**ETL Project: S&P 500**

**Team:**

**Baljit Alang**

**Maria Javier**

**Daniel Kline**

**Data Sources:**

* 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:**

***Data Extraction:***

From the **URL** - url = 'https://pkgstore.datahub.io/core/s-and-p-500-companies/constituents\_json/data/64dd3e9582b936b0352fdd826ecd3c95/constituents\_json.json'

We directly extracted data and got the following information:

Company Names, Sectors and Ticker Symbols

From **Wikipedia**, we extracted the following information:

Ticker Symbols, Location of the company headquarters, Date first added and the Founded year.

***Transform:***

Both the above datasets were converted into Pandas Dataframe

We then merged the above 2 dataframes into a third dataframe which has the following information:

Company Names, Ticker Symbols, Sectors, Headquarters, Date first added and the Year founded

This dataframe was then cleaned by replacing the empty cells with “Value Unavailable” so as to preserve the list of total stocks.

***Data Load***

We used Postgresql DB – We chose this DB because it is easy to work and our data was very much organized.

PostGres Connection –

We connected to postgresql Database.

Firstly a new database has to be created in Postgresql.

Then using Jupyter Notebook, the final dataframe was converted into an SQL Table - st\_table.

The select query was executed from jupyter notebook to confirm that the data was loaded into the DB

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