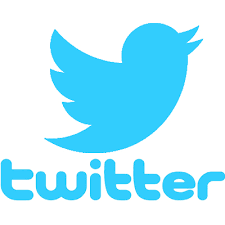
## horizontal line



Boot Camp Project II

University of Toronto SCS - Data Analytics

Stock Data ETL - Top Ten Stock from Twitter

April 17, 2019

**─**

**Project Members:**

Quan Shuang

James Jean-Jacques

# Overview

The objective of the project is to extract stock tickers name from Twitter user’s tweets using the Twitter Search API. The stock tickers names will be used to extract various additional data from Yahoo Finances. We will scrape from Yahoo Finance websites using the stocker name list of companies to gain access to extract data such as stock summary (price), profile (address and key executives). The data will be transform and store it to tables within a MySQL Database.

**Twitter user Feed:** “Stock Watch List”**,** “Stock Market”, “Biostocks”, “Stock Tweets”, “Stock Picks”

**Yahoo Finance URL:**  <https://finance.yahoo.com/>

**Database**: MySQL

# Goals

1. Acquire the stock ticker data from Twitter using API
2. Acquire the financial data from Yahoo Finance using Bootstrap
3. Transform the data and merging into tables in one Database
4. Load the data into python Pandas (Jupyter Notebook)
5. Load Pandas to MySQL Database
6. [LATER] Using Flask Create HTML pages and routes
7. [LATER] User should be able to lookup company details using HTML interface

# Specifications

Using Jupyter Notebook, Pandas, SQLite, Twitter API, Yahoo Finance, Json, bsoup, Flask, (Google Api)

# 

# Milestones

## Extraction

Extraction of the Data occurred from two Data Sources Twitter API Data and Yahoo Finance. Import the following library modules such as requests, json, pandas, datetime, re (regex), bs4 (BeautifulSoup), Browser (splinter), Searchtweets (ResultStream). Use for-loops to acquire the data and then create lists of the data values extracted.

For Twitter API we needed to follow the following steps to create an account and eventually pull the data.

Twitter API new account setup and user access guidance:

* Create a App
* Create a Project
* Apply for access keys (key and secret)
* We generated the Bearer Token using access key and secret key once Twitter approved API
* Use Bearer Token to construct headers for requests.get

Twitter API Limitations of (free) account

* Access to tweets from the past 7 days (instead of unlimited access from 2006 for premium and enterprise accounts
* 15 tweets per query (instead of 200 queries for paid accounts)
* Max 100 queries per day (limiting the number of keywords we can choose)

Twitter Data Format is JSON.

For Yahoo Finance we needed to use splinter browser and “ChromeDriver” to allow us to use BeautifulSoup to pull the specific data we needed. Yahoo Finance provides free financial data; therefore, we did not require creating any user access or security token.

Yahoo Finance Data Format is BeautifulSoup.

## 

## Transformation

The following approach were taken for the data transformation from both data sources Twitter and Yahoo, Data. The following transformation was completed to the data.

* Removed extra “$” characters from stock Tickers
* Removed extra “.” characters from stock Tickers
* Removed extra “B” characters from Market Caps
* Removed extra “N/A” characters from missing Earning Date
* Removed extra “,” characters from Volume values
* Removed extra “-” characters from the Telephone number
* Removed extra “M” characters from Key Execs Salary
* Convert Market Caps values from string to float
* Convert Volume values from string to float
* Convert Telephone number values from string to float
* Convert Key Executives Salary values from string to float
* Combine DataFrame using Pandas concatenate

The data in the Pandas data frame were create using dictionary.

## Loading

The loading of the data involves the creation Relation Database on MySQL. Creation of Table and define values types. We create four Tables,

1. Stock Data
2. Summary
3. Profile
4. Sustainability

For Twitter API data, we created a pandas list and sorted with pandas groupby.

For Yahoo Finance data, we created four Pandas data frames using the created list of values. We then created a connection engine for data load to MySQL.

**Twitter Stock Database Schema**

