International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 4, April 2016

# Stock Market Prediction and Analysis using Hadoop

[1] Pawar Manali V. [2] Pathwe PrajaktaK [3] Pawar Suhas S [4] Thakre Sanjay [5] Poman Sushant S
[1][2][3][4] Student [5] Assistant Professor
[1][2][3][4][5] Department of Computer Engineering, JSPM'S Rajashri Shahu Collegeof Engineering Tathawade Pune, SavitribaiPhule University Pune, India.
[1]Manalipawar97@gmail.com [2] prajaktapathwe1994@gmail.com [3] shspwr6@gmail.com
[4] mail2sbt@gmail.com [5] sushantpoman94@gmail.com

***Abstract:*** **Staying updated with latest business news from Indian stock market and making a profitable investment is difficult. The purpose of stock market is to facilitate the exchange of securities between buyers and sellers. Our system provides exchanges with information on the listed securities, facilitating price discovery. Existing system has drawback that they cannot provide prediction over share's threshold with performance parameter over continuously growing market's data, however, our proposed system will enable prediction module and will be implemented parallel on Hadoop framework with MapReduce which will not only enhance the speed up factor with performance enhancement but also automate profitable investment and additionally enabling inexperienced investors to make a sound investment.**

***General Term*****: Naive Bayes, yahoo finance, HDFS.**

***Keywords*****: Big data, stock market prediction, Hadoop, Map Reduce, automation**

## I. INTRODUCTION

### *1.1 Stock Market*

Stock Market or equity market is one place where buyers and sellers meet and invest their securities i.e. money to make profit. Investing in shares for company buyer tends to buy little part of that company. Company however makes profit due to this investment. Equity Market is previously implemented with ANN (Artificial Neural Network) which increases system delay overhead because of its serial architecture. Our proposed system removes this overhead by parallel architecture with its interface of Hadoop enhancing its performance.

### *1.2 Hadoop*

Hadoop can be one of the solutions for delay in performance issue as Hadoop enables parallel processing with HDFS (Hadoop Distributed File System). Hadoop is deployed over cloud which has maximum processing capacity as many secondary nodes can be created. The racking property of Hadoop enables fault tolerance with yarn scheduling over HACE (Heterogeneous, Autonomous, Complex and Evolving Relationships in dataset) database with security. Our database will be fetched online from yahoo finance and historical data will also be maintained at the admin module.

## II. PROBLEM DESCRIPTION

To implement stock market prediction system using Hadoop with Naïve Bayes probabilistic classifier in co-relation with HACE theorem to increase the performance based on factors like company economic growth, inflation, unemployment, earnings.

An accurate prediction of stock market movement is crucial for investors to make effective market trading decisions. However, because of the high fluctuations of the stock market, it is difficult to reveal the inside law of stock market movement. To overcome such difficulties, data mining techniques have been introduced and applied for this financial prediction. This study we attempt to develop a stock market prediction system for sound investment using Naïve Bayes probabilistic classifier. Ten microeconomic variables were chosen as inputs of the proposed model.

## III. RELATED WORK

The ideal system is the one which reflects the change in market as soon as they happen i.e. the stock movement [7]. In our system we will fetch values from News feed. Managing huge amount data i.e. HACE [1]. Data will be from heterogeneous sources and complex.

Parallel architecture over serial [2]. In our system we will adapt the parallel architecture with HDFS. Artificial neural network, genetic algorithm, working & training of ANN [13]. From this paper we have referred -ANN (Artificial Neural Network), Neurons, pattern recognition. A framework - PESMiner which scales data in parallel and sequential manner at scale. Whereas most existing sequential mining algorithms can only find sequential orders of temporal events [9]. From this paper we have referred - Parallel and quantitative mining of sequential patterns at scale. Hadoop DistributedFileSystemthat allows the distributed processing and fast access to large data sets on distributed storage platforms [8].From this paper we have referred -Hadoop Distributed File System. Parallel Processing Technique (PPT) that characterizes the features of big data revolution, reduces complexity, and processes this data perspective [10].From this paper we have referred –the HACE (heterogeneous, autonomous, complex and evolving relationships in dataset) concept. Applicable parallel concept and that can be applied to many data mining algorithm [11].From this paper we have referred-map-reduce, performance enhancement. MapReduceissues and challenges in handling Big Data with the objective of providing an overview of the field [12]. From this paper we have referred-map reduce, challenges of big data. Nature of big data and various sources of Big Data [3]. From this paper we have referred: - 1.Types of big data and challenges in big data. 2. Three Vs in Big Data (Variety, Velocity and Volume).Evolution of DFS from the history, current state of the art design and implementation of the DFS and replica placement in Hadoop DFS [4]. From this paper we have referred: - 1.Distributed File System, challenges and Design issues 2.Comparison of Distributed File System and HDFS architecture 3. Anatomy of HDFS File writes operation using pipelined and parallel replication approach. Evaluating the performance of a Hadoop single node cluster with respect to Big Data [5]. From this paper we have referred: - Apriori based Association rule mining algorithm was used to find the frequent patterns and then their rules.

Hadoop Distributed File System that allows the distributed processing and fast access to large data sets on distributed storage platforms [6].From this paper we have referred-Hadoop Distributed File System. Parallel architecture over serial [2]. In our system we will adapt the parallel architecture with HDFS. Artificial neural network, genetic algorithm, working & training of ANN [13]. From this paper we have referred -ANN (Artificial Neural Network), Neurons, pattern recognition. A framework - PESMiner which allows parallel and quantitative mining of sequential patterns at scale. Whereas most existing sequential mining algorithms can only find sequential orders of temporal events [9]. From this paper we have referred - Parallel and quantitative mining of sequential patterns at scale. Hadoop DistributedFileSystemthat allows the distributed processing and fast access to large data sets on distributed storage platforms [8].From this paper we hav referred -Hadoop Distributed File System. Parallel Processing Technique (PPT) that characterizes the features of big data revolution, reduces complexity, and proposes a big data processing model from the data mining perspective [10].From this paper we have referred -Complexity, data mining, heterogeneity, autonomous sources volume, velocity, variety, variability of big data. Applicable parallel programming method, one that is easily applied to many different learning algorithms [11].From this paper we have referred-map-reduce, performance enhancement. Map Reduce issues and challenges in handling Big Data with the objective of providing an overview of the field [12]. From this paper we have referred-map reduce, challenges of big data. Nature of big data and various sources of Big Data [3]. From this paper we have referred: - 1.Types of big data and challenges in big data. 2. Three Vs in Big Data (Variety, Velocity and Volume).Evolution of DFS from the history, current state of the art design and implementation of the DFS and replica placement in Hadoop DFS [4]. From this paper we have referred: - 1.Distributed File System, challenges and Design issues 2.Comparison of Distributed File System and HDFS architecture 3. Anatomy of HDFS File writes operation using Pipelined and parallel replication approach. Evaluating the performance of a Hadoop single node cluster with respect to Big Data [5]. From this paper we have referred: - Apriori based Association algorithm was used to find the frequent patterns and then their rules. Hadoop Distributed File System that allows the distributed processing in parallel and single scan to large data sets on HDFS platforms [6].From this paper we have referred- Hadoop Distributed File System.

## IV. SYSTEM ARCHITECTURE

In this architecture the main components are User, Admin, Automation for purchase, Automation for sale, Prediction.

1. User registers his details.
2. Admin authenticates user's details.
3. User uploads his/her data.
4. Server verifies the Username and stores it in the database.
5. Prediction analysis is displayed with comparison.
6. User sets his/her threshold value for shares.
7. According to profit or loss alert will be given and preferred action will be taken.

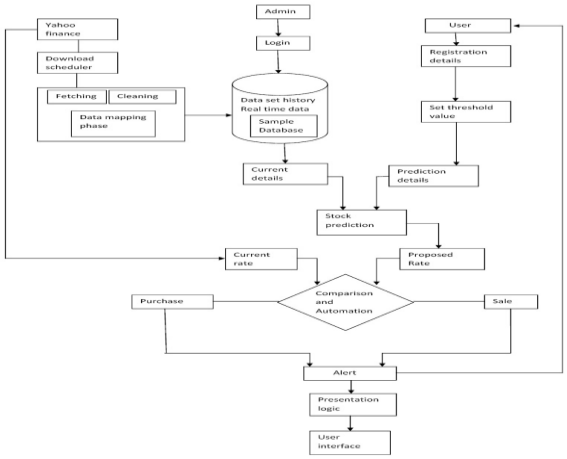


*Fig 1: System Architecture*

### *A. Modules*

Module 1: Register New User: A new user needs to enter all the required details to get registered to our application. Registration includes all personal details of user.
Module 2: Admin will register company:
Admin will enter all company details.
Module3: Prediction: By comparing current details and threshold value.
Module 4: Automation for purchase or
Automation for sale: Value of shares will be calculated with respect to real time data and it will be notified for purchase or sale.

## V. PROPOSED SCHEME

Our proposed architecture is collaborated with Hadoop with Naïve Bayes algorithm.

### *A. NBE learning algorithm*

1. Fetch number of day's previous data from the dataset as entered by user.
2. Mean calculationof parameters to be considered (coloumnwise).
3. Fetch today's Open value from the dataset (maybe online).
4. Classify mean coloumn wise based on above and below the mean value calculated in step 1.
5. Calculate Mean for specific company given by user and classify according to step 3.
6. Bayesian classifiers use Bayes theorem, which says,

$P(c_j|d) = P(d|c_j) * P(c_j) / P(d)$

Where,

$P(c_j|d)$ = probability of instance d being in class $c_j$.

$P(d|c_j)$ = probability of generating instance d given class $c_j$.

$P(d)$ = probability of instant d occurring.

$P(c_j)$ = probability of occurrence of class $c_j$.

7. Calculate threshold,

Max. Value = Open value + P (above|Company)
Min. value = Open value - P (above|Company)

- ❖ As many parameters will be considered more the accuracy will be achieved. For our system Minimum value is safer for setting of threshold.
- ❖ Includes Fast to train i.e. Single scan, Fast to classify.
- ❖ Handles real and discrete data.
- ❖ Handles streaming data well.
- ❖ Alert messages will be sent to user according to the equity market movement.

### *B.* Proof of Concept

*Table 1. Dataset*

| Company | Open | High | Low | Close |
|---|---|---|---|---|
| TCS | 30.79 | 31.19 | 30.66 | 30.96 |
| HP | 30.80 | 31.20 | 30.35 | 30.85 |
| DELL | 28.62 | 30.70 | 28.43 | 30.71 |
| DELL | 28.95 | 29.00 | 28.44 | 28.91 |
| HP | 28.65 | 29.11 | 28.49 | 28.91 |
| TCS | 28.58 | 29.23 | 27.85 | 28.26 |
| HP | 29.03 | 29.22 | 27.20 | 27.60 |
| DELL | 29.66 | 29.71 | 28.91 | 29.13 |
| HP | 29.47 | 29.57 | 28.85 | 29.34 |
| TCS | 30.56 | 30.57 | 29.63 | 29.74 |

(Here we calculate only for Open parameters from above dataset)
Suppose new Open value for TCS is 30.85 … (1)
Mean = 29.51
We classify above Open values (above and below)

*Table 2. Classification according to Mean*

| Open | |
|---|---|
| **Below** | **Above** |
| **28.62** | **30.79** |
| **28.95** | **30.80** |
| **28.65** | **29.66** |
| **28.58** | **30.56** |
| **29.03** | |
| **29.47** | |

Now,
Mean (TCS) = (30.79+28.58+30.56)/3
=29.97
We classify above Open values (above and below) for TCS.

*Table 3. Classification according to Mean (TCS)*

| Open | |
|---|---|
| **Below** | **Above** |
| **28.58** | **30.79** |
| | **30.56** |

Our Open value falls in "above" category from (1)
Therefore, according to formula
P (h|X) = P (X|h) *P(X) /P (h)
Where,
P (h) = Posterior hypothesis (In our example we consider Question tuple)
P(X) = Data set.
P (above|TCS) = P (TCS|above) * P(above)/ P(TCS)
= (2/3) * (4/10) / (3/10)
= 0.88

### *C. Calculation of threshold,*

Therefore, Maximum value for profitable investment can be calculated as,
Open value + P (above|TCS) = 30.85+0.88
= 31.73
Minimum value for profitable investment can be calculated as,
Open value – P (below|TCS) = 30.85-0.66
= 30.19
However, for safer investment user should set minimum value or the average of both as threshold.
Similar way above High, Low, Close parameters can be calculated.

## VI. MATHEMATICAL MODEL

Z={S, E, X, Y, F, DD, NDD, Success, Failure}.
Where,
**S: Initial/Start State**
After successful registration user should log in his account.
**E: End state/Final State**
Setting his threshold value according to system predicted value.
**X: Input for prediction**.
User enters company name and number of days for prediction request.
**F: Functions**
P (h|X) = P (X|h) *P(X) /P (h)
Where,
P (h) = Posterior hypothesis (In our example we consider Question tuple)
P(X) = Data set.
**DD: Deterministic Data**
Number of previous day's user requests prediction basis for.
Name of the company user wish to invest his money.
Rough amount he wishes to invest.
**NDD: Non-Deterministic Data.**
Open value for every new day.
**Success: Desired output is generated**
User gets the predicted value within his budget and he makes profitable transaction.
**Failure: Desired output is not generated**.
System fails to predict or user makes loss.

## VII. CONCLUSION

Our proposed system for stock market prediction and analysis will be implemented in parallel with Hadoop and Naïve Bayes algorithm. Prediction module will enable stakeholder to make transparent investment and accuracy up to 70-80% with increased rate of performance. Features of

Map Reduce will help to overcome the current issues like delay and make the system fault tolerant.

In this paper we have proposed prediction and performance analysis on the basis of manually entered value and system calculated value for tomorrow's shares. Comparison of these two values will be analyzed for system accuracy. Inexperienced investors can make a sound investment by analyzing the output. Therefore, the system will not only enhance the speed up factor but also automate profitable investment and additionally

enabling inexperienced investors to make a sound investment.

## ACKNOWLEDGMENTS

Our sincere thanks to Professor S.B. Thakare who guided us in selection of algorithm and technical support towards developing this system.

## REFERENCES

[1]. Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, Data mining with big data, senior member, ieee, ieee transactions on knowledge and data engineering, vol. 26, no. 1, January 2014

[2]. Kushagrasahu, revatipawar, sonalitilekar, reshmasatpute, Stock exchange forecasting using hadoop map-reduce technique,Department of computer, aissmssioit, pune, india, International Journal of advancement in Research and Technology,Volume 2, Issue4, April2013 ISSN 22787763

[3]. Bharti Thakur and Manish Mann, Data Mining for Big Data: A Review, Computer Science Department LRIET, Solan (H.P), International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

[4]. Neha M Patel, Narendra M Patel, Mosin I Hasan and Mayur M Patel. Improving Data Transfer Rate and Throughput of HDFS using Efficient Replica Placement. International Journal of Computer Applications 86(2) : 4-7, January 2014. Published by Foundation of Computer Science, New York, USA.

[5]. A.Asbern, P.Asha, Performance Evaluation Of Association Mining In Hadoop Single Node Cluster With Big Data, Department of Computer Science, Sathyabama University, Chennai, India, 2015 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[6]. Daniel Lowd, Pedro Domingos, “Naive Bayes Models for Probability Estimation”, Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, USA.

[7] Stock Price Prediction Using News Articles, Qicheng Ma CS224n.

[8] Tomasic, I. Ugovsek, J. ; Rashkovska, A. ; Trobec, R.Multicluster Hadoop Distributed File System IEEE Conference on 21-25 May 2015.

[9] GuangchenRaun,HuiZhang,Plale B, Parallel and quantitative sequential pattern mining for large-scale interval-based temporal data,Big Data ,2014 IEEE Conference.

[10] J.JosephaMenandas,J.JakkulinJoshi,Data Mining with Parallel Processing Technique for Complexity Reduction and Characterization of Big Data,Global Journal Of Advanced Research.

[11] Scholkopf ,B Platt J,Hofmann T Map-Reduce for Machine Learning on Multicore.

[12] Ms. Sonali. B. Maind Ms. Priyanka Wankar Research Paper on Basic of Artificial Neural Network International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 1.

[13] Grolinger, K. Hayes, M. ; Higashino, W.A. ; L'Heureux, A. Challenges for MapReduce in Big Data 2014 IEEE World Congress on June 27 2014-July 2 2014.