Project description

Project Title

Passing streaming data of stocks and predicting changes in stock data using historic data and its affecting stocks of other companies.

Project Objective

Streaming the data into a limited size windows to process data to work parallel over huge data.

Then on processing data we analyze it to divide the stocks into different clusters based on the pattern of data processed in that particular window.

Now that we have the data which is already analyzed and stored into various clusters we now can analyze changes in new stock and of those whom so ever are getting affected.

And if a particular stock has a high variance will get into a new cluster and the behavior would be interdependent on the cluster in which it is present.

All of this is shown through a visualization tool which would animate the patterns of stocks data.

Project Description (Methodology)

The data of stocks would generally be millions per minute but we would rather consume data by randomly generating and sending in to buffers.

Then we use spark- streaming where RDDs forms a DStream (discretized stream)

Here input data stream is passed to Spark streaming which would produce batches of input data to the spark engine in turn returning the batches of processed data, as we are already sending the data in to different batches through DStreams it would give us the data of a particular ‘window’ in this case as we discussed earlier.

Receiver receives the data from sources and store in spark memory (buffer) for processing, so that operations takes place parallel on a huge data.

Then we use k-means clustering algorithm for working on clusters on the analyzing data as mentioned above. It is vector quantization which partition n observations on to k clusters.

Clusters are formed based on the normalized result of variance of overall stocks for a particular window, So now when one window is processed and have stored its clusters based on normalized result we could put our new batch in to some clusters which have already formed i.e. let’s say we have a cluster which has stocks with just say like within 3-5 units then new data which has same values get in to that cluster.

So now we could be able to predict other stocks in same cluster as in if we have Microsoft and Google in same cluster and we find that Microsoft stocks have raised so we can tell that even Google stocks also would be raised.

We show all of those changes in the stocks using some animation which would make it easier to understand.

Visualization tools which are available for spark could come in handy for the animation work for the changes in data.