**Data Analytics broad plan**

1. New York Stock Exchange Datasets (4).
2. **Introduction:** This public dataset has stock exchange market data available for all corporate shares that sold through New York Stock Exchange market. The end-of-day financial data is historical for corporations whose stocks are sold through NYSE over period of 2012 through 2016. There are 4 datasets (comma separated) which are called 1. Fundamentals.csv, 2. Prices-split-adjusted.csv, 3. Prices.csv, 4. Securities.csv. The first dataset (Fundamentals.csv) has 79 columns of various financial data (10-K) about participating corporations that were filed over quarters during years 2012 through 2016. The second dataset (Prices-split-adjusted.csv) has information about prices, adjustments and splits on stock for each day in 7 columns. The third dataset (prices.csv) has US Dollar prices (opening, closing, highest and lowest) and volume of stock units of various corporations for each day in 7 columns. The fourth dataset (securities.csv) has further non-financial information about each client company. Each client company whose stock sells in this market has a unique ticker symbol using which the client can be identified.
3. The web-link is : <https://www.kaggle.com/dgawlik/nyse#prices-split-adjusted.csv>

In order to download, create a Kaggle account and download the datasets into S3 bucket.

1. In Athena, check if we can query the data in bucket.
2. The dataset for our most common usage will be prices.csv (or Prices-split-adjusted.csv) that carries end-of-day stock price information and volume (i.e. volume of stock units traded on the day).The detailed description of all the fields on all datasets can be found on above web-site. All data is in basic structured format.
3. **Broad steps of data processing:**
4. **Initial check:** Data is generally clean.



1. **Exploratory Analysis of data:**
2. **Problems using machine learning:** On SageMaker ML tool from AWS, we need to analyze the data.
3. Can we try to predict the closing price & volume for Ticker symbol = ‘YHOO’ that will occur after 5 days? Train thru all the data and then predict closing price & volume after five days given the opening price. Let us separate data into three parts the last part being say 20 rows of batch data (all for ‘YHOO’ ). Let us do a batch transform and see the values. We may use linear regression model for this.
4. Can we use XGBR model instead of linear regression in above problem?

1. Can we find top 10 profitable stocks as of Dec, 2016?
2. **Data visualization:** Can we plot average price for all stocks over last one year 2016?