**Machine Learning Task**

Analyze the NYSE Data Set for the year 2000-2001.

You can download this file here:

<https://s3.amazonaws.com/hw-sandbox/tutorial1/NYSE-2000-2001.tsv.gz>

The Dataset has over half a million records.

By using K-means Clustering divide the dataset into 3 clusters (Low, Medium, and High. ) Low would have all the values lower than a particular threshold, Medium would have a set of values and

High would have a different set of values.

Map stage: calculates the distances between samples and centroids; match samples with the nearest centroid and assigns them to that specific cluster. Each map task is processed with a data block.

Reduce stage: recalculates the centroid point using the average of the coordinates of all the points in that cluster. The associated points are averaged out to produce the new location of the centroid. The centroids configuration is feedback into the Mappers. The loop ends when the centroids converge.

The results would include complete analysis of the increase or decrease in the stocks over a period of years. Also if there was a high at a particular point or a low at a particular point we can gather it from the data set.

Now that we have different clusters with the required information, we can predict whether that

particular stock would give us a profit of not.

Map Reduce will have to perform several steps in order to gather all the required information like price\_low, price\_high price\_maximum etc. about the ups and downs in the stock values and predict the values of stock in future.

Say suppose in a certain period of time the cost of the stock is high and sometimes it goes down to low, with the help of the output we get from k means regarding the prices of the stocks, we get.

The stock names in a particular months divided in categories say high profit low profit and medium profit.

We also get the outliers for the stocks which cannot be categorized in any of the above three clusters