**ETL project report – Bitcoin vs. Ethereum**

The theme we have chosen for this project is finance and after a discussion on what sector to look into we decided to go for cryptocurrency. The findings we are searching for are that of which of the 2 cryptocurrencies is more reliable to invest in. Using Extract, Load & Transform on the data sources we are able to gain professional findings.

Diagram

Description automatically generatedThis is the process we followed.

**EXTRACT:**

The data we decided to use is that of the tokens Bitcoin and Ethereum as these 2 are the most well-known coins in the cryptocurrency market in the last 5 years.

The 2 original data sources we decided to extract from were from the 2 URLs below:

|  |  |
| --- | --- |
| **Alphavantage.co** | **Kaggle.com** |
| 2 JSON Files | 2 CSV files |
| *URLs:* https://www.alphavantage.co/query?function=DIGITAL\_CURRENCY\_DAILY&symbol=ETH&market=USD&apikey=  https://www.alphavantage.co/query?function=DIGITAL\_CURRENCY\_DAILY&symbol=BTC&market=USD&apikey= | *URL:*  <https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory?resource=download&select=coin_Aave.csv> |

In order to get the most up to date information we had to use an API, as the CSV files only went up to 2021.

The 2 CSV files followed the same format they both had 10 columns and over 2000 data entries, while the JSON files consisted of a dictionary of dictionaries as shown below.

*Text

Description automatically generated*

Graphical user interface, application, table, Excel

Description automatically generated

**TRANSFORM:**

The CSV files were much more organised by the data being in columns, rows and included headings which stated what the data was meant for. The JSON files had a different structure to the CSV files as well as a different range of data. The following Transformation changes were made on the files:

* From the CSV file we dropped the unrequired columns, we also dropped any data entry before 08/08/2015 for the bitcoin CSV so the two datasets matched.
* We also formatted the date columns to a date format so that we could use it for data manipulation.
* The JSON Files did not have names and symbols columns and were organised differently. There were headings within the data that were not necessary such as Highs and Lows of the Cryptocurrency rate. We iterated through the files and collected the data we required in lists. These lists were then used to create a dictionary which was then used to create a data frame.
* We went on to add name and symbol columns.
* We limited the data to after 07/07/2021 so that we could append it to the data from the CSV files.

**Table

Description automatically generated**

**A picture containing text

Description automatically generated**

**LOAD:**

As the data we have collected is relational and contains the identical columns for each currency, we decided to use SQL to display our data. In SQL we created a database (crypto\_db) which contains the two final dataframes we created. Both tables held the date as the primary key, the name, symbol, open and close values. We then went on to join the two tables on the date in order to be able to compare their values. The below ERD illustrates this.

Using SQLAlchemy we made a connection to pgAdmin and appended the two tables to the database.

**Graphical user interface, application, table, Excel

Description automatically generated**Graphical user interface, application

Description automatically generatedThis topic of which cryptocurrency is more sustainable/reliable, Ethereum or Bitcoin, was chosen by our team because it changes the way people look at investments and the investment process. With this information we have gathered and sorted through, we can find whether Ethereum or Bitcoin is more reliable, therefore helping us determine which of the two is a better investment. The data we have decided to keep can be manipulated to find information, such as the daily change and % growth of the currencies, in order to find the better currency of the two. Another reason our team wanted to investigate the two cryptocurrency is because of the worldwide popularity with how fast, digital, and secure it is.

**Table

Description automatically generated**

**Limitations:**

While doing our project our team incurred 2 major issues. The first issue was that because we were using 2 different data sets there were some slight differences in the figures. With some research we found that these discrepancies occur because Ethereum and bitcoin are their own currency used around the world this makes the data slightly different depending on which country you are recording it from. To fix this problem we decided to choose the figures from one data set so they could match as the discrepancies where not that large. The second problem was that the CSV had capital letters and the JSON file did not. SQL did not accept the fact that the CSV had capital letter and would not work, it took a lot of trial and error for us to realise and rectify this, making it take a lot longer than we had initially expected.