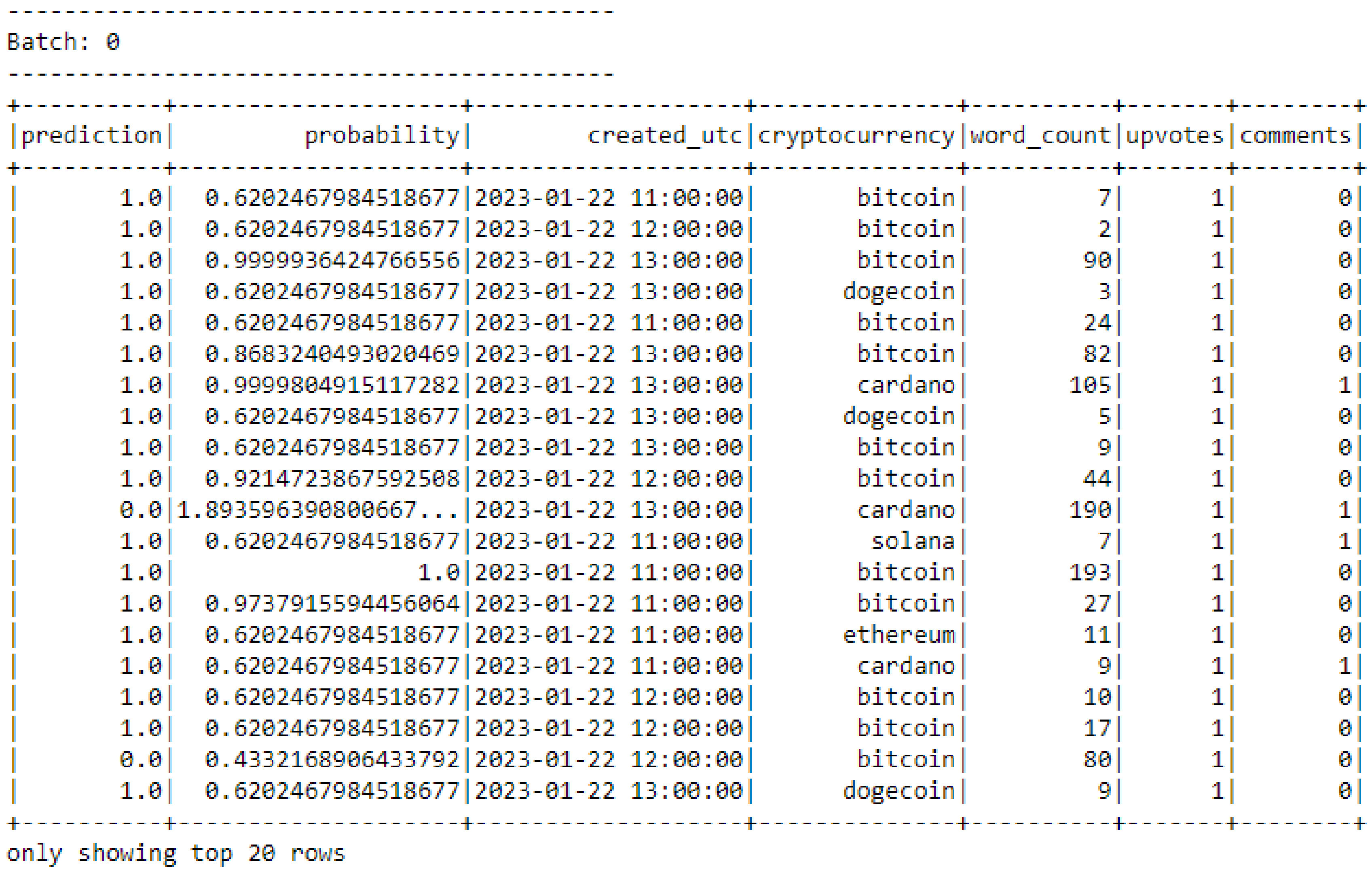
Pub/Sub for Data Streams

(speed layer)

- **The use of Pub/Sub for ingesting real-time data:**
  - **The already mentioned NiFi processor publishes messages to Pub/Sub topic.**
  - **A Python script subscribes the topic and retrieves messages every 30 seconds to publish them to the corresponding Pub/Sub Lite topic due to the compatibility with PySpark**
  - **Next, PySpark Structured Streaming is used to read the stream and process obtained messages**
    - **In case of Coinbase data, the records are prepared to show the current prices**
    - **In case of Reddit data, there is a word count, sentiment prediction and number of upvotes**





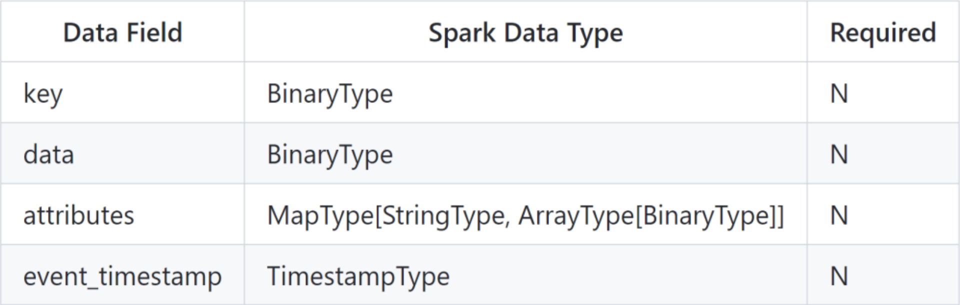


```
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+--------------------+-------------------+--------------+----------+-------+--------+
|prediction|         probability|        created_utc|cryptocurrency|word_count|upvotes|comments|
+----------+--------------------+-------------------+--------------+----------+-------+--------+
|       1.0|  0.6202467984518677|2023-01-22 11:00:00|       bitcoin|         7|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 12:00:00|       bitcoin|         2|      1|       0|
|       1.0|  0.9999936424766556|2023-01-22 13:00:00|       bitcoin|        90|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 13:00:00|      dogecoin|         3|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 11:00:00|       bitcoin|        24|      1|       0|
|       1.0|  0.8683240493020469|2023-01-22 13:00:00|       bitcoin|        82|      1|       0|
|       1.0|  0.9999804915117282|2023-01-22 13:00:00|       cardano|       105|      1|       1|
|       1.0|  0.6202467984518677|2023-01-22 13:00:00|      dogecoin|         5|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 13:00:00|       bitcoin|         9|      1|       0|
|       1.0|  0.9214723867592508|2023-01-22 12:00:00|       bitcoin|        44|      1|       0|
|       0.0|1.893596390800667...|2023-01-22 13:00:00|       cardano|       190|      1|       1|
|       1.0|  0.6202467984518677|2023-01-22 11:00:00|        solana|         7|      1|       1|
|       1.0|                 1.0|2023-01-22 11:00:00|       bitcoin|       193|      1|       0|
|       1.0|  0.9737915594456064|2023-01-22 11:00:00|       bitcoin|        27|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 11:00:00|      ethereum|        11|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 11:00:00|       cardano|         9|      1|       1|
|       1.0|  0.6202467984518677|2023-01-22 12:00:00|       bitcoin|        10|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 12:00:00|       bitcoin|        17|      1|       0|
|       0.0|  0.4332168906433792|2023-01-22 12:00:00|       bitcoin|        80|      1|       0|
|       1.0|  0.6202467984518677|2023-01-22 13:00:00|      dogecoin|         9|      1|       0|
+----------+--------------------+-------------------+--------------+----------+-------+--------+
only showing top 20 rows
```

Reddit stream processed using Spark Structured Streaming

Pub/Sub Lite for BigTable

- All the messages are transformed to the Pub/Sub Lite message format and published to other topics for processed data.



Data Field	Spark Data Type	Required
key	BinaryType	N
data	BinaryType	N
attributes	MapType[StringType, ArrayType[BinaryType]]	N
event_timestamp	TimestampType	N

- Because the BigTable instance may not be available, the messages are published to Pub/Sub Lite.
- PySpark Structured Streaming does not support putting data directly to BigTable without any workarounds
- We prepared Python scripts that appends BigTable table via Google Cloud BigTable API. This would be the endpoint for the client application.

Overview of machine learning models

Machine learning models:

1) sentiment classification – trained on sentiment data, will be used in the future to periodically predict sentiment of reddit data, output: *(reddit_sentiment)*

- Trained & tested on: sentiment data
- Used for predicting: reddit sentiment (historical + real-time)

2) price prediction – takes into account past trends in cryptocurrency prices and senitment from *reddit_sentiment*

- Trained & tested on: crypto data (historical + real-time): once with once without an aggregated sentiment.
- Used for predicting: future crypto prices (project's main goal)

Sentiment Classification

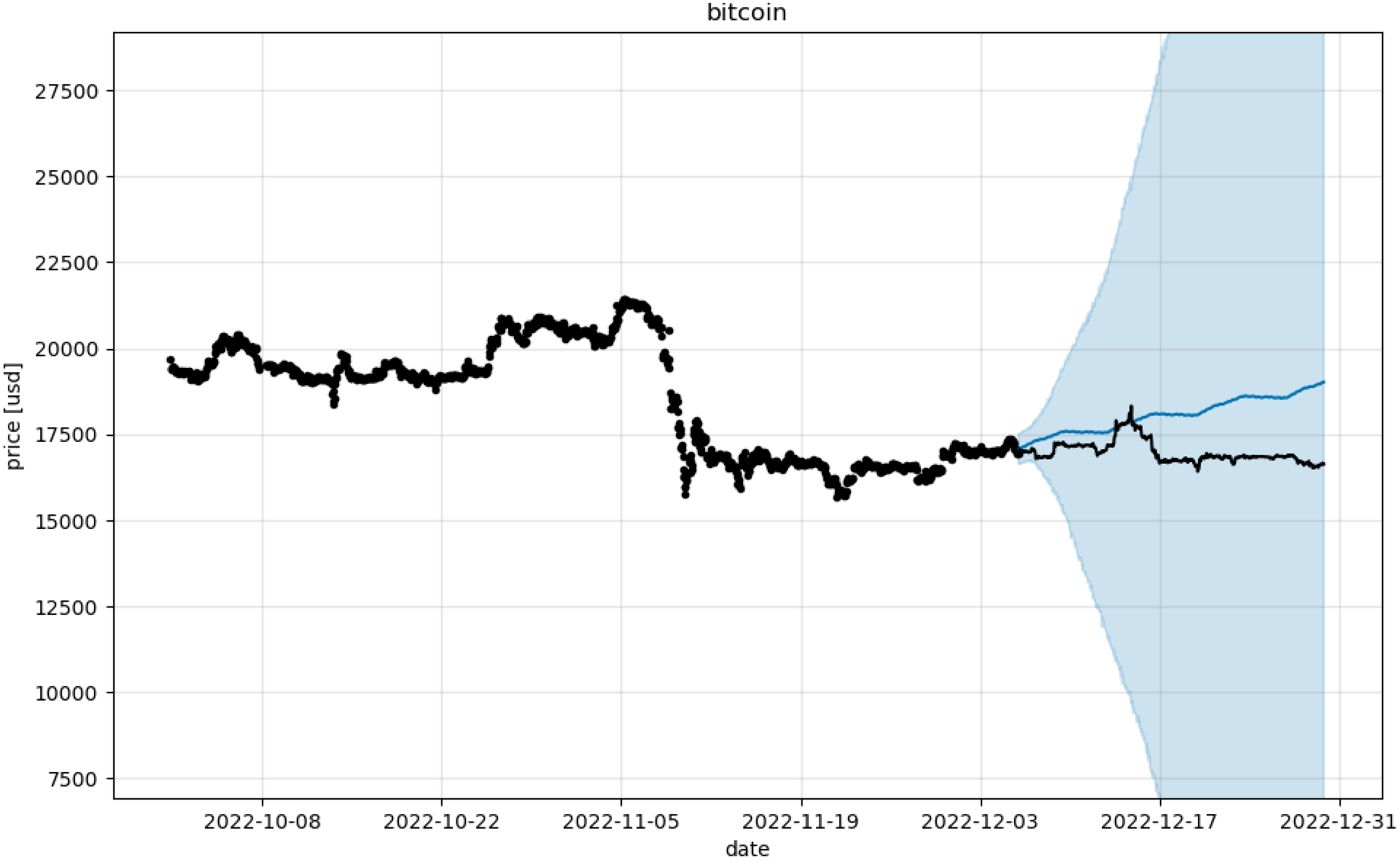
- Normalization of the text
- Embedding via string hashing
- Logistic Regression model
- Splitting the dataset on train and test 7:3
- Accuracy: **74%**

Cryptocurrency Price Prediction

- Model based only on past cryptocurrencies' prices
- **Prophet** model
  - Created for time series forecasting
  - It may be multivariate
  - Optional seasonality and different options
- We were evaluating model by training it on **hourly** data form 65 days and testing on 25 following days for each cryptocurrency.

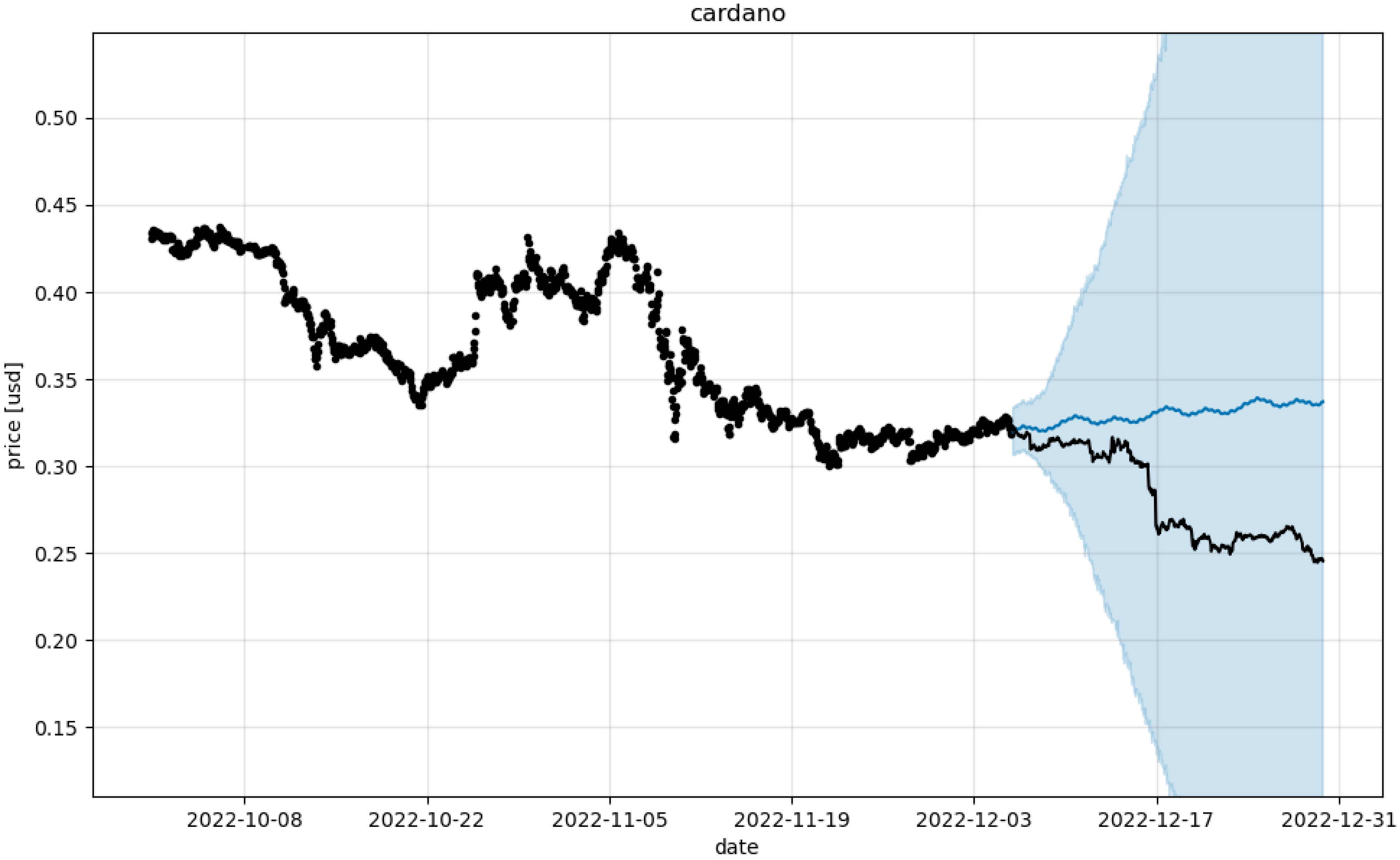
MSE	291346.41
MAE	218.76
R^2	0.99

Bitcoin



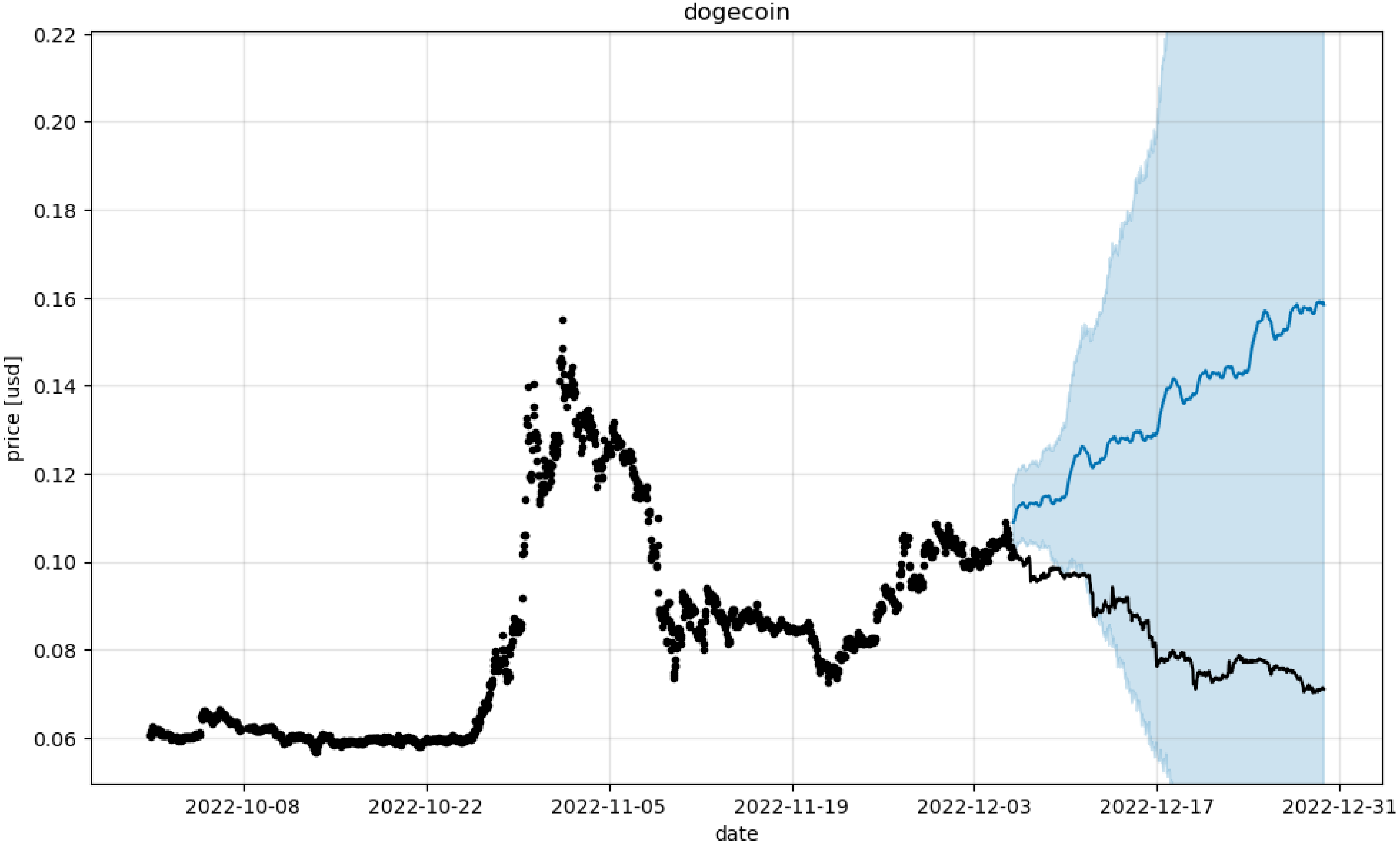
MSE	1693117.12
MAE	1096.00
R²	-15.21

Cardano



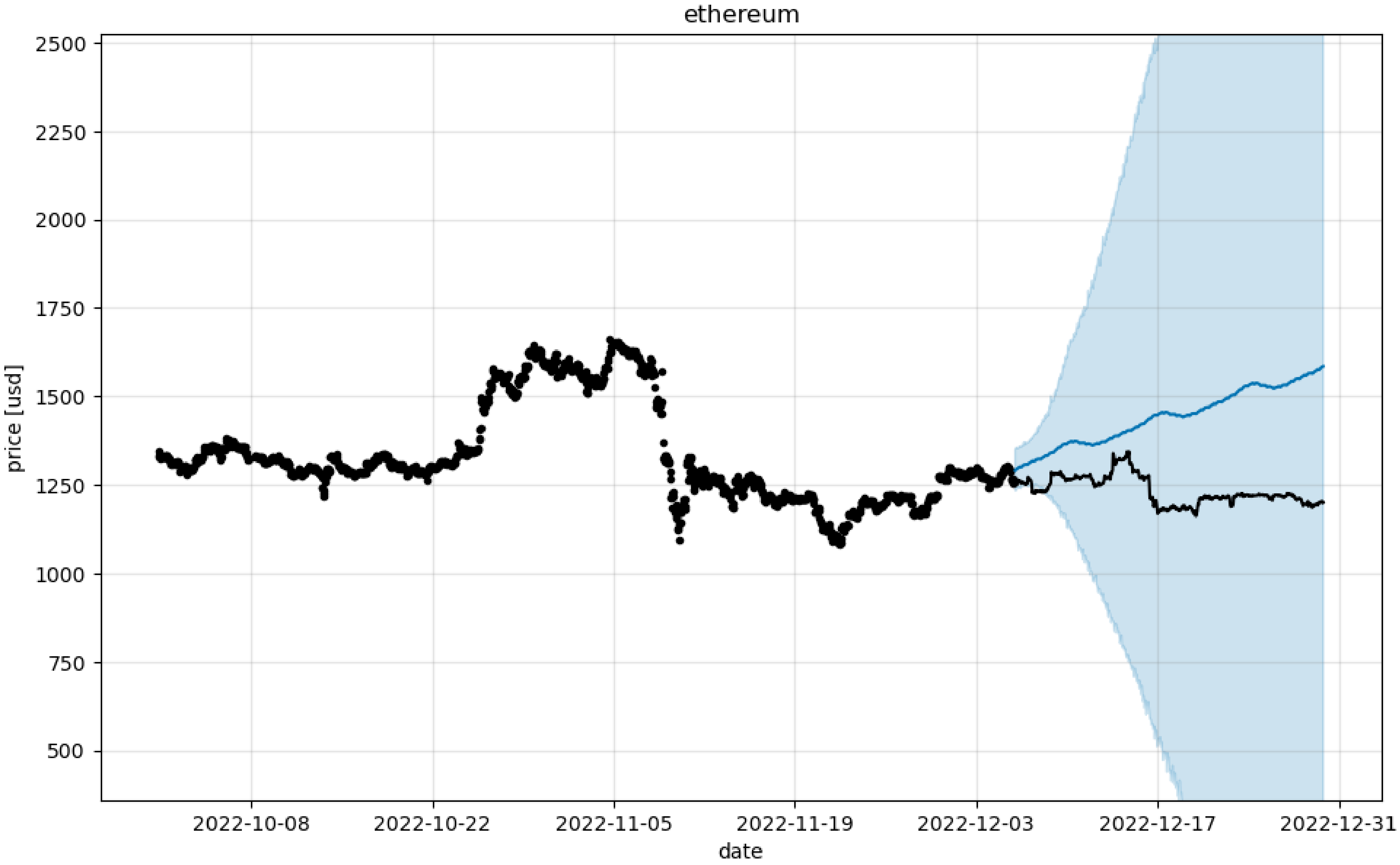
MSE	0.0032
MAE	0.048
R^2	-3.71

Dogecoin



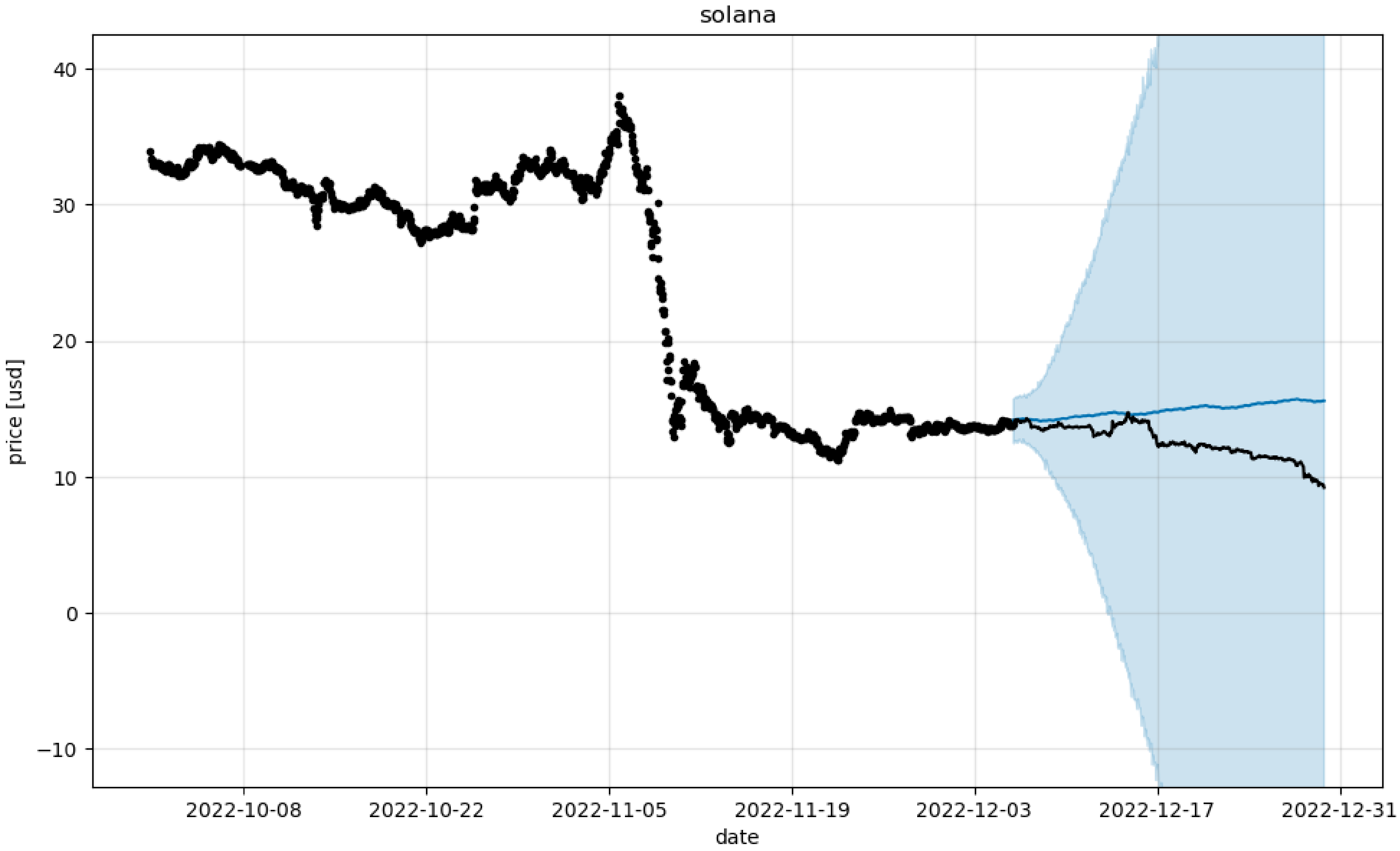
MSE	0.0032
MAE	0.051
R²	-32.21

Ethereum



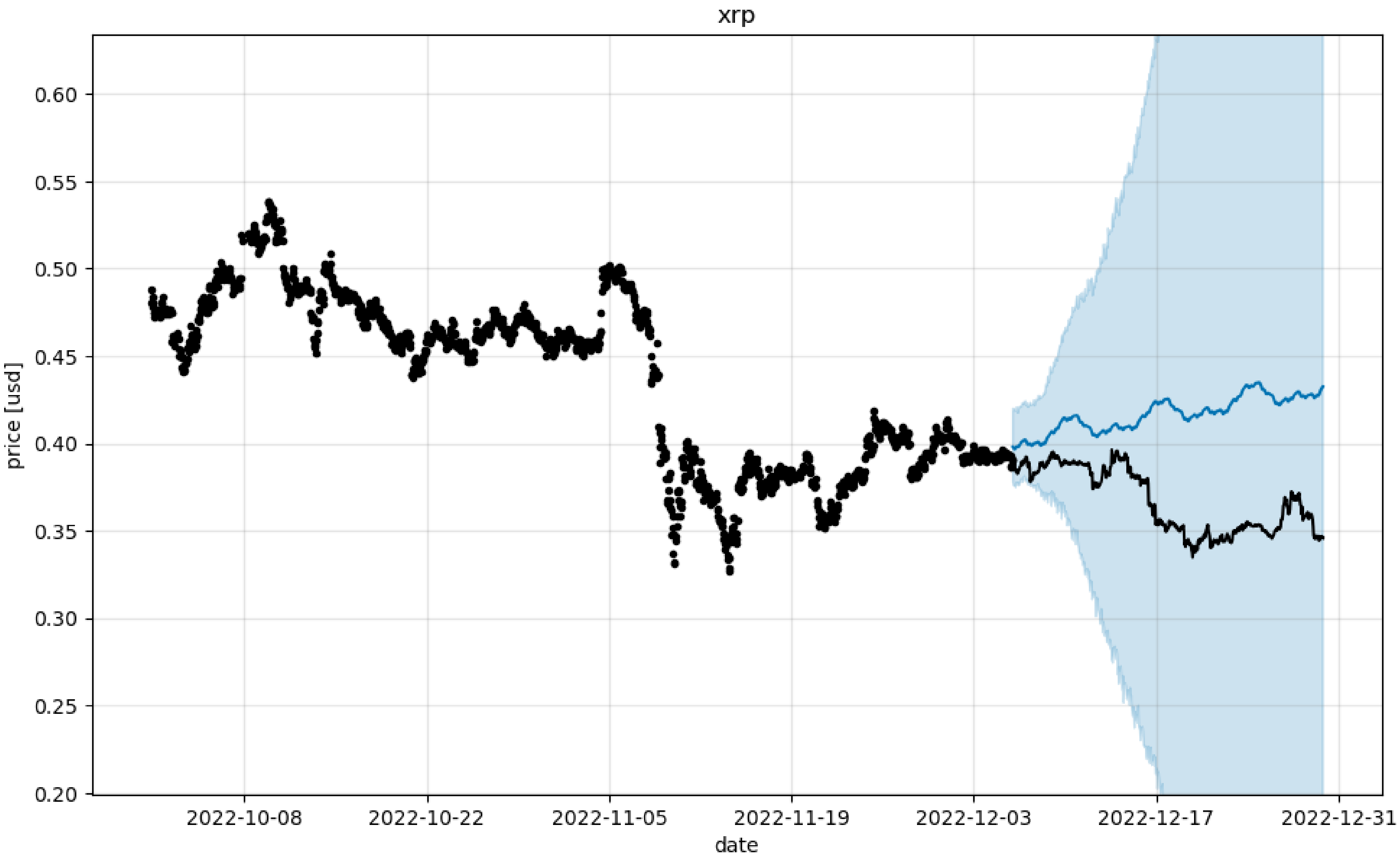
MSE	54953.07
MAE	208.08
R²	-34.21

Solana



MSE	8.18
MAE	2.31
R^2	-4.28

XRP



MSE	00.0032
MAE	0.05
R²	-8.5

Model based on both sentiment and cryptocurrency

- Prophet as before but with added sentiment regressor now
- Both model and regressor are multiplicative
- Depending on type of trend
  - Linear did not show much different results from before but in general seemed to predict the trend better/more naturally
  - Flat differed a lot depending on sentiment but in very chaotic way which did not seem very natural
- In general, similar overall results (MSE, MAE) to before
- Probably just not the best model

Predicting future prices

It demands following steps

1. Predicting sentiment based on trained beforehand model on available Reddit data
2. Changing all timestamps for Crypto and Reddit data by rounding it to the full hour.
3. Aggregating Reddit sentiment results to one per date by prediction (also tried probability)
4. Shifting dates from reddit by N days to future
5. Aggregating columns by date and cryptocurrency (crypto based)
6. If no price is available for given hour it is taken from neighboring rows and repeated
7. It is saved to HDFS as training data for crypto-price model
8. Since we shifted the data forward there will be some "future sentiment", or rather based on point of view "sentiment from N days ago".
9. Based on this sentiment and date we will predict future crypto prices. It is saved separately as "testing" data.

Results - quality of the solution from a business perspective

1) Reviewing the sentiment and popularity of particular cryptocurrencies as the public opinion influences the cryptocurrency prices

2) Prediction of the cryptocurrency prices based on a fixed window of time

3) Possibility of changing of business model thanks to saved raw data in master data set

4) Thanks to Pub/Sub, the solution will adapt to increasing inflow of data streams

5) The solution could be scaled horizontally, involving more worker nodes, besides, used BigTable NoSQL database assures reliable and consistent batch views storage.

6) Thanks to GPC security policies (e.g. IAM), the securiy of user's data is kept

7) The data is also replicated in Google Storage to enable fast recovery

Potential impact of the solution on the society

1) Possibility of verifying the influence of public opinion on financial instruments, possible cascade of changes? (if the opinion is already bad, it will get worse)

2) Proving the high variance and unpredictability in cryptocurrency prices to describe the risks of cryptocurrency transactions, so that risk-takers could invest in cryptocurrency (**Solana** and **dogecoin** revealed to be the most changing cryptocurrencies)

3) In case of relying on this model, each investment would impact the real price of the cryptocurrency and its popularity will increase, so probably the decreasing tendency from 2022 will stop

4) The solution provides an easy way of monitoring the cryptocurrency current prices, and could be extended to investing modules, so the automation of investments could be increased, which would save peoples' time.

Thank you for your attention!

Should you have any questions, please ask them now ☺