**ETL Project: S&P 500**

**Data required:**

* JSON : stock.json <https://datahub.io/core/s-and-p-500-companies>
* CSV: all\_stocks\_5yr.csv <https://www.kaggle.com/camnugent/sandp500>
* CSV: wiki.csv <https://en.wikipedia.org/wiki/S%26P_500_Index>

**Description:**

From the stock.json ('https://pkgstore.datahub.io/core/s-and-p-500-companies/constituents\_json/data/64dd3e9582b936b0352fdd826ecd3c95/constituents\_json.json') we obtained the name, sector, and symbol.

We extracted the open and close from the second csv (all\_stocks\_5yr.csv) and merged it to the first set of dataFrame (obtained from the json data) resulting in ‘df2’ dataframe.

We then loaded a 3rd set of data, ‘wiki.csv’ from Wikipedia.com to obtain the headquarters, Date first added and the Founded year. An inner join with df2 was performed using the symbol in order to get the “final\_df”.

Rather than deleting the rows containing the empty cells, “final\_df” was cleaned by replacing the empty cells with “No Data Available” so as to preserve the list of total stocks.

We then connected to pgAdmin4. In order to do this, an empty database must first be created within pgAdmin4 by right clicking on the Server and clicking on ‘Create database’. You can call this database anything, but for this project, the database name is called ‘stocks2’.

The client or API may use this PostgreSQL data as they wish.

For completeness, we also brought our data back from PostgreSQL to csv and into jupyter notebook.