

**Project Name** – ETL (Extract, Transform and Load) technical report

**Submitted by**- Amber Sidhu, Raj Sisodia

**Project Proposal** – Currently, as world is facing a pandemic called Covid-19, the economies across the globe are fluctuating. There is a variation in the market which is leading to uncertainty across different sectors. These changes are leading to unemployment, lack of demand of the nations good’s and services and decrease in productivity. In contrast to that our company figured a bunch of data which includes USA covid-19 and NYC stock exchange files and our team is tasked to migrate it to production data base.

**Sources of Data –** a) data. World b) Kaggle c) Yahoo finance website

Data Links are provided below:

<https://www.kaggle.com/sudalairajkumar/covid19-in-usa#us_states_covid19_daily.csv>

<https://www.kaggle.com/dgawlik/nyse#prices.csv>

<https://ca.finance.yahoo.com/quote/SPGI/history?p=SPGI>

**Main Procedure-**

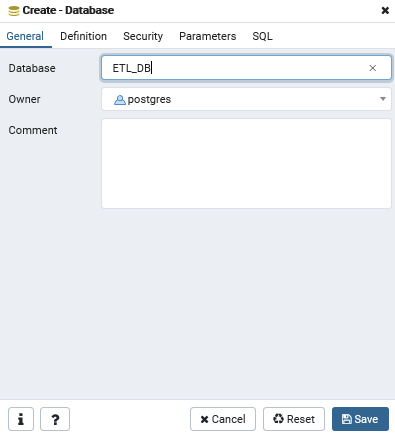
ETL - ETL stands for Extract, Transform and Load, which is a process used to collect data from various sources, transform the data depending on business rules/needs and load the data into a destination database.

The project started with lots of brainstorming within the team where we discussed different scenarios and ways to receive the data. We went through various websites like data.world, Kaggle and yahoo finance to finalise the data which looked productive and appealing to our eyes. The **USA Covid-19 csv** and **New-York stock exchange csv** files used for this project are taken from the above-mentioned sources of data. Even to make this project more challenging we scraped a stock table from yahoo finance website.

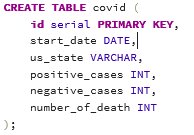
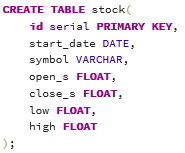
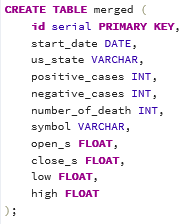
**Extraction Process** -

The individual csv files were visualised by team members to discover if the columns and rows are well structured with proper values. The team after mutual discussion finalised the columns in both the data sets and started working ahead.

The process started with creating a **database in pgAdmin named ETL\_db**



After which, tables are created to which the data is loaded after the execution and transformation process. In the below pictorial three tables are created in pgAdmin i.e. covid, stock and merged table.

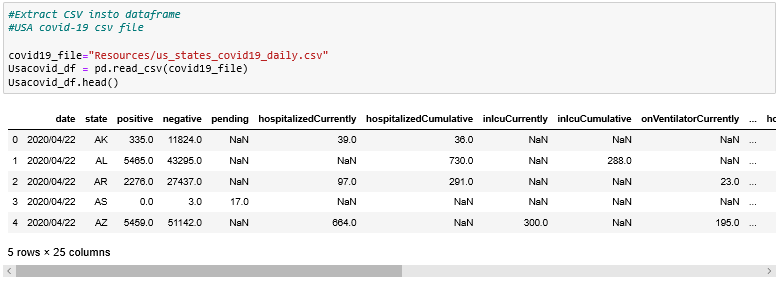
  

Select\* from covid, select\* from stock and select\* from merged commands are used to run and create tables in the database.

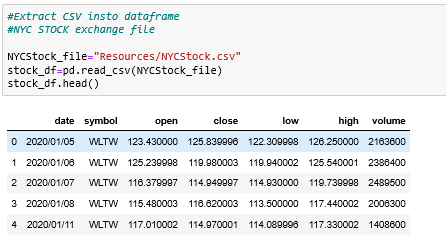
**Jupyter Notebook**

The next extraction process is to extract the excel csv files (**USA Covid-19 csv** and **New-York stock exchange csv**) in Jupyter notebook. It begins by importing the correct dependencies. For this project, the following dependencies were imported, import pandas as pd, from sqlalchemy import create\_engine and import datetime as dt.

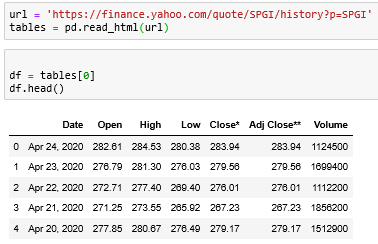
Extract the **USA covid-19 csv file** by giving the path where the file is situated and using **pd. read\_csv file** command. Use**. head( )** command to visualize the data frame as shown in the picture,



Now, extract the **New-York stock exchange csv** file by giving the path where the file is situated and using **pd. read\_csv file command**. **Use. head( )** command to visualize the data frame as shown in the picture,



The second different approach is directly scraping the table from yahoo finance website. The website URL along with pd.read\_html(URL) command is used to perform the task.



**Transformation Process** – The main objective of the transformation process is to mould the data according to the specific needs. The data executed can provide unnecessary information which is not required. Transformation process in Jupyter notebook, can help to modify the data and clean according to the business requirement.

For the following two files imported, the transformation will be done separately.

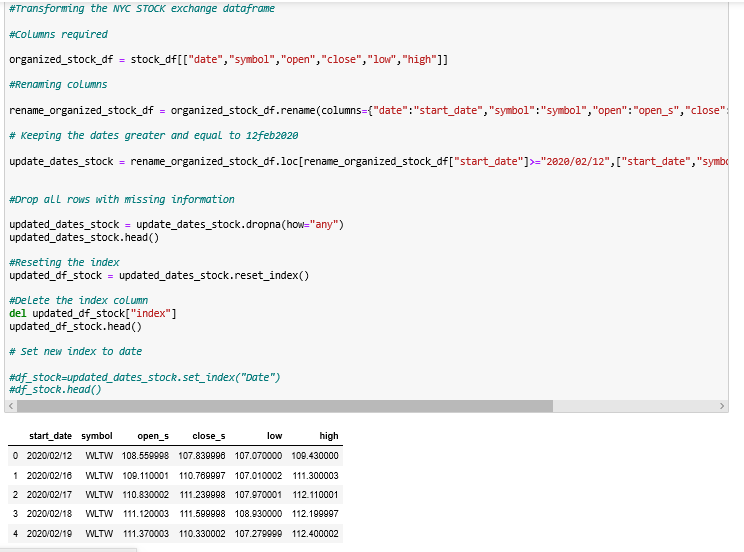
In the **USA covid-19 data frame,** there are certain columns which are not required. Use **df [[]]** command to eliminate unnecessary columns and columns are renamed using **df. rename(columns={})** command.

In the following data frame **start\_date, us\_state, positive\_cases, negative\_cases, number\_of\_death** columns are used. According to the project requirements the data is sorted on **dates** as the number of cases spiked in USA after February 28,2020. The dates are then arranged in ascending order to keep the dates in line and index is reset to have another column with serial numbers. In the data there were many rows with NaN values, all those values were converted to zero.

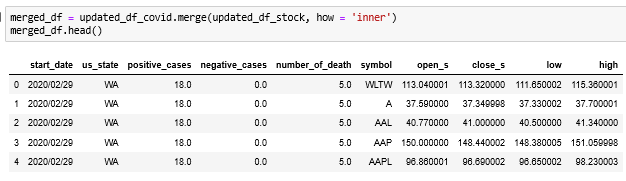


In the **New-York stock exchange data frame CSV file,** there are certain columns which are not required. Use **df [[]]** command to eliminate unnecessary columns and columns are renamed using **df. rename(columns={})** command.

In the following data frame **start\_date, Symbol, open\_s, close\_s, low, high** columns are used**.** The main objective is also to determine how the stock market fluctuated as the number of cases spiked in USA. So, data is sorted on **dates** to identify the stocks in USA after February 28,2020. The dates are then arranged in ascending order to keep the dates in line and index is reset to have another column with serial numbers. The stock csv is a big file with many rows with missing information, so all the rows with missing information is dropped from the dataset.



Both the data( **USA Covid-19**  and **New-York stock exchange** ) are merged using inner joint and on date.

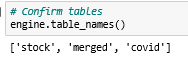


**Loading Process –** The process is to load the data after the transformation/modification process into the tables created in pgAdmin initially.

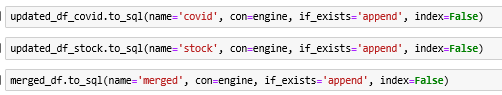
The process starts by creating connection string which consists of username and password along with local host details of the database followed by creating engine and giving the connection string to PostgreSQL.



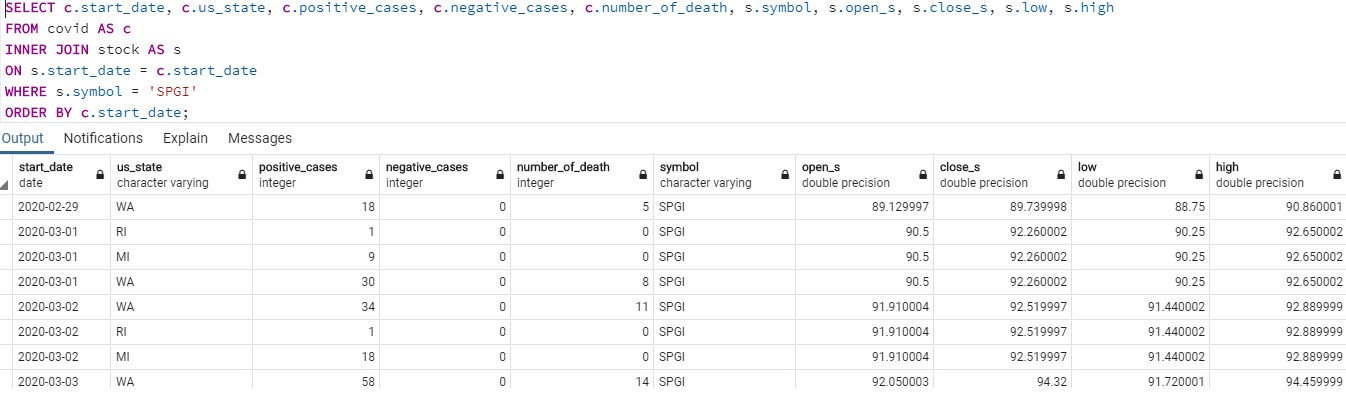
Confirm if the connection is made by checking the tables using **engine.table\_names()** command.



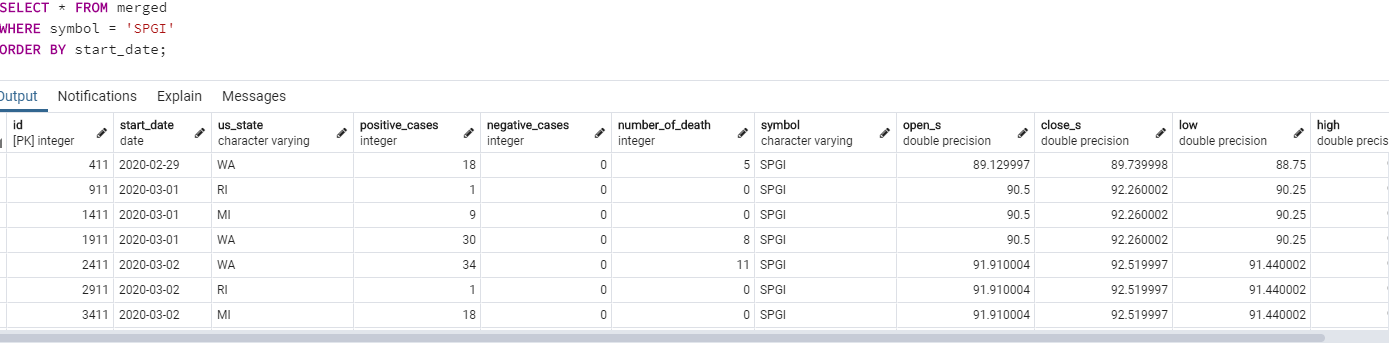
Convert the individual tables formed in the jupyter notebook using pandas to sql using **.to\_sql** command and append the data into sql tables.



**QC & Data Validation** – Check the SQL database, if tables persists and data holds it value from the loading process by running the following query in Query Tool. Here we are checking data for S&P Global Inc. using its symbol “**SPGI**” to identify its stock price variation from dates starting from 28th Feb, the oldest date our combined data holds for covid-19 and stock tables.



We can validate the results with merged data table, by running following query. Both queries returned 1895 rows. Interesting to note that, above query returned result in 337msec whereas this one took 417msec.



This data is selected to determine the fluctuation in the stock market due to covid-19 pandemic. It is wisely said that any form of major crisis can result in fluctuation. The data can help investors to follow the stocks closely during the pandemic and can help the major industries to compare their results with their competitors.