Bitcoin price analysis with integration of social media and news data

Data Management - Q1 (2023)

Group 25

Full Name	Student ID
Davide Ghilardi	857384
Lorenzo Lobosco	851289
Fabio Salerno	861419

Milan, June 12, 2023

Contents

1 Introduction **1**
- 1.1 Problem statement . . . 1
- 1.2 Aim of the project . . . 1

2 Data Acquisition **2**
- 2.1 Bitcoin prices data . . . 2
- 2.2 Google Trends . . . 2
- 2.3 Reddit posts data . . . 3
- 2.4 Articles web scraping . . . 3
- 2.5 Other sources . . . 4

3 Pre-processing and Integration **5**
- 3.1 Bitcoin Data . . . 5
- 3.2 Reddit . . . 5
- 3.3 Articles . . . 6
- 3.4 Google Trend . . . 6
- 3.5 Sentiment analysis . . . 6

4 Data quality **8**
- 4.1 Completeness . . . 8
- 4.2 Temporal Dimension . . . 9

5 Data storage **10**
- 5.1 Modelling . . . 10
- 5.2 Daily pipeline . . . 12

6 Fear and Greed Index **13**
- 6.1 Index calculation . . . 13
- 6.2 Results . . . 14

7 Conclusions and further developments **16**

8 References **18**

1 | Introduction

1.1 | Problem statement

In this digital era, cryptocurrencies have emerged as a revolutionary concept, reshaping the way we perceive and interact with money. At the forefront of this financial revolution is Bitcoin, the pioneering and most widely recognized cryptocurrency to date. But what exactly is Bitcoin? In simple terms, according to Forbes [3], "Bitcoin is a decentralized digital currency that you can buy, sell and exchange directly, without an intermediary like a bank". The backbone of Bitcoin is its innovative technology called blockchain, which guarantees transparency, security, and the permanence of transactions. Bitcoin is fundamental in the world of cryptocurrency, it serves as a store of value, medium of exchange and unit of account, all contained in a single digital asset. Bitcoin has captured the imagination of enthusiasts and investors, sparking a global movement toward digital currencies. One of the fascinating aspects of Bitcoin is its volatility and the factors that influence its price movements. Bitcoin is not backed by a central bank or government, so its value is determined by various factors such as supply and demand dynamics, market sentiment, regulatory developments, macroeconomic indicators, and technological advancements. In this context, the role of social networks, news, and media becomes crucial. Social networks provide platforms for discussions, debates, and the sharing of insights among the Bitcoin community. They can influence market sentiment and create waves of enthusiasm or panic. News outlets and media coverage play a significant role in shaping public perception and investor sentiment towards Bitcoin.

1.2 | Aim of the project

Cryptocurrency markets are influenced by a combination of factors, including both rational and emotional elements. While traditional financial markets are driven by various economic indicators, company performance, and other fundamental factors, cryptocurrency markets, including Bitcoin, tend to be more susceptible to emotional and speculative behavior. So especially for crypto markets, assessing the market sentiment become really important for better support the investment decisions. A possible method to do it is through the Alternative.me Crypto Fear and Greed Index, a tool that attempts to measure and quantify the overall sentiment or emotional state of the market participants towards cryptocurrencies. According its website, it is built by combining data coming from many sources such as financial calculations, social sentiment, google trends and online polls. This index ranges from 0 to 100, with higher values indicating extreme greed and lower values indicating extreme fear [1]. The aim of the project is to collect, combine, process and store different Bitcoin information coming from multiple data sources such as brokers, social media, newspapers, and search engines, to build a unified database that can serve as a platform to create a quantitative measure of Bitcoin market sentiment similar to the one developed by Alternative.me. Finally, to help traders and investors gauge the potential risks and opportunities in the market and to support them in the decision making process, it was decided to build a pipeline that every day collects data from the different sources and outputs the value of the index, ultimately providing a snapshot of the market sentiment. To achieve this result, it was initially decided to work on a historical data sample that served to study and structure the process of data collection, cleaning and storage. The time window of 2022 was selected as subject of study for the implementation of the process that would later be used to collect data on a daily basis and produce the fear and greed index.

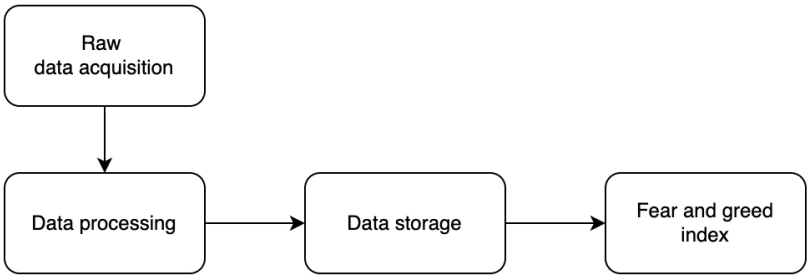

Figure 1.1: Summary data workflow.

2 | Data Acquisition

For the purposes of the project, data was collected from many sources and through different techniques. Sources can be divided in the following way:

- Bitcoin financial time series data (via API)
- Reddit posts data (via API)
- Articles related to Bitcoin (via Web scraping)
- Google trends (via API)
- Other sources (downloaded)

2.1 | Bitcoin prices data

The financial data related to Bitcoin were extracted through API requests. API (Application Programming Interface) is a set of protocols and tools that enable two software applications to communicate with each other. Binance API provides access to the exchange's data. Through Binance API, we can access real-time trading information and also retrieve historical data, including price, volume, and trade history, for various time intervals and for any crypto traded on the exchange. To acquire data through Binance API requests, an account on the Binance exchange is firstly needed to create an API key. With the API key, Python programming language was used to make API requests and retrieve the data needed. In particular the data was extracted with the command `.get_historical_klines()` by setting the symbol token (in this case BTCUSDT), the time interval of interest (in this case for the purposes of the project to make all the calculation needed the time period of data was from 1 September 2021 to 1 June 2023) and frequency of the data (in this case daily).
The dataset contains 639 records, and has 12 columns, the feature available are:

- Open time
- Open
- High
- Low
- Close
- Volume
- Close time
- qav (quote asset value)
- num_trades (number of trades)
- taker_base_vol (taker buy base asset volume)
- taker_quote_vol (taker buy quote asset volume)
- ignore (This variable is present in the klines command but is a deprecated feature, so that specific column was dropped.)

For the purposes of the project and to save space on disk, it was decided to keep only the following columns: Open, High, Low, Close, Volume, and Close time.

2.2 | Google Trends

To analyze trends in searches related to the term 'Bitcoin' in the year 2022, the Google Trends API was used along with the Python pytrends library. Google Trends API provides access to a vast amount of search data, allowing the exploration of popularity and interest in various topics over time. The goal of this analysis was to gain a deeper understanding of the interest and fluctuations of Bitcoin-related searches throughout time. The pytrends library in Python acted as a convenient interface to interact with the Google Trends API, enabling the retrieval of search data for 'Bitcoin' keyword, firstly in the time range of 2022, and then for present data.

2.3 | Reddit posts data

Text-based social networks like Twitter and Reddit are online platforms that prioritize the sharing and exchange of text-based content. These platforms provide spaces for users to express their thoughts, engage in discussions, and connect with others based on shared interests. Considering the daily users volume of these platforms, they play a very important role in making worldwide connections and instantly spreading information worldwide. In particular, for this project it was decided to focus only on the data coming from Reddit, and the main reason is due the fact that from February 9 2023 Twitter started charging non-basic access to its API. Reddit is a forum-style platform organized into various communities called "subreddits". Each subreddit focuses on a specific topic, interest, or theme, and it's a place where users can join and participate in these communities, posting text-based submissions or engaging in discussions through comments. Reddit encourages longer-form content and fosters in-depth conversations on a wide range of subjects. The platform's upvoting and downvoting system allows users to collectively curate and moderate content, ensuring that the most valuable contributions rise to the top. The text data sourced from Reddit posts was extracted from the three primary Bitcoin-related subreddits:

- **r/Bitcoin**, with a membership of 5 million users.
- **r/btc**
- **r/BitcoinBeginners**, also with a membership of 1.1 million users.

There are two primary methods for retrieving data from Reddit: the Reddit API and the Pushshift API and for this project both of them were used. The Reddit API allows users to fetch a limited amount of recent comments or submissions from different streams within a subreddit, such as hot, new, top, and more. However, it is not ideal for creating large datasets due to its limitations in retrieving historical data. On the other hand, Pushshift is a service that collects new comments and submissions from Reddit, stores them in a database, and provides an API endpoint for querying the data. It offers an alternative solution for accessing Reddit data beyond the limitations of the native Reddit API, and it's optimal to retrieve historical data. It's important to note a few drawbacks of Pushshift, which, as a third-party service maintained by a single individual, may incur in occasional instances of downtime when comments and submissions are not ingested or stored. While these gaps are sometimes backfilled with data, it is not always guaranteed. Additionally, since not proprietary, the data provided by Pushshift is not real-time, and the delay in capturing comments or submissions can be up to 2 days according to the documentation. For the purposes of the project the best option for retrieving historical data from Reddit was through Pushshift, while for collecting the posts with the daily request, Reddit API was preferred.
The Reddit posts were extracted from each subreddit and in particular considering the aim of the project they were extracted for each post the following features:

- subreddit: The belonged subreddit name of the post.
- id: Unique identifier for each post given by Reddit.
- author: The nickname of the post.
- url: The link to the Reddit post.
- title: The title of the post.
- selftext: The text contained in the post.
- utc_datetime_str: The datetime version of the timestamp.

2.4 | Articles web scraping

To understand the impact of bitcoin news on its price, two important websites were considered: CoinTelegraph and CoinDesk. Cointelegraph is an independent digital media resource covering a wide range of news on blockchain technology, crypto assets, and emerging fintech trends. The focus of our research was bitcoin news which in the site have a dedicated section. Cointelegraph bitcoin section has all the articles listed one under the other with the most recent to the top. Since the page doesn't load all the articles all at once but it loads them in chunks while scrolling down, the need of browser automation was clear. It was decided to use Selenium to automatically scroll down the page and load articles until

the desired date was reached. The implementation in python was adapted from the GitHub repository dub-basu/cointelegraph-scraper. The same procedure was followed for CoinDesk, an integrated platform for media, events, data and indices for what concerns the blockchain. Also CoinDesk offers a dedicated section for bitcoin news that has a simpler structure than CoinTelegraph's one. There, articles are disposed in separated sections accessible by a navigation bar at the bottom of each page, whose url is simply the bitcoin news page url plus the number of news section. The CoinDesk scraper was adapted from the one for CoinTelegraph and didn't need browser automation. Both the two scrapers have similar structure which resembles the original one in the repository. They work in two steps:

1. The first step is dedicated to scrape the bitcoin news homepage to get all the urls of all articles of interest.
2. The second step goes through each article (by the url), fetches all the useful information about it, and saves the results in a json table format.

The step that differs the most from the two websites is the first since it requires browser automation for CoinTelegraph, while only an iteration through sections for CoinDesk. Finally, information extracted from the articles was about: url, author, date, title, description, and content.

2.5 | Other sources

For the construction of the index mentioned before and better discussed later in the index section, some data other than the one previously mentioned was required. The first is the bitcoin dominance, which is the ratio between the market capitalization (market cap) of Bitcoin and the market cap of the entire cryptocurrency market. While present dominance value can be easily accessed by making a request to CoinGecko API, historical data of the Bitcoin dominance is more difficult to obtain. The best solution was exploiting two datasets downloadable on the internet:

- Bitcoin market capitalization was downloaded from this link, and it is a daily time series starting from 2009 and ending May 29 2023.
- Overall crypto market capitalization was downloaded from this link, and it is a weekly time series. Considering the fact that we needed the daily Bitcoin dominance for all 2022 it was decided to proxy the weekly data for all the days in the week.

With this information, the dominance time-series can be computed as the ratio of the two.

3 | Pre-processing and Integration

To ensure the highest quality and suitability of the data for the aim of the project, the raw data obtained from various sources underwent a pre-processing phase. Each source was treated individually, with dedicated attention given to refining the data to better fit in the database system and enhancing its overall quality. All collected data were based on certain time instants, some sources like the articles or Google trends data had the standard readable date format (year, month, day); while the data given by the API where in the form of epoch timestamps, also known as Unix timestamp or POSIX time, is a representation of time as the number of seconds that have elapsed since a specific reference point called the "epoch." The epoch is commonly set as January 1, 1970, at 00:00:00 Coordinated Universal Time (UTC). Epoch timestamps have several advantages in computer systems and programming because they provide a simple and standardized way to represent time. In order to integrate those different data sources all together, it was crucial to format and standardize all the timestamps coming from different sources. During the pre-processing phase of data sources that contain text, particularly Reddit articles and posts, we applied a text cleanup, removing elements not needed for our analysis, such as links, emoticons, and many others. By cleaning the data of these elements, we wanted to ensure that the text used for our analysis was consistent, standardized, and free of unnecessary noise. By doing this, the quality of the text increased significantly, and allowed to work more easily using these sources. Subsequently, we conducted a sentiment analysis on the processed data. This analysis involved determining the sentiment or emotional tone expressed in the text, in particular whether it was positive, negative, or neutral. By performing sentiment analysis, we aimed to gain a deeper understanding of the overall sentiment and attitudes reflected in the articles and Reddit posts, providing valuable insights into the perception and opinions surrounding the subject of Bitcoin.

3.1 | Bitcoin Data

The raw data coming from the Binance API in json format were first of all loaded into a tabular format. The `close_time` variable was converted from the POSIX timestamp format to the readable UTC date format (year-month-day) and the name was converted to "date". Additionally to the Bitcoin price time series it was decided to enrich the table with dominance value. The dominance was computed by dividing the Bitcoin market cap with the overall crypto market cap. The two data coming from different sources were joined together through their date, the daily dominance was then computed and finally merged with the Bitcoin price time series dataset through a full outer join. This procedure avoided the creation of another table just for the dominance calculation. One downside of this operation was the introduction of NULL values due the different time spans available for each measure. Since Bitcoin price time series required a larger time span to compute the price momentum, it was decided to take Bitcoin closes from September 1 2021 to June 1 2023, thus allowing all the calculations that covers the 2022 and the ones for daily database update. The dominance value instead was just needed for all the 2022 historical calculation of the index, while for the daily calculation that specific value is extracted every day from CoinGecko API. For these reasons, for some periods of time we don't have information about the dominance, but this doesn't compromise the calculations.

3.2 | Reddit

In the pre-processing phase of our work, we took several measures at the source of data from reddit to ensure higher quality, consistency, and relevance. We initially considered the fact that the data came from 3 different subreddits ('r/bitcoin', 'r/Btc', and 'r/BitcoinBeginners') and implemented practices to manage them. To consolidate the data from these subreddits and enable optimal analysis, we performed a join operation, merging the datasets from all three subreddits into a single unified dataset. Next, we reset the index of the dataset and created a new numeric index starting from 100 million. This approach was chosen strategically to allow for future iterations and the continuous addition of new observations to the dataset. By establishing a higher index range, we aimed to ensure scalability and accommodate the potential growth of the dataset as more data points are collected. The time variable required specific processing to optimize its usability and relevance. We refined the data by removing the seconds, minutes, and hours from the timestamps, retaining only the date component. This transformation allowed us to utilize the date variable as a key for connecting and correlating data across different sources. By standardizing the time format, we simplified the analysis and ensured consistency in our subsequent operations. Cleaning the text data was an essential step to refine the quality of the posts and titles from

the Reddit data source. We deleted non-alphanumeric characters that were considered useless for our specific analyses and sentiment analysis. Additionally, we removed links, images, and wallet-related codes that were predominantly present within the text. By eliminating these extraneous elements, we aimed to extract the core textual content and enhance the accuracy of our sentiment analysis and subsequent analyses. Following the preprocessing steps, we conducted a sentiment analysis on the cleaned text data. The sentiment analysis aimed to identify and categorize the emotional tone expressed in the posts and titles. This analysis provided valuable insights into the overall sentiment, opinions, and attitudes prevalent within the Reddit discussions regarding Bitcoin. Finally, to distinguish the posts' source subreddits, we created an encoding scheme. We assign a numerical value (0, 1, or 2) to each post, indicating the subreddit it originated from (respectively 'Btc', 'BitcoinBeginners' or 'bitcoin'). This encoding allowed us both to differentiate the data based on the subreddit source, and to perform targeted analyses specific to each subreddit or compare the sentiments and trends across the different communities.

3.3 | Articles

In the preprocessing phase of the articles, similar steps were undertaken as described for the Reddit data source. Initially having data from two different websites and having two different datasets, we conducted a join operation to consolidate and combine articles from these sources into a single dataset. Assuming that articles have to be written following specific rules, and, unlike social posts, that they should't contain unnecessary information, we decided to avoid any text cleaning techniques and send the corpus directly to the sentiment analysis. For indexing, we created a new numeric index that acts as the primary key for articles. This index started at 100 million, allowing for future scalability and the addition of new articles. In the same way as for Reddit data, we conducted a sentiment analysis on titles, subtitles, and contents of articles that aimed to measure the sentiment expressed within the text and to understand the prevailing attitudes and opinions regarding Bitcoin within the news sources. Finally, to indicate the source of each article, we created a variable for encoding the origin site. We assigned numerical values of 0 or 1 based on the site of news origin. Specifically, we used the value 0 for the 'cointelegraph' site and 1 for the 'coindesk' site.

3.4 | Google Trend

Google Trends is a free online tool provided by Google that allows users to explore and analyze the popularity of specific search terms or topics over a given period of time. The tool displays the data on a scale from 0 to 100, with 100 representing the peak popularity of a term or topic normalized for the specified time range. For this reason, so to avoid a close-world perspective and to have a better understanding of how the Bitcoin trend moved during the period of interest, which was 2022, data was firstly collected from 2020, and then filtered for the 2022. This procedure allowed to have data normalized for a longer time window thus increasing its informative power. Another issue that had to be faced was that it is only possible to download daily search data for the last three month, while for requests that spans longer time frames, Google Trend gives only weekly data. To solve this problem and obtain daily-level information, we took an approach to infer daily observations using the weekly data provided. We started by converting the dataframe index to an object of type DatetimeIndex to allow time-based operations. Next, it was used the "resample" method to change the frequency of the data from weekly to daily. This allowed us to have a dataset with values for each day in the period of interest. Values for the intermediate days of each week were not available directly in the weekly dataframe. To solve this problem, we decided to assign the missing days the value of the corresponding week using the forward fill (`ffill`) method. Although this procedure required simplification by assuming that the weekly value is representative of each day within the week, it allowed us to obtain a more detailed view of the daily trends. The final dataset consisted of 365 observations, each representing a single day, with the corresponding normalized trend value. By converting the original weekly data into daily observations and assigning the weekly value to each corresponding day, we were able to create a comprehensive dataset that captured the daily fluctuations in the search interest for the term 'Bitcoin' throughout the year. In addition, by having daily data, we can integrate this source with other relevant datasets and conduct comprehensive analyses.

3.5 | Sentiment analysis

Advances in machine learning and deep learning has brought through the years to more and more capable models to analyse text and perform multiple operations with it. Sentiment analysis is the task of recognisign the opinion of a text and classifying it into desired classes such as positive, neutral, and

negative. Early sentiment detectors was based only on syntactic information such as part-of-speech tagging and opinion words [4], without exploiting all the potential of semantic information. Syntax is only the tip of the iceberg when it comes to natural language processing, so further models based on deep neural networks (LSTM, and later autoencoders) and knowledge representation capable of handling semantic information rapidly outperformed all the previous ones. Nowadays another class of deep learning models is taking the lead; they're based on the autoencoder architecture [7] and they're pre-trained on huge amounts of textual data. Their power lies in the ability to automatically learn how to represent natural language in a way that can be fine-tuned for every kind of NLP task. A popular pre-trained model is BERT [2], it was developed by Google in 2018 and trained on predicting masked words in sentences, but can be adapted to a wide range of NLP tasks. In the project, sentiment analysis was performed on all the available text by leveraging FinancialBERT, a version of BERT fine-tuned on sentiment analysis for financial-related sentences. More precisely, sentiment analysis was performed in two distinct ways:

- For historical data, FinancialBERT was downloaded on Cloab and loaded on GPU; the model took about 1h to process both articles and posts.
- For the pipeline, all the text of each day was sent to an HuggingFace Inference API endpoint that directly returned the results

The second approach avoided the need to download the model on our or hosted machines, making the processing much more efficient and faster. Even though the model is the same for both methods, the outputs were different, in particular:

- The model loaded on GPU returned only the confidence of the most probable score; for this reason the final score was computed as

$$\text{Score} = map(\text{label}) * \text{score\_confidence}$$

  where map(label) is maps 'positive' to 1, 'neutral' to 0, and 'negative' to -1.

- The model called by API returned the confidence scores for all the labels, so the final score was computed by firstly finding the most probable among the 3.

4 | Data quality

Data quality refers to the fitness of the data for its intended use. A piece of data is considered of good quality if it meets the requirements and is useful in representing reality, as it should be. Therefore, the concept of quality is not absolute, but rather depends on the purpose for which the data is intended. Data quality often involves a series of trade-offs, where the choices made are influenced by the intended use. In this context, there are several key concepts related to data quality:

- **Accuracy:** The degree to which data reflect the actual values or facts they are intended to represent.
- **Completeness:** The extent to which all required data elements are present and available in a data set.
- **Consistency:** The consistency and conformity of data across different sources or within a single dataset.
- **Temporal dimension:** The consideration of time-related aspects, such as timeliness, relevance, and validity of data in a specific period.

These dimensions or metrics are used to assign a numerical value to data quality, indicating the level of satisfaction they provide for a particular use case. We mainly analyzed data quality metrics regarding completeness and temporal properties as they were more appropriate for our project and for our data structure.

4.1 | Completeness

Completeness corresponds to the "coverage with which an observed phenomenon is represented within a dataset." We need to analyze two types of completeness:

- Data completeness within sources: it refers to the presence of missing data within the dataset obtained from various sources. In this case, we needed to assess if there are any gaps or missing values in the data that we have collected.
- General Data Availability: this aspect focuses on the overall completeness of the data in our possess compared to the total data available on the web. We needed to determine how much of the available data we have managed to collect and process for our analysis.

Completeness metrics:

- **Tuple Completeness** $\rightarrow \frac{\text{number of null values}}{\text{number of columns}}$.
  This metric is applied to each tuple (row), and then averaged.
- **Attribute Completeness** $\rightarrow \frac{\text{number of null values in the column}}{\text{number of rows}}$.
  This metric is applied to each column, and then averaged.
- **Table Completeness** $\rightarrow \frac{\text{number of null values in the table}}{\text{number of columns} \times \text{number of rows}}$.

The results are presented in the following table:

Table 4.1: Data Completeness Results

	Reddit	Google Trends	Articles	Prices
Tuple Completeness	0.06	\	0	0
Attribute Completeness	0.06	\	0	0
Table Completeness	0.053	0.86	0	0

From the results of the completeness matrices calculated for the reddit data, it can be seen that there is a not too excessive number of missing values in the dataset. For Google Trends we consider just the table completeness since at the end of pre-processing the database being composed by just one variable. This is why the completeness is very low. In the dataset of articles and prices there were no missing values, so the rates are obviously equal to 0. After data pre-processing, we recalculated the data quality metrics to

Table 4.2: Data Completeness Results after processing

	Reddit	Google Trends	Articles	Prices
Tuple Completeness	0.004	\	0	0
Attribute Completeness	0.004	\	0	0
Table Completeness	0.0037	0	0	0

see if there had been any improvements, and the results obtained are as follows:

We can see that we have had significant improvements. The number of null values in the reddit data have decreased a lot as extensive cleaning of the dataset has been done. While we notice how now the google trend dataset has no missing data as a result of the data adjustment explained above. In the end, in order to assess completeness, we analyze the general data availability by comparing the data we have obtained and utilized for our work with the entire data present on the web. It is crucial to determine if we are analyzing a substantial portion of the available data or if our analysis is based on only a small fraction of it. This evaluation allows us to ascertain the representativeness and reliability of our findings. Considering the data collected from Reddit, according to [6], "in 2022, redditors created more than 430 million posts on Reddit.com. Reddit.com had 2.5 billion comments as of 2022". By analyzing posts from only 3 subreddits and getting 114656 posts, we can calculate that we have analyzed approximately 0.027% of the total available posts on reddit. While this percentage may seem low, it is justified when considering the specific scope of our analysis. By analyzing more specific information from [5], we can approximate that the total amount of the post about crypto in 2022 was approximately 7 million. Additionally, assuming the number of posts related to a specific crypto being proportional to its market share, and considering that Bitcoin had a dominance of 40%, we can estimate that we have analyzed about 4,1% of the total amount of Reddit posts about Bitcoin. This estimation is very approximate but provides us with a useful understanding of the extent of our analysis.

4.2 | Temporal Dimension

For the temporal properties we consider frequency as the main metric. Since we have based all our database on date, and in particular daily data, it is therefore appropriate to check frequency by day, that is, how much data we have on average each day for each source.
The results are hinted in the following table:

Table 4.3: Data Frequencies

	Reddit	Google Trends	Articles	Prices
Frequency (by day)	314.13	1	17.04	1

As we can see, the datasets concerning Google Trends and Prices have a frequency of 1 as they have a daily value, so just one record for each day. Meanwhile, we observe a rather valid value of 17 articles per day, which is a plausible and good number. Furthermore, there is a substantial amount of Reddit posts being generated every day, with an approximate count of 314. The frequency of data collection per day indicates the amount of data we will collect daily. Considering the timeliness which refers to the measure of time between when data becomes available and when it becomes useful. In the context of our dataset, timeliness refers to the time lag between the collection of data (such as posts, prices, articles) during the day and when the data is processed, typically on the following day. Timeliness is crucial as it ensures that the data we analyze remains relevant and up-to-date. In our specific case, the time lag is at most of one day, indicating that there is a maximum delay of one day between the creation of data and the process that exploits those data. This time lag allows us to capture a comprehensive and accurate representation of the data available for analysis. Currency instead refers to the frequency at which the data is updated in relation to real-world events, it measures the speed at which the data reflects the current state of affairs. In our case, the currency level is 1, indicating that the data is updated daily. This means that the dataset is refreshed every day to capture the most recent information and align with the current state of the world. The daily updates ensure that the data remains current and relevant for analysis, reflecting the ongoing developments in the subject matter.

5 | Data storage

5.1 | Modelling

After data has been processed it was made the decision to store it in a Relational Database Management System (RDBMS). This kind of storage was preferred instead of other NoSQL data structures such as MongoDB for many reasons: firstly, RDBMS ensures data integrity through the enforcement of relational constraints, such as primary keys and foreign keys, which maintain data consistency and accuracy. For the aim of the project, which has the final goal to deliver every day an index combining data from different sources, the queries and the calculations remain the same each day. Since the usage of the data is pretty clear and defined, it was decided to store the data in a systematic way with a specific scheme that doesn't have the necessity to be changed for some time. Moreover, RDBMS offers robust query capabilities with the support of structured query language (SQL) which was exploited to perform data manipulations, aggregations, and joins useful to retrieve the data necessary for the index calculation. On the other hand, RDBMS also has certain drawbacks: one limitation is the difficulty in accommodating unstructured or semi-structured data so before loading the data on the database a pre-processing phase is necessary to better fit the raw retrieved data in a formatted and predefined scheme; this tends to limit the usability of the data for other purposes than the one initially in mind during the designing phase of the scheme. Storing the data in a RDBMS implies also to deal with the closed-world assumption and the NULL value. The rigidity of the scheme may also close the opportunity to easily add in a second time new type of data because it requires to update the structure of the scheme. Even though these limitations can impact on the scalability of the project, considering the pros and the cons previously discussed, it was decided to store the data in a RDBMS, and the strucutre of the databse can be summed up in the following scheme:

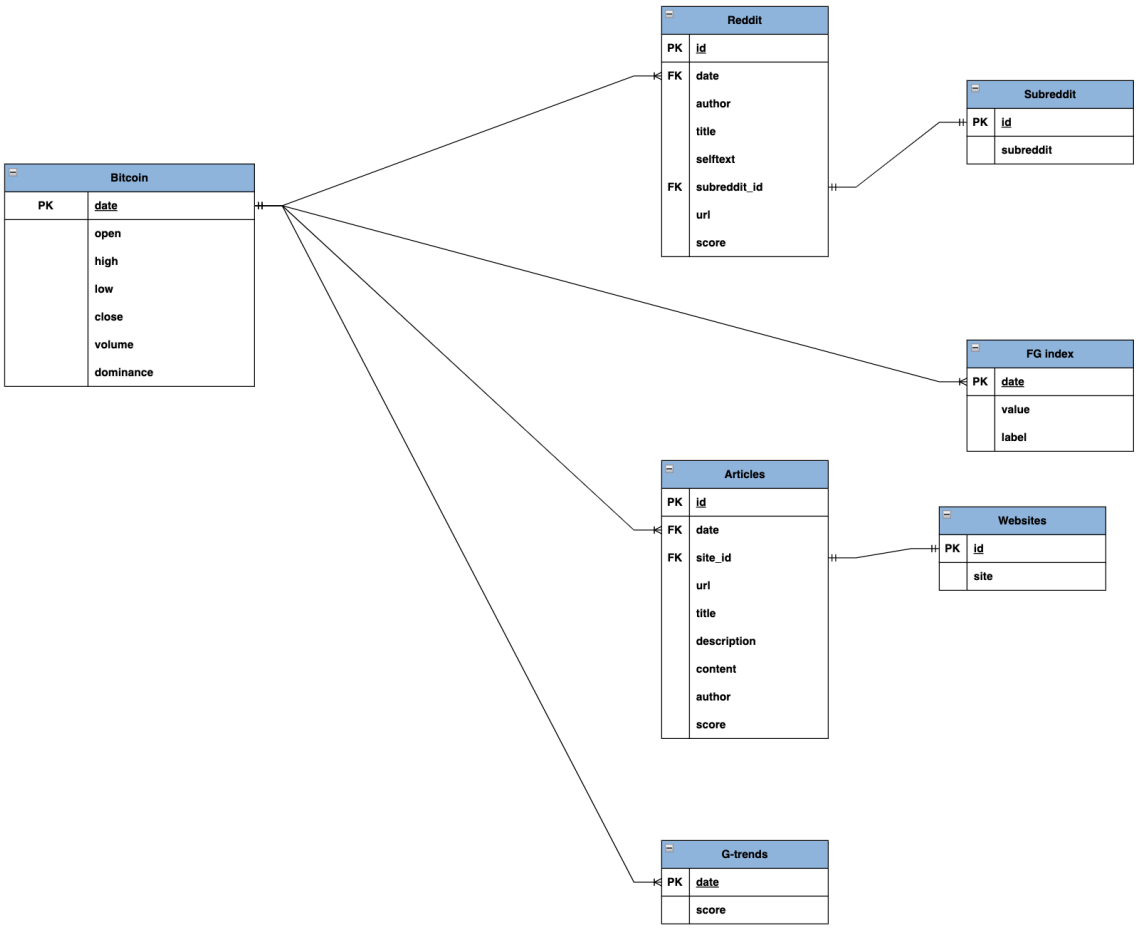

Figure 5.1: The scheme of the RDBMS structured to store the retrieved data.

The database can be queried through SQL language and, for example, the following code permits to retrieve for each day the Bitcoin close price and the article with the highest sentiment score:

Listing 1: SQL query 1

```
SELECT A.title as title, B.date as date, MAX(A.score) as best_score, B.close as price
FROM Bitcoin as B, Articles as A
WHERE B.date = A.date
GROUP BY A.date
```

The obtained output:

	title	date	best_score	price
1	Kevin O'Leary says his crypto holdings could reach 20% o...	2021-12-30	0.4367749227	47135.98
2	Cryptocurrencies Approach New Year in Positive Mood	2021-12-31	0.9980431795	46213.76
3	The year for Bitcoin: A 2021 roundup of the flagship crypto	2022-01-01	0.4345249547	47737.6
4	President Bukele predicts Bitcoin rally to $100K,further ...	2022-01-02	0.3289014101	47339.15
5	El Salvador to Build New Stadium in Collaboration With ...	2022-01-03	0.9972599745	46438.64
6	Nexo co-founder targets Bitcoin at $100K by mid-2022	2022-01-04	0.7113773482	45849.98
7	El Salvador prepares 20 bills to provide legal framework f...	2022-01-05	0.457289925	43438.22
8	US Congress to Hold Oversight Hearing on Crypto Mining:...	2022-01-06	0.9492229819	43142.34
9	Bitcoin-based tracking platform Eggschain partners with ...	2022-01-07	0.6231956556	41544.89
10	Will this time be different? Bitcoin eyes drop to $35K as	2022-01-08	-0.4779682438	41691.66

Figure 5.2: Output of the first query.

This second query is the one actually used to retrieve from the database all the data needed to make the calculations for the Fear and Greed index:

Listing 2: SQL query 2

```
SELECT B.date, close, volume, dominance, AVG(article_score) as article_score,
    AVG(reddit_score) as reddit_score, AVG(gtrend_score) as gtrend_score
    FROM Bitcoin as B LEFT JOIN (
        SELECT A.date, A.score as article_score, R.score as reddit_score, G.score as
            gtrend_score
        FROM Articles as A, Reddit as R, Gtrend as G
        WHERE A.date = R.date AND A.date = G.date
    ) as C ON B.date = C.date
    GROUP BY B.date
```

The obtained output:

	date	close	volume	dominance	article_score	reddit_score	gtrend_score
177	2022-02-24	38297.1	788.816769	0.398908779527991	-0.219058855744828	0.041646558515936	39.0
178	2022-02-25	39221.12	255.70831	0.3948194769792	0.17002643736842	0.142064384316434	39.0
179	2022-02-26	39111.11	142.971765	0.393122488640106	0.051108440033333	0.132238421016618	39.0
180	2022-02-27	37691.03	282.099937	0.372691853610715	0.0929729113499994	0.113253643851424	37.0
181	2022-02-28	43170.95	585.939547	0.404466896996282	0.0658050173133291	0.0986628184261082	37.0
182	2022-03-01	44396.56	242.320383	0.4186695531272	0.249100718914283	0.161409703560319	37.0
183	2022-03-02	43925.39	289.671327	0.413378641597326	0.134716061876462	0.133787504345951	37.0
184	2022-03-03	42466.67	141.903472	0.398108088228599	0.0221822942900017	0.148832941185804	37.0
185	2022-03-04	39123.41	158.787156	0.389163898913648	-0.0741172976117612	0.171798679649536	37.0
186	2022-03-05	39405.13	76.470685	0.379715630086922	0.0373696973750007	0.123938624098319	37.0
187	2022-03-06	38403.71	124.69034	0.377668422375148	-0.274166499166666	0.114026858933491	30.0

Figure 5.3: Output of the second query.

5.2 | Daily pipeline

To observe index movement outside the range of 2022, it was decided to implement a data pipeline that could compute index value at the end of each day by collecting all the necessary data. The result can be best resumed by the following schema:

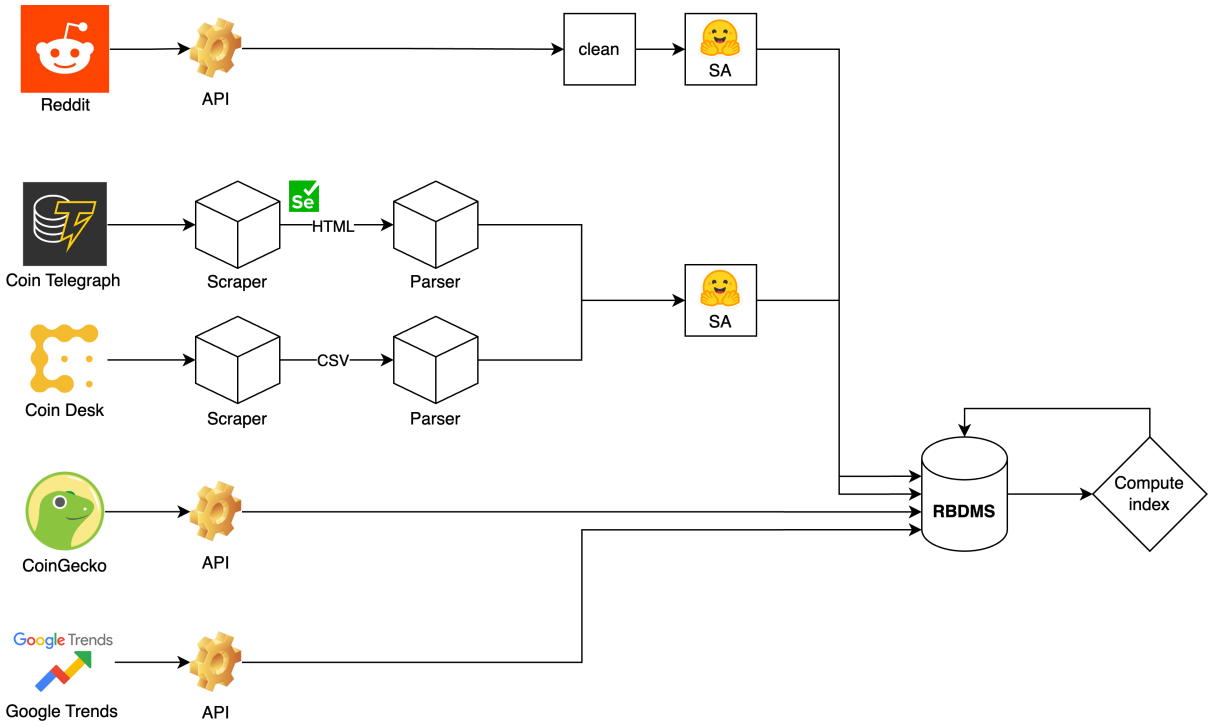

Figure 5.4: The scheme of the pipeline used to update the database every day.

The pipeline is composed by four main modules:

- Reddit module: it connects to Reddit API and gets all posts of the last month for each subreddit; the posts are then filtered for the day of interest, cleaned, and sent to Hugginface Inference API to be analyzed with sentiment analysis.
- Articles module: it scrapes CoinTelegraph and CoinDesk websites getting all the articles of the day of interest; the latter are sent to HF API to be analyzed with sentiment analysis.
- Bitcoin module: it connects to CoinGecko API and gets Bitocoin data about ohlc, volume, and dominance.
- Google Trends module: it connects to Google API and gets GTrend data about bitcoin over the last month; these data are then normalized with respect to the one already stored.

After the data fetching phase, all the data frames are then uploaded in the RBDMS. Finally, a function retrieves from the DB the specific data necessary for the computation of the index and returns it ready to be uploaded in the RBDMS.

6 | Fear and Greed Index

As mentioned multiple times before, the aim of the project is to collect, combine, process and store different data sources to develop an indicator which attempts to gauge the overall market sentiment and investor psychology surrounding Bitcoin at a given time, trying to follow what Alternative.me has done with the Bitcoin Fear and Greed Index. Alternative.me's Bitcoin Fear and Greed Index is an indicator built by combining data coming from many sources such as financial calculations, social sentiment, google trends and online surveys. This index ranges from 0 to 100, with lower values indicating fear, so that investors are worried and that this could be a buying opportunity. Higher values instead indicate that investors are getting too greedy, and this means the market is due for a correction [1].

Figure 6.1: Alternative.me Fear and Greed index range

The factors included in the Alternative.me's index are the following:

- Market momentum/volume (25%)
- Volatility (25%)
- Social media (15%)
- Online surveys (15%) - currently paused
- Dominance (10%)
- Google trends (10%)

6.1 | Index calculation

Following the criteria given by Alternative.me, first of all the single factors were computed, then they were normalized and the final index was the weighted average of the normalized single factors. Before deep diving on the single factors used to build the index, two important calculations need to be mentioned: the logarithmic returns ($\log(P_t) - \log(P_{t-1})$) and the momentum. Price momentum is a widely used technical analysis indicator that measures the strength and persistence of a price trend over a given period of time, one possible way to compute momentum is through the Rate of Change ($MOM_{t,n} = (P_t - P_{t-n})/P_{t-n} \cdot 100$). Momentum is an indicator of the "direction" of the market, which can be positive or negative. The single factors used for the index are:
Market momentum-volume (35%):

$$MOMV_{t,30} = MOM_{t,30} \cdot V_t$$

It was obtained by multiplying the daily volume with the 30 days market momentum. When we see high buying volumes in a positive market (positive momentum) on a daily basis, we conclude that the market

acts overly greedy / too bullish, while if we see high buying volumes in a negative market (negative momentum) it is assumed that the market acts moved by fear.
Directed volatility (25%):

$$\sigma_{t,30}^{adj} = \sigma_{t,30} \cdot MOM_{t,30}$$

It was obtained by multiplying the 30 days volatility as the standard deviation of the log returns with the 30 days market momentum. Market momentum gives a "direction" of the volatility value.
Daily average sentiment score of the Reddit posts (5%):

$$s_t^{Reddit}$$

For each day is computed the average sentiment score assigned to each Reddit post.
Daily average sentiment score of the Articles (15%):

$$s_t^{Articles}$$

For each day is computed the average sentiment score assigned to each article published.
Bitcoin dominance index (10%):

$$D_t = \frac{MarketCap_{BTC,t}}{MarketCap_{tot,t}}$$

The Bitcoin dominance is the ratio between the market capitalization (market cap) of Bitcoin to the market cap of the entire cryptocurrency market. It's also known as the Bitcoin dominance ratio. The dominance of a coin resembles the market cap share of the whole crypto market. A rise in Bitcoin dominance is caused by a fear of (and thus a reduction of) too speculative alt-coin investments, since Bitcoin is becoming more and more the safe haven of crypto. On the other side, when Bitcoin dominance shrinks, people are getting more greedy by investing in more risky alt-coins, expecting high profits.
G-trends score (10%):

$$G_t$$

Which is the score of Bitcoin researches from Google trends.
The final index is the weighted average of the normalized single factors:

$$FG_t = 0.35 \cdot MOMV_{t,30} + 0.25 \cdot \sigma_{t,30}^{adj} + 0.5 \cdot s_t^{Reddit} + 0.15 \cdot s_t^{Articles} + 0.10 \cdot D_t + 0.10 \cdot G_t$$

How much does each score weigh over the total score? It is important to consider the weights used to construct the index: there are numerous weighting schemes and strategies, they were tried many approaches like exploiting a linear regression of the returns using as covariates the factors lagged to the previous time period and exploit the estimated coefficients to understand the level of linkage between the returns to the previous period factors. But the best results were obtained following as near as possible the weighting scheme used by Alternative.me which of course may come from deeper studies and calculations.

6.2 | Results

To evaluate the results obtained by the computed index it were two main analysis:

- Index compared with the prices
- Index compared with a benchmark

In general, fear can be a sign that investors are too worried and that could lead to buying opportunities, while greed could mean that investors are getting too greedy and that could lead to a possible future market correction (which is a market decline that is more than 10%, but less than 20%). Looking at the price time series which was normalized, is it possible to appreciate how the index is evaluating each day of 2022. For example, in April 2022 the price it's on its peak for the year and the index gives a value of around 55 considering the sentiment of the investors neutral. The first days of July the index assigned a sentiment score of around 12 so according to the index the market was moved by fear. From the end of July to all the month of August the index gives a score of around 70/75, which means that the market is acting greedy, and that led to a decrease in prices during the next months.

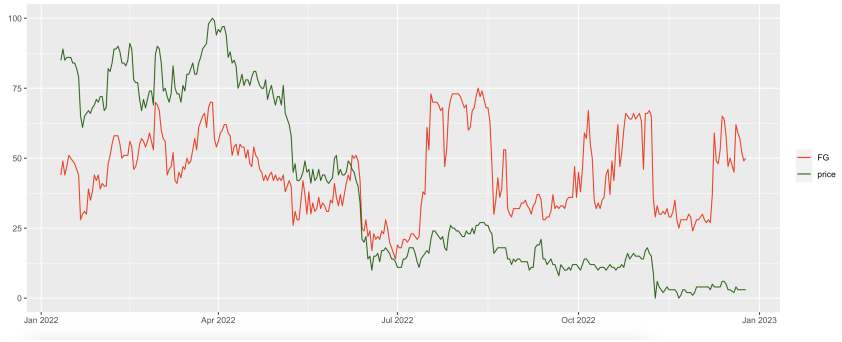

Figure 6.2: Fear and Greed index built 2022 time series and the 2022 price time series.

To better understand the goodness of the index, it was compared with the benchmark, the Alternative.me's Greed and Fear Index that can be easily retrieved using their API. From a graphical point of view, what mainly stands out is the fact that our index tends to assess a higher level of greed of about 20%. For some months of the year our index followed quite well the benchmark (January, February, July, September and November). While for the remaining months, our index tended to have a 20% higher value. Overall our index tends to perceive the market greedier than the benchmark does.

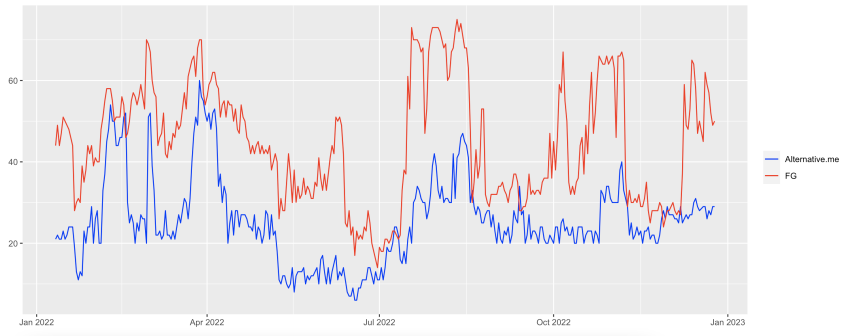

Figure 6.3: Fear and Greed index built 2022 time series and the 2022 Alternative.me's Fear and Greed index time series (benchmark).

Therefore, it should be noted that Alternative.me's index is the result of a more comprehensive, competent and in-depth study than ours. Surely analyzing, studying and combining the available data better will lead to better and more refined results.

7 | Conclusions and further developments

This project shown how retrieving and combining data from different sources gives the opportunity to obtain a snapshot of the sentiment of the crypto market that can help investors in the decision-making process. The data coming from reddit posts, web articles, Google Trends were processed and combined with the Bitcoin financial data to obtain an index, inspired by the job done by Alternative.me, that can explain the sentiment of the Bitcoin market. After comparing our results with the already existing Fear and Greed index, we found them satisfactory but not perfect. In fact, much work and adjustments focused on the calculation and the weighting of each factor can be done to improve the index.
We think our project has also lot of potential of future improvement from different aspects. In particular, the following will be discussed:

- Content scalability
- Currencies scalability
- Purpose scalability

In the project the text data come from 2 websites and 3 subreddits. A possible improvement involves expanding the data sources: adding for example more subreddits, including the content inside the comments of each post to capture the interactions between users but also going beyond Reddit and articles to include additional social media platforms like Twitter. By incorporating data from Twitter, we can capture a wider range of real time conversations and opinions surrounding cryptocurrencies. Furthermore, we can consider scraping information from new websites to gather diverse perspectives on the crypto market. Currently, our project focuses on analyzing Bitcoin, however, a natural progression is to extend the index to other prominent cryptocurrencies such as Ethereum, Litecoin, or Ripple. By expanding our scope to multiple cryptocurrencies, we can gain insights into the market dynamics and interrelationships among various digital assets. This broader perspective will provide to the investors a more comprehensive view of the crypto market sentiment. One of the drawbacks of storing data in an RDBMS is the difficulty in accommodating unstructured or semi-structured data, so before loading the data on the database, a pre-processing phase is necessary to better fit the raw retrieved data in a formatted and predefined scheme; this tends to limit the usability of the data for purposes other than the one initially had in mind during the designing phase of the scheme, thus hindering the development of new analysis and purposes that may come out in the future. A solution can be storing all the data retrieved every day in a data-lake. A data lake serves as a centralized repository that stores diverse data sources in their raw format. By establishing a data lake, we can integrate and store data from various sources, including social media platforms, news articles, and other relevant data. This centralized repository will enable to conduct future analyses efficiently, explore new research questions, and apply advanced analytics techniques to gain deeper insights into the crypto market.

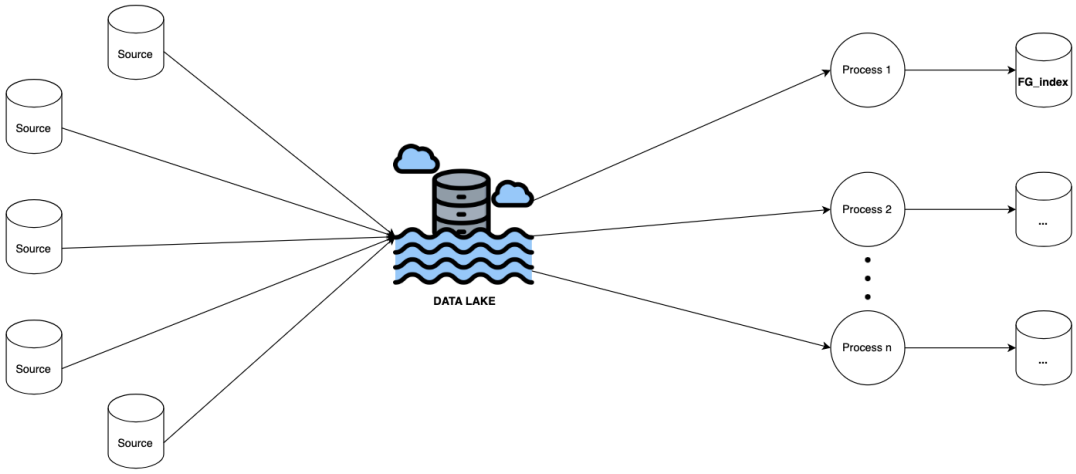

Figure 7.1: Example schema of a data lake storage system.

The members of the group and their specific contribution to the project are listed below:

- Salerno Fabio:
  - Data acquisition: Reddit API, Bitcoin data API
  - Pre-processing: Bitcoin data
  - Data modelling
  - Fear and Greed index construction
- Ghilardi Davide:
  - Data acquisition: Web Articles scraping
  - Sentiment analysis
  - Daily pipeline
  - Fear and Greed index construction
- Lobosco Lorenzo:
  - Data acquisition: Google trend API, Web articles scraping
  - Pre-processing: Reddit, Google trend
  - Data quality

8 | References

[1] Alternative.me. Crypto fear greed index. https://alternative.me/crypto/fear-and-greed-index/, 2018.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Forbes. What is bitcoin? https://www.forbes.com/advisor/investing/cryptocurrency/what-is-bitcoin/, 2021.

[4] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015.

[5] Indian Express. Cryptocurrency was the most popular conversation on reddit in 2021. https://indianexpress.com/article/technology/crypto/cryptocurrency-was-the-most-popular-conversation-on-reddit-in-7667454/, 2021.

[6] Use Sign House. Reddit stats. https://www.usesignhouse.com/blog/reddit-stats, 2022.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.