International Journal of Current Engineering and Technology
©2015 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Prediction of BSE Stock Data using MapReduce K-Mean Cluster Algorithm

R. Lakshman Naik[†*] **and B. Manjula**[†]

[†]Department of Computer Science, Kakatiya University, Warangal, Telangana, India

Accepted 05 May 2015, Available online 12 May 2015, **Vol.5, No.3 (June 2015)**

Abstract

Bombay Stock Exchange (BSE) Limited, established in 1875 as the Native Share and Stock Brokers' Association is considered to be one of Asia's fastest stock exchanges and oldest stock exchange in the South Asia region. On 31 August 1957, the BSE became the first stock exchange to be recognized by the Indian Government under the Securities Contracts Regulation Act 1956. In this paper, we developed a novel framework that can achieve parallel time series prediction using Hadoop. By implementing the proposed framework, the system should be able to deal with massive amount of time series data, either regular or irregular. The proposed system can handle the optimization, parameter selection and also model combination through K-mean clustering. In this paper, experiment is carried to forecast the company's next bid accurately based on the other companies that have similar trend with it.

Keywords: *BSE, Hadoop, MapReduce, K-mean Clusters, Prediction, Stock;*

1. Introduction

Bombay Stock Exchange Limited, established in 1875 as the Native Share and Stock Brokers' Association is considered to be one of Asia's fastest stock exchanges, with a speed of 200 microseconds and one of India's leading exchange groups and the oldest stock exchange in the South Asia region. On 31 August 1957, the BSE became the first stock exchange to be recognized by the Indian Government under the Securities Contracts Regulation Act 1956. In 1980, the exchange moved to the Phiroze Jeejeebhoy Towers at Dalal Street, Fort area. BSE provides an efficient market, upholding the interests of the investors. Bombay Stock Exchange is the world's 10th largest stock market by market capitalization at $1.7 trillion as of 23 Jan 2015(WFE, 2015).

It operates a fully integrated exchange, offering the complete range of exchange-related services including trading, clearing, settlement and depository services that are traded on day-to-day. The Prices of the trade are determined by the market forces. The buyers and sellers quote the bid and ask prices and if prices are matched, in the case of BSE, by its automated trading. Due to the BSE trade is carried out every day, so there is a dynamic data for the BSE day-by-day. This big data need to be stored, processed and analyzed so that investors able to see the trend of the stock exchange, and they able to identify when and what stocks to buy and sell, by aware the track of upswings and downswings over the history of one's company according to the sector. For a long time, stock market prediction is long esteemed desire of investors, speculators, and industries. Although several studies investigated to predict price movements in stock market, financial time series too complex and noisy to forecast. Many researchers predicted the price movements in stock market using data mining techniques such as neural networks, artificial intelligence (AI) and Genetic algorithms (Naik R. Lakshman *et al*.,2012)(Manjua B. *et al*., 2012, 2011)during past decades.

The past decade has seen tremendous advances in application of parallel computing to various fields. New principles and standards are being created to address different requirements, and algorithms undergo many changes to become scalable. This requires not only an understanding of these algorithms, but of principles and techniques for parallel programming. To achieve an efficient approach for analyzing time series data in a parallel architecture, Hadoop is currently considered as the most appropriate option to try. Apache Hadoop, originally derived from the work of Google's MapReduce (Ronald, 2010), has become the standard way to address Big Data problems. MapReduce is used to process files on each node simultaneously and then aggregate their outputs to generate the final result. Hadoop allows for more scalable, cost effective, flexible and fault tolerant parallel programming (IBM, 2015). Despite all of its advantages, the original MapReduce

*Corresponding author: **R. Lakshman Naik** methodology of Hadoop is not ideally suited for time series analysis. This is due to the implicit dependencies among time series data observations (George *et al*., 2008). The best algorithm for performing prediction depends on the data and a considerable amount of expertise is required to design and configure a good predictor. In addition, the issue of predictor algorithm selection and optimization is critical, as is the implementation of an efficient platform that scales with time series data size.

The main aim of this paper is to develop a novel framework that can achieve parallel time series prediction using Hadoop. By implementing the proposed framework, the system should be able to deal with massive amount of time series data, either regular or irregular. Furthermore, the proposed system can handle the optimization, parameter selection and also model combination through K-mean clustering.

2. Related Work

The categorization of companies in the stock market is very useful for managers, investors, and policy makers. It can be performed based on several factors, such as the size of the companies, their annual profit, and the industry category. However, these features usually change over the course of time; thus, they are improper for categorization purposes. Industry-based categorization is also not preferable due to evidence that financial analysts are biased by industry categorization (Naik R. Lakshman *et al*.,2012)(Manjua B. *et al*., 2012, 2011) (P. Kruger *et al*., 2012). Identifying homogeneous groups of stocks where the movement in one market affects the stock prices in another market. The literature shows that the similarity of stock market in a country is affected by the movement of other stocks in that country or in other regions (D. Collins *et al*., 2003, A. Antonion, 2003, A. Masin *et al*., 2001). As a result, numerous studies have been performed on the recognition of co-movements among different countries (A. Rua *et al*., 2009, M. Graham *et al*., 2011, L. Norden *et al*., 2009). Most of these studies consider the co-movement of the stock market between different reigns or countries but not among different industries or companies in a stock market.

Previous research using AI techniques almost predicted the price of every trading day, week, and month. It is more important, however, to determine stock market timing, when to buy and sell stocks, than to predict the price movement for everyday because investors in stock market generally do not trade every day. If investors trade their stocks every day, they are charged to tremendous amount of fee for trade. Market timing is an investment strategy which is used for the purpose of obtaining the excess return. Traditionally excess return is achieved by switching between asset classes in anticipation of major turning points in stock market (Naik R. Lakshman *et al*., 2012, Waksman *et al*., 1997). In refer. (Manjua B. *et al*., 2012, 2011, Trippi *et al*., 1992) Executed daily prediction of up and down direction of S&P 500 Index Futures using ANN. Generating a composite recommendation for the current day's position. Input variables in this study were technical variables for the two-week period to the trading day, open, high, low, close price, open price and the price fifteen minutes after the market opening of the current trading day. The output variable was long or short recommendation. They performed composite rule generation procedure to generate rules for combining outputs of networks. They reported prediction accuracy was 45.3% - 52.8%.

In the time series literature review, (T. W. SR. Aghabozorgi *et al*., 2012), (Aghabozorgi *et al*., 2009), (Aghabozorgi and Wani *et al*., 2011), (Aghabozorgi and Ten, 2014), (Aghabozorgi and Wah 2014), (Aghabozorgi and Shirkhorshidi *et al*., 2014) (Nassirtonssi *et al*., 2014)(saeed *et al*., 2014), the author tries to cluster the time series of data efficiently by customer segmentation and developing a novel method for clustering time series incrementally based on its ability to accept new time series and also able to update the underlying clusters. While in the other time series literature review (Aghabozorgi and Wani *et al*., 2011), the author stated the significant problem of traditional clustering – defining prototype and explained the benefits of the proposed prototype by customer transaction clustering as well as present a prototype for time series clusters efficiency that can be updated based on a fuzzy concept through iterations.

There are several numbers of literatures that has been published about the Big Data and Hadoop as well as the stock market over the Internet. Among these publications, one of the literatures is about Evaluation of Data Processing Using MapReduce Framework in Cloud and Stand-Alone Computing (Daneshyar *et al*., 2012). This article described about the comparison of the data processing speed and time in the cloud computing environment and the stand alone system environment. To establish the experiment, the authors compare and concluded that the Linux environment is more suitable to develop the MapReduce than the windows as the windows had problem connection to a distributed cluster (Daneshyar *et al*., 2012).

3. Proposed System

It is determined that the Hadoop MapReduce is more suitable to install on the Linux environment than the windows environment. The k-Nearest Neighbor technique was implemented on a setup consisting four nodes connected over a private LAN. One node was used as a Namenode and Job Tracker the other three nodes were used as Datanodes and Task Trackers. All the four nodes had Intel i3 processors with 2.70 GHz and 4 GB memory. The Operating System running on all the 4 nodes was CentOS 6. The programming language used to code the k-Means Algorithm in MapReduce Implementation was JAVA. Apache Hadoop version 2.0 was installed on all the nodes and the single node and consequent multi-mode configurations were done according to the guide found at (Article1 & 2, 2015).

The install and configure the Hadoop MapReduce in the personal computer. Then storing and processing the BSE Bigdata stock market price components can be loaded into the Hadoop MapReduce and user needs to key in the Java coding to extract the desired data such as company name, date and closing bids of the BSE stock as the output.

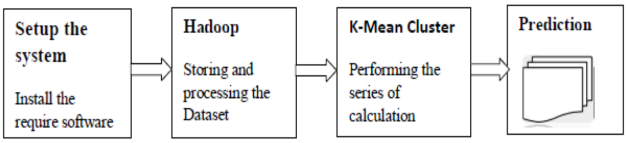

Setup the system
Install the require software
Hadoop
Storing and processing the Dataset
K-Mean Cluster
Performing the series of calculation
Prediction

Fig.1: framework for proposed system

4. Prediction of BSE Stocks

The extracted output from the Hadoop MapReduce will be passing to the k-means clustering for further analysis by performing series of calculation on the closing bids, to determine the similarities among companies. Conventional clustering and similarity measures which are applied to static data are not practical for the time-series datasets because they are essentially not designed for time-series data. Hence, various techniques have been recommended for the clustering of time-series data. Most of them try to customize the existing conventional clustering algorithms such that they become compatible with the nature of time-series data. In these cases, usually the distance measure is modified to be well-matched with the time-series data

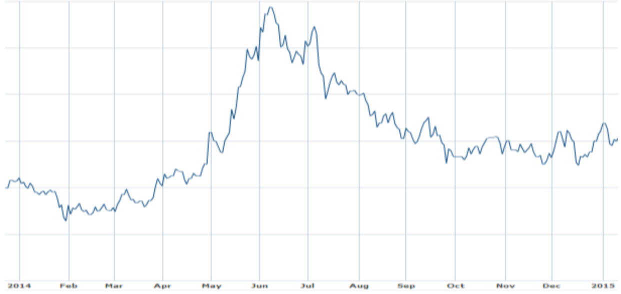


Fig. 2: Simulated ARIMA Series BSE components (From Jan, 2014 to Jan, 2015)

The starting of the identification stage is to specify the input data set in the k-means clustering. The input data set is the BSE components. Then use an identify statement to read the BSE close bids in time series and plot a graph. The graph that has been plotted is shown in the figure 2 above and the table 1 of the data below shows the example of BSE components data set.

Table 1: Company IOB components data set

Ticker	Date	Open	High	Low	Close	Volume
IOB	01-01-15	0.62	0.64	0.6	0.63	533
IOB	12-01-14	0.57	0.63	0.5	0.62	284
IOB	11-03-14	0.58	0.61	0.5	0.57	203
IOB	10-01-14	0.56	0.61	0.6	0.58	210
IOB	09-01-14	0.6	0.66	0.5	0.56	227
IOB	08-01-14	0.69	0.72	0.6	0.6	144
IOB	07-01-14	0.82	0.85	0.7	0.7	278
IOB	06-02-14	0.78	0.89	0.8	0.81	407
IOB	05-01-14	0.61	0.86	0.6	0.77	0
IOB	04-01-14	0.51	0.64	0.5	0.61	615
IOB	03-03-14	0.45	0.53	0.4	0.51	111
IOB	02-03-14	0.46	0.47	0.4	0.45	63
IOB	01-01-14	0.51	0.54	0.4	0.46	107

The estimate statement next prints a table of correlations of the parameter wanted, as shown on the table 2 below.

Table 2: Company IOB close bids and company INDIANB close bids are extracted

Ticker	Date	Close
IOB	01-01-15	0.63
IOB	12-01-14	0.62
IOB	11-03-14	0.57
IOB	10-01-14	0.58
IOB	09-01-14	0.56
IOB	08-01-14	0.6
IOB	07-01-14	0.7
IOB	06-02-14	0.81
IOB	05-01-14	0.77
IOB	04-01-14	0.61
IOB	03-03-14	0.51
IOB	02-03-14	0.45
IOB	01-01-14	0.46

Ticker	Date	Close
INDIANB	01-01-15	2.12
INDIANB	12-01-14	2.18
INDIANB	11-03-14	1.89
INDIANB	10-01-14	1.66
INDIANB	09-01-14	1.54
INDIANB	08-01-14	1.36
INDIANB	07-01-14	1.48
INDIANB	06-02-14	1.83
INDIANB	05-01-14	1.7
INDIANB	04-01-14	1.28
INDIANB	03-03-14	1.15
INDIANB	02-03-14	0.88
INDIANB	01-01-14	0.99

When the output is extracted from the Hadoop MapReduce, then use formulas to perform the calculation to calculate the entire closing bids distances between companies.

$$\text{Distance } (t_1, t_2) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \ldots\ldots + (y_n - x_n)^2}$$

$$= \sqrt{\begin{array}{l}(2.12-0.63)^2 + (2.18-0.62)^2 + (1.89-0.57)^2 + (1.66-0.58)^2 + (1.54-0.56)^2 \\ + (1.36-0.6)^2 + (1.48-0.7)^2 + (1.83-0.81)^2 = (1.7-0.77)^2 + (1.28-0.61)^2 \\ + (1.15-0.51)^2 = (0.88-0.45)^2 + (0.99-0.46)^2\end{array}}$$

$$= \sqrt{12.9385} = 3.597$$

When the distances between the two companies are known, next is to normalize the distance to the values between 0 and 1 for the standardization purpose.

$$\text{Normalized Distance} = \frac{\text{Distance}(t_1, t_2)}{\sum(x_1 + x_2 + ... + x_n)}$$

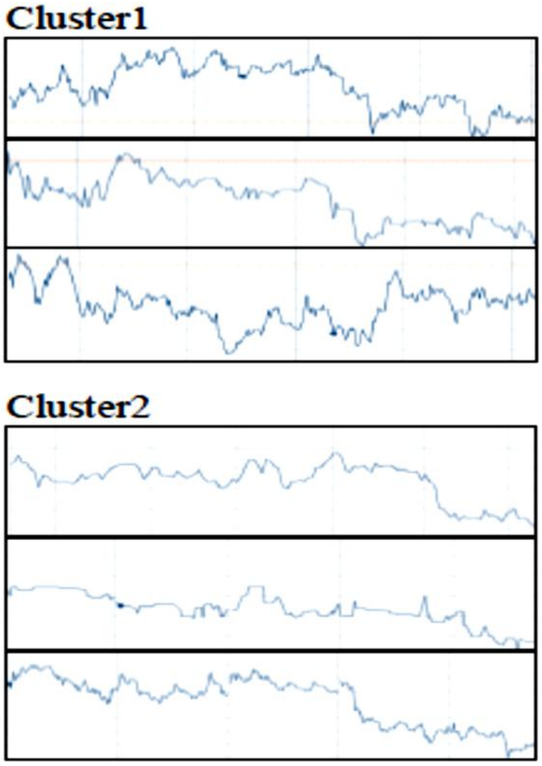
$$= \frac{3.597}{(0.63 + 0.62 + 0.57 + 0.58 + 0.56 + 0.6 + 0.7 + 0.81 + 0.77 + 0.61 + 0.51 + 0.45 + 0.46)}$$

$$= \frac{3.597}{7.87} = 0.457$$

When the distance values between companies had been standardized, the similarities between the companies can be determined.

Similarities (t_1, t_2) =1 – Normalized Distance=1– 0.457=0.54

After a series of calculation, it can be seen that the similarities between both company IOB and company INDIANB are 0.54. Therefore, it can be concluded that the smaller the similarities (t_1, t_2) between both companies, the both companies' trends are similar, in contrast, the larger the similarities (t_1, t_2) between both companies, the both companies' trends are not similar. From the similarities (t_1, t_2) between company IOB and company INDIANB, it can be concluded that the both companies are neither similar nor not similar. We expect to see the clusters as shown in the figure 3 below.



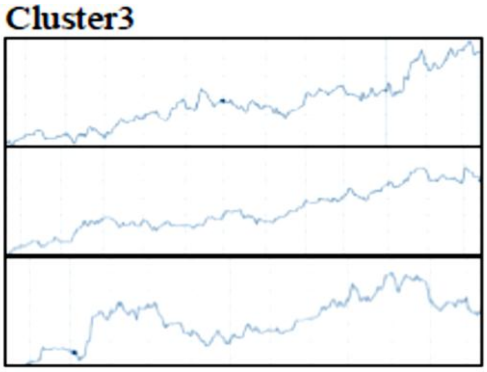


Fig.3: Three sample stock market of three clusters of BSE datasets

That has the most similar trend with Syndibank are the Axis bank (most similar trend), followed by the Indian bank (similar trend), Public bank (similar trend) and Hdfc (less similar trend). To produce forecast, company X's next bid will be predicted based on the other company such as company Y that has most similar trend with company X because they have the similarity shape of the stock price or they are co-movement that move together in the same trend. The figure 4 below shoes the daily stock price index prediction of BSE in the graph form.

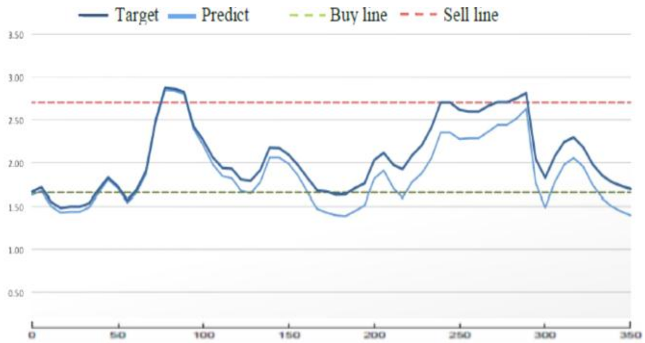


Fig. 4: Daily stock Price Index Prediction of BSE

Conclusion

In this paper, BSE components as an input of Hadoop MapReduce, the output is the company's closing bids, then passing to the ARIMA model for the series of calculation, and determined the companies' similarities. A simple and clear methodology is used to investigate the similar trends of the BSE for companies. From the calculation, we found out that the series of the calculation should be integrated into one algorithm to facilitate the calculation, and it should be insert it into the Hadoop MapReduce's reducer part; to minimize the time and get the accurate output in the shortest possible time. In this paper, prediction techniques are useful to the investors in the future as it able to forecast the company's next bid accurately based on the other companies that have similar trend with it.

Reference

WFE, 2015, (World Federation of Exchanges) *URL:http:// www.world-exchanges.org/ statistics/ monthly-reports*

Naik, R. Lakshman, *et al*., 2012, Prediction of Stock Market Index Using Neural Networks: An Empirical Study of BSE. *European Journal of Business and Management* 4.12, pp. 60-71.

Naik, R. Lakshman, *et al*., 2012, Prediction of Stock Market Index Using Genetic Algorithm. *Computer Engineering and Intelligent Systems* 3.7, pp. 162-171.

Manjula, B., R. Lakshman Naik, and S. S. V. N. Sarma, 2012, Tracking the Trends of Financial Applications Using Genetic Algorithm. *International Journal of Computer Applications* 48.16, pp. 36-40.

Manjula, B., *et al*., 2011, Stock Prediction using Neural Network. *International Journal of Advanced Engineering Sciences and Technologies* 10.1, pp. 13-18.

Naik, R. Lakshman, D. Ramesh, and B. Manjula, 2012, Instances Selection using Advance Data Mining Techniques. *International Journal of Computer Engineering & Technology (IJCET)* 3.2, pp. 47-53.

Ronald C Taylor, 2010, An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *In Proceedings of the 11th Annual Bioin-formatics Open Source Conference* (BOSC),

IBM Software, 2015. What is Hadoop? *URL www.ibm.com/software/data/infosphere/hadoop*

George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel., 2008, Time Series Analysis: Forecasting and Control *Wiley Series in Probability and Statistics,* 4th edition, June 30.

P. Krüger, A. Landier, and D. Thesmar, 2012, Categorization Bias in the Stock Market, *Available SSRN 2034204*.

D. Collins and N. Biekpe, 2003, Contagion and interdependence in African stock markets, *South African J. Econ.*, vol. 71, no. 1, pp. 181–194,

A. Antoniou, 2003, Modelling international price relationships and interdependencies between the stock index and stock index futures markets of three EU countries: a multivariate, *J. Business, Financ. Account.*, vol. 30, pp. 645–667,

A. Masih and R. Masih, 2001, Dynamic modeling of stock market interdependencies: an empirical investigation of Australia and the Asian NICs, *Rev. Pacific Basin Financ. Mark. Policies*, vol. 4, no. 2, pp. 1323–9244.

A. Rua and L. Nunes, 2009, International comovement of stock market returns: A wavelet analysis, *J. Empir. Financ.*, vol. 16, no. 4, pp. 632–639.

M. Graham and J. Nikkinen, 2011, Co-movement of the Finnish and international stock markets: a wavelet analysis, *Eur. J. Financ.*, vol. 17, no. 5, pp. 409–425.

L. Norden and M. Weber, 2009, The Coâ€•movement of Credit Default Swap, Bond and Stock Markets: an Empirical Analysis, *Eur. Financ. Manag.*, vol. 15, no. 3, pp. 529–562.

Waksman, G., *et al*. 1997, Market timing on the Johannesburg Stock Exchange using derivative instruments, *Omega, International Journal of Management Science*, Vol.25, No. 1, pp. 81-91.

Trippi, R. R. and DeSieno D., 1992, Trading equity index futures with a neural network, *The Journal of Portfolio Management*.

T. W. Sr Aghabozorgi, Mr Saybani, 2012, Incremental Clustering of Time-Series by Fuzzy Clustering, vol. 688, pp. 671–688.

S. Aghabozorgi and T. Y. Wah, 2009, Dynamic Modeling by Usage Data for Personalization Systems, *2009 13th Int. Conf. Inf. Vis.*, pp. 450–455.

S. Aghabozorgi and T. Y. Wah, 2009, Using Incremental Fuzzy Clustering to Web Usage Mining, in *2009 International Conference of Soft Computing and Pattern Recognition*, pp. 653–658.

S. Aghabozorgi, T. Y. Wah, A. Amini, and M. R. Saybani, 2011, A new approach to present prototypes in clustering of time series, in *The 7th International Conference of Data Mining*, vol. 28, no. 4, pp. 214–220.

S. Aghabozorgi and Y. Teh, 2014, Clustering of Large Time-Series Datasets, *J. Intell. Data Anal.*, vol. 18, no. 5,

S. Aghabozorgi and T. Wah, 2014, Effective Clustering of Time-Series Data Using FCM., *Int. J. Mach. Learn. Comput.*, vol. 4, no. 2, pp. 170–176

S. Aghabozorgi, A. S. Shirkhorshidi, T. Hoda Soltanian, U. Herawan, and T. Y. Wah, 2014, Spatial and Temporal Clustering of Air Pollution in Malaysia: A Review, in *International Conference on Agriculture, Environment and Biological Sciences*, pp. 213–219.

A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. N. C. Ling, 2014, Text Mining for Market Prediction: A Systematic Review, *Expert Syst. Appl*.

Saeed Aghabozorgi and T. Y. Wah, 2014. Shape-based Clustering of Time Series Data, *J. Intell. Data Anal.*, vol. 18, no. 5.

S. Daneshyar and A. Patel, 2012, Evaluation of Data Processing Using MapReduce Framework in Cloud and Stand - Alone Computing, *Int. J. Distrib. Parallel Syst.*, vol. 3, no. 6, pp. 51–63.

Article1, 2015: Single Node Cluster Setup at: http://www.michaelnoll.com/tutorials/running-hadoop-on-ubuntu-linux-single-nodecluster/

Article2, 2015: single node cluster setup http://www.cloudera.com/content/cloudera/en/documentation/cdh4/v4-2-0/CDH4-Quick-Start/CDH4-Quick-Start.html