

Coursework Title: ETL Project

Team: Team #3

Class: RU-HOU-DATA-PT-04-2020-U-C-TTH

Program: Data Visualization Part-Time

Instructor: Siri Surabathula

Date: August 1, 2020

# Project Title: ETL of S&P 500 Data from 2015-2020 (current)

# Team Members: Ross Davidson, Brodie Armstrong, Mujgan Guner

# Project Description:

S&P 500 is a stock market index that measures the stock performance of currently 505 companies listed on stock exchange in the United States. In this project, our objective was to obtain the daily data of S&P 500 companies from various sources in the web and deploy the data in a relational database.

# ETL – Extract-Transform-Load

For this project, ETL procedures were used to achieve a final relational database as an end-product. In the following sections, the steps and methods are described.

#### **EXTRACT**

S&P 500 company list is updated many times within a year by company additions and removals. Therefore, while extracting the data, our goal was to get all the relevant companies’ daily data, company names, company tickers, the current situation of the company if they are still in the S&P 500 list or not, the reason of the addition/removal, dates of addition/removal. For the data extraction phase, python jupyter notebooks (libraries: pandas, requests, data reader, tqdm, datetime) were used.

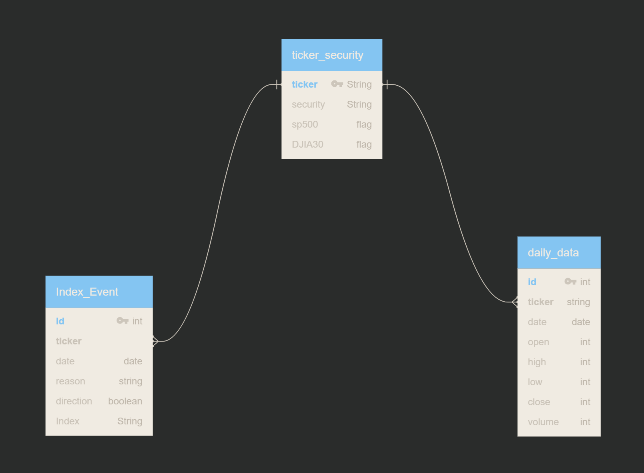
# Data Sources for S&P 500 Data Extraction

S&P 500 index daily data per companies are widely available in the internet. To start with, the daily data from between years 2013 to 2018 are downloaded from Kaggle. (<https://www.kaggle.com/camnugent/sandp500/data>). The initial plan was to obtain the remaining data from 2018 to date from APIs to populate the daily dataset. However, checking the data integrity of Kaggle csv file, it was observed that all companies in S&P 500 in relevant years were not fully reflected to the dataset. There were plenty of missing companies for a given year. Therefore, some steps were considered to make sure all relevant company daily data are available in the daily dataset.

1. Current S&P 500 company list was found from the internet and their daily data was extracted from two sources:
   1. <https://www.alphavantage.co/documentation/>
   2. Yahoo finance by using “pandas data reader” library
2. After checking the data integrity of the daily data from both sources, alphavantage API data source was selected to populate the database. (We did not use the Kaggle data for final database)
3. Alphavantage API data was checked and the missing companies to cover all the target years were identified. Then, a list of remaining tickers was obtained to pull the remaining data to cover all the companies in the S&P 500. Daily data consists of date, open, close, high, low, volume and tickers columns.
4. Other datasets obtained for this project were csv files with tickers and company names, company add/remove reasons and dates.

#### **TRANSFORM**

Once the necessary datasets were extracted, an Entity Relationship Diagram (ERD) was drawn. ERD is shown in the figure below. ERD was used as a guide to transform all the data from multiple data sources.



While transforming data, jupyter notebooks were used (libraries: pandas). Finally, the cleaned up and transformed datasets were converted into csv files ready to upload to the database.

#### **LOAD**

The load phase of the project was to load the final csv tables to a relational database. The relational database is selected due to the nature of our data. Our datasets consist of a large set of daily data, by using SQL database, we have avoided repeating some informative data such as company name, add/remove reason and normalized the database structure to second normalized form. For this project, PostgreSQL database was selected since it is open source. Jupyter notebooks with SQL Alchemy library was utilized for the database deployment. Final database was loaded to postgres, necessary SQL queries were written to make sure the load phase is successful.