DEVELOPING PREDICTION MODEL FOR STOCK EXCHANGE DATA SET USING HADOOP MAP REDUCE TECHNIQUE

Mrs. Lathika J Shetty[1], Ms. Shetty Mamatha Gopal[2]

[1] *Computer Science & Engineering, Sahyadri College of Engineering and Management, Mangalore, Karnataka, India*
[2] *Computer Science & Engineering, Sahyadri College of Engineering and Management, Mangalore, Karnataka, India*

---***---

ABSTRACT- *Stock Market has high profit and high risk features which tells why its prediction must be close to accurate. The main issue about such data sets is that these are very complex nonlinear functions and can only be learnt by a data mining methods to recognize the future market trend. Companies provide daily statistics of their market trend and in time, generating a great deal of information which is dumped into their database. Forecasting stock price is an important task for investment and financial decision making process. This is considered as one of the biggest challenges. In this paper the proposed system goal is to develop a prediction model using MapReduce with the help of time-series analysis in Hadoop which can be used to predict the future stock closing price. This system will be a Hadoop based Stock Prediction Model generator for the people interested to know the future market trend of a particular company. The target clients are shareholders and the company officials. The developed model can be deployed and used by companies and shareholders to adjust their strategies based on the results of the analysis done.*

Key Words: *Hadoop, HDFS, Map reduce, prediction model, stock exchange data set, prediction model*

1. INTRODUCTION

Stock Market has high profit and high risk features which tells why its prediction must be close to accurate. The main issues about such data sets are that these are very complex nonlinear functions and can only be learnt by a data mining methods to recognize the future market trend. Companies provide daily statistics of their market trend and in time, generating a great deal of information which is dumped into their database. The project goal is to develop a prediction model using Map Reduce technique with the help of time-series analysis in Hadoop which can be used to predict the future stock closing price.

Hadoop is a Java based open source framework which uses simple programming models to allow storing and processing of Big Data in a distributed computing environment across clusters of computers. It is a part of the Apache Software foundation. It provides a design to scale up from single servers to thousands of machines and each offering local computation and storage. One of the most efficient solutions for processing of large data sets is Hadoop.

The main components of Hadoop framework is HDFS and Map reduce. HDFS (Hadoop Distributed File System) is a block structured distributed file system which holds large amount of Big Data. It provides high throughput access to application data. HDFS uses master/slave architecture. Master consists of a single name node that manages the file system metadata and one or more slave data nodes that store the actual data. Hadoop MapReduce-Hadoop runs applications using map reduce algorithm which is a programming framework for distributed computing. Here divide and conquer method is used to break large complex data.

A time series is a sequence of numerical data points in successive order, usually occurring in uniform intervals. In general, a time series can be defined as a sequence of numbers collected at regular intervals over a period of time.

A stock is a share in the ownership of a company. Stock represents a claim on the company's assets and earnings. As you acquire more stock, your ownership stake in the company becomes greater.

2. LITERATURE SURVEY

Various contributions are made in the direction of building prediction models for historical stock exchange data set. Works carried out in this field are as follows:

[1] Time Series Forecasting Of Nifty Stock Market(NSE) Using Weka by Raj Kumar, Anil Balara in 2010, JRPS: Here the researchers utilized Weka 3.7.8 tool to obtain more accurate stock prediction price using time series forecasting package of weka for the work. Weka tool has been used and it analyzed and compared results by plot graphs.

Comparision among various prediction functions were made throughout the work. After comparision it is found that the SMO regression function predicted the stock prize of NSE more accurately than the other prediction functions such as Gaussian process, linear regression.

[2]Applying Data Mining Techniques to Stock Market Analysis by Gabriel Fiol-Roig, Margaret Miro-Julia, and Andreu Pere Isern-Deya in 2010, Springer:

Here the researchers viewed the stock market analysis as an artificial intelligence problem. First, data mining techniques will be used to evaluate past stock prices and acquire useful knowledge through the calculation of some financial indicators. Next they applied artificial intelligence strategies to construct decision making trees and predictions are made.

[3] A Decision Tree- Rough Set Hybrid System for Stock Market Trend Prediction by Binoy.B.Nair, V.P Mohandas and N. R. Sakthivel in 2010, IJCA

They have used hybrid decision tree rough set based system for prediction. First, the relevant feturese selected using decision tree and then a rough set based classifier is used for stock price trend forecasting. Overall, it is based on soft computing technique.

[4] The use of data mining and neural networks for forecasting stock market returns by Thawronwong in 2012, Journal of Computers:

Here in this study an attempt has been made to investigate the predictive power of financial and economic variables by adopting the analysis technique in machine learning for data mining. Selecting the appropriate variable for analysis is the major task.

[5] Machine Learning in Stock price trend forecasting, by Yuqing Dai, Yuning Zhang in 2012:

Here the researchers have used Machine learning techniques to obtain their result. Machine Learning techniques are very popular due to the capacity of identifying stock trend from massive amounts of data that capture the underlying stock price dynamics. In this work, they used supervised learning methods to stock price trend forecasting.

[6] Predicting stock prices using data mining techniques by Qasem a. Al-radaideh, Adel abu assaf and Eman in 2013,ACIT:

Here the method is based on decision tree classifier which is one of the data mining technique. Also the CRISP-DM(Cross-Industry Standard process for data mining) is applied over the historical stock exchange data to predict the future stock market price .

[7] Stock Price Prediction Using *K*-Nearest Neighbor (*k*NN) Algorithm by Khalid Alkhatib Hassan Najadat Ismail Hmeidi Mohammed K. Ali Shatnawi in 2013 , IJBHT :

In this methodology, the researchers applied k-nearest neighbor algorithm and non-linear regression approach in order to predict stock prices. In classification approaches, a data set is divided into two sub sets such as training data set and testing set. *k*NN algorithm uses similarity metrics to compare a given test entity with the training data set to aid in the prediction process.

[8] Clustering-Classification Based Prediction of Stock Market Future Prediction by Abhishek Gupta, Dr.Samidha D Sharma in 2014, IJSCIT

The methodology implemented here for the prediction of stock market is Clustering classisfication based prediction such as applying clustering algorithm such as K-means and decision tree algorithm. The particular methods has been applied in term of two stages and prediction is made.

[9] Predicting Closing Stock Price using Artificial Neural Network and Adaptive Neuro Fuzzy Inference System (ANFIS) by Mustain Billah, Sajjad Waheed andAbu Hanifa in 2015 , IJCA :

An attempt has been made in this study to discover the efiicient soft computing technique. Two techninques such as ANN(Artificial Neural Network) and ANFIS(Adaptive Neuro Fuzzy Inference System) Has been applied for the historical stock exchange dataset. The outcomes of both the method is compared and it is found that the ANFIS technique is more efficient and effective in predicting the future stock closing price.

Overview:

All the above researchers have been successful in analyzing the dataset related to stock exchange using various approaches like Neural Network, Regression , clustering, classification and different algorithms like K-means clustering. Most of them used Neural Network approach to build the prediction model but an attempt to build prediction model using hadoop is not made yet.In this project an attempt to analyze the dataset using Hadoop map reduce technique and time series algorithm is done.

3. SYSTEM ANALYSIS

3.1 Existing system

As discussed many researchers have been successful in analyzing the dataset related to stock exchange using various approaches like Regression based datamining , clustering, classification and different algorithms like K-means clustering. Most of them used Nueral Network approach to build the prediction model . Thus all these prediction models were developed using data mining tools like Oracle data Miner, Weka etc.

3.2Limitations of Existing Systems

All the researchers have been successful in analyzing the stock data set and developing good prediction models. But most of them used tools like weka and oracle data miner and few of them used tools like Tanagra. But these methods aren't efficient. Also an attempt to build prediction model using hadoop is not made yet. Hadoop can work using a cluster of nodes to work on a large unstructured data set in parallel and produce faster results.

3.3Proposed System

The proposed system (fig 3.1) focuses on building a prediction model using Hadoop map reduce technique. The Hadoop Software library gives a major advantage of distributed processing of bulky data sets and it also provides high availability of data. Hence, Hadoop has been taken as a framework for developing the prediction model. In this system, a company's daily stock data set is chosen as a training data set, , and after analyzing the data set completely, a prediction model is developed, it can be used to predict how the stock will be for the next quarter.

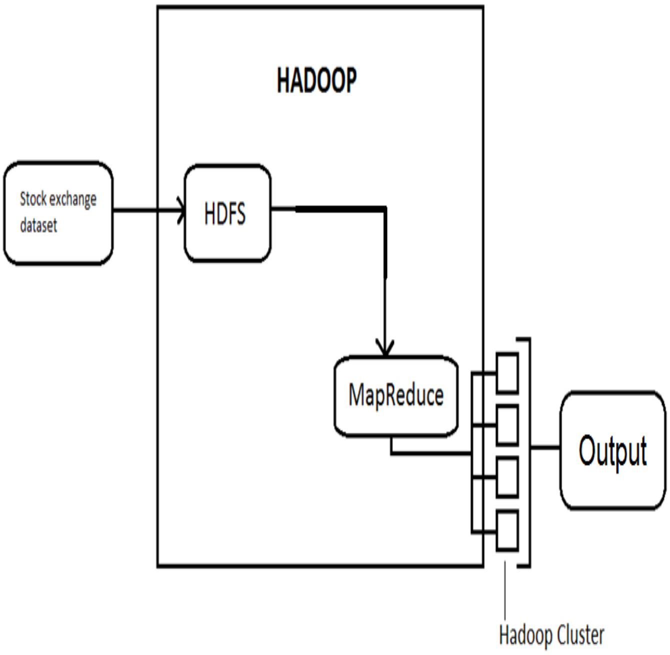


Fig 3.1: Proposed System

3.4Objectives

There are 3 major requirements to be satisfied in the order:
1.Acquiring the stock exchange dataset.
2.Designing an appropriate prediction algorithm.
3.Applying prediction algorithm on test input data and getting output.

3.5Specific Requirements

3.5.1 Hardware Requirements

This system will be designed to work on general purpose computers with processor Intel dual core and above. The data is acquired from finance.yahoo.com. It is a .csv file (comma separated value file). It is a structured dataset. It contains daily stock data of a company which includes attributes – date, low, high, open, close and volume. The nodes are interconnected using an algorithm which implements the MapReduce model developed by Google. Hadoop will take care of how the data and processing are managed through multiple nodes.

3.5.2 Software Requirements

- The system uses historical stock dataset acquired from finance.yahoo.com to predict a company's future quarter closing prices.
- Operating system used is Ubuntu which is based on Linux
- Terminal is used for the commands.

3.5.3 Functional Requirements

1) **3.5.3.1 Create Prediction model**

- Introduction: This function will take a stock dataset and provide a prediction model
- Inputs: Stock dataset of a company
- Processing: Workload will be shared among multiple nodes using an algorithm which implements MapReduce and using time series analysis algorithm, it will produce the prediction model.
- Outputs: Prediction Model

Figure 3.2 depicts the activity diagram for the proposed system.

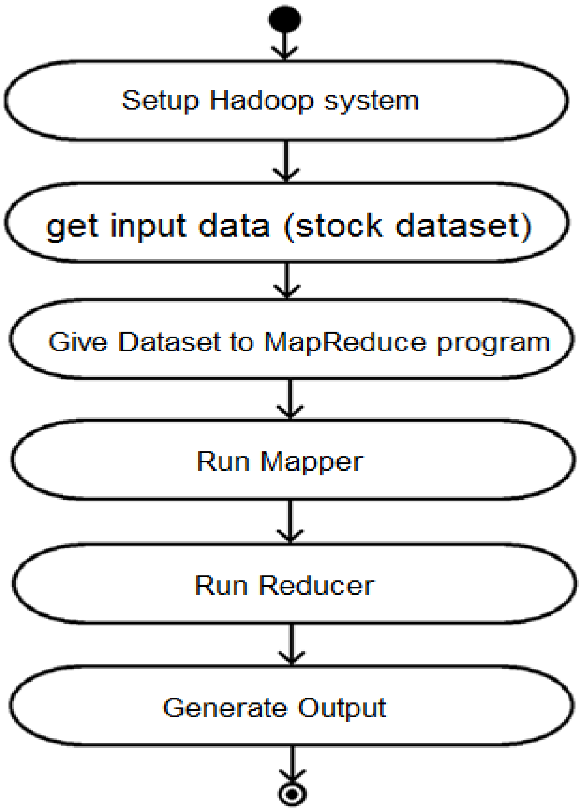

Setup Hadoop system
get input data (stock dataset)
Give Dataset to MapReduce program
Run Mapper
Run Reducer
Generate Output

Fig 3.2: Activity diagram

4. IMPLEMENTATION

4.1 Module Implementation

4.1.1 Time series analysis

Time series analysis technique consists of methods to analyze the time series data such as in the proposed system the historical stock exchange data, which is helpful in obtaining the statistics and various characteristics that are useful in decision making process. This particular technique make use of the observed values in order to predict the future values.

Natural temporal ordering being one of the feature of time series data, makes time series analysis distinct from cross-sectional studies. In the cross-sectional studies there is no natural ordering means individual's data could be entered in any order. Also time series is different from that of spatial data analysis where in later analysis the focus is given on geographical locations. It has been observed that the stochastic model for a time series will generally reflect the fact that observations made with time are close related to the accurate results. In addition to all the above features, in time series the values for a given period has been derived by using some methods from the past (historical) values rather than making use of future values.

Time series is considered to be a sequence of data points with following features:

1) Over a continuous time interval

2) Successive measurements across the particular interval

3) By making use of equal spacing between every two consecutive measurements

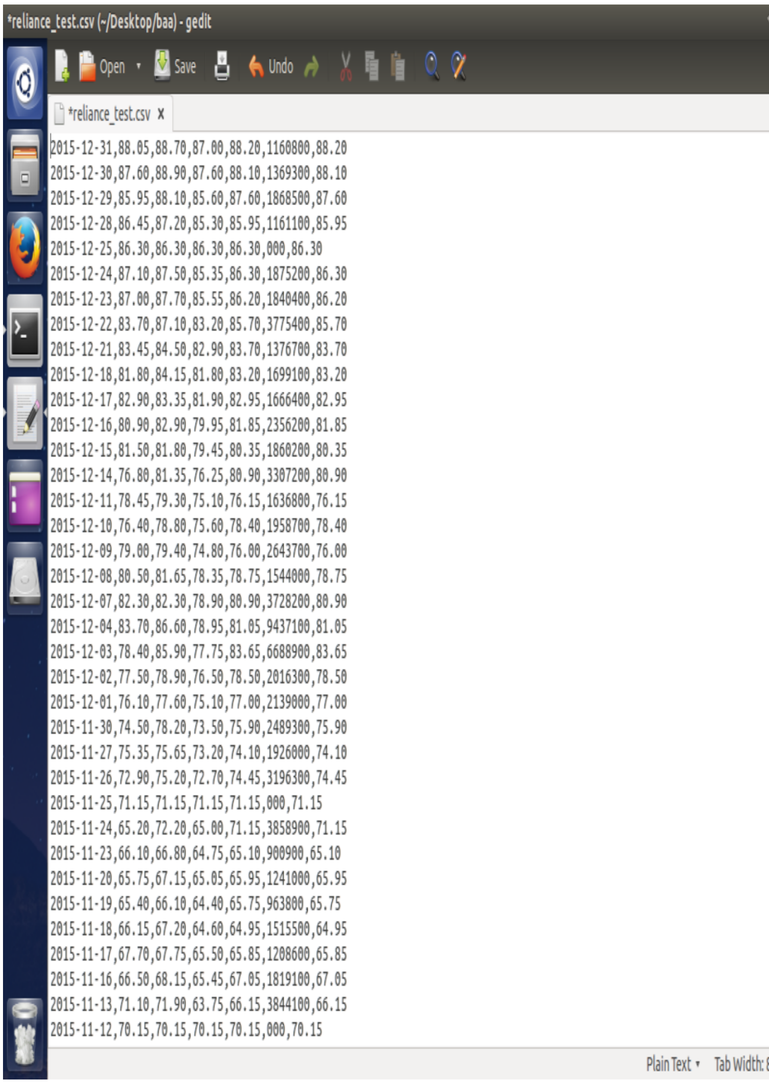
4.1.2 Time series algorithm

- y is a variable. It is collected data over time t.
- Time (t) is in quarters.
- y is the average closing price for that quarter.
- y=TSI
- Where T is Trend.
- I is the irregularity constant.
- S is the seasonal change in the y value.
- T is the trend calculated using variables b0 and b1.
- Sumt is the total summation of t values.
- Sumy is the total summation of y values
- Sumty is calculating t*y then total summation of t*y
- Sumt2 is total summation of t*t
- We Calculate all the variables (sumt, sumy, etc)
- Then we calculate b0 and b1.
- We calculate the future Trend using formula: T=b0+tb1
- Then we correct the prediction by applying the seasonal and irregular effects
- Predictiondone.

5. RESULTS AND SNAPHSHOTS

5.1 Screenshots

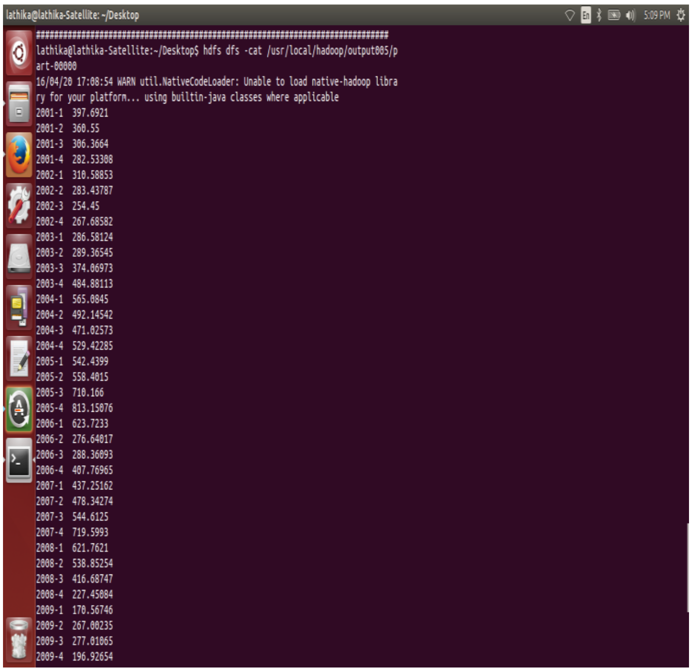
The figure 5.1 is the test dataset. It is a .CSV file (Comma Separated File)



```
2015-12-31,88.05,88.70,87.00,88.20,1160800,88.20
2015-12-30,87.60,88.90,87.60,88.10,1369300,88.10
2015-12-29,85.95,88.10,85.60,87.60,1868500,87.60
2015-12-28,86.45,87.20,85.30,85.95,1161100,85.95
2015-12-25,86.30,86.30,86.30,86.30,000,86.30
2015-12-24,87.10,87.50,85.35,86.30,1875200,86.30
2015-12-23,87.00,87.70,85.55,86.20,1840400,86.20
2015-12-22,83.70,87.10,83.20,85.70,3775400,85.70
2015-12-21,83.45,84.50,82.90,83.70,1376700,83.70
2015-12-18,81.80,84.15,81.80,83.20,1699100,83.20
2015-12-17,82.90,83.35,81.90,82.95,1666400,82.95
2015-12-16,80.90,82.90,79.95,81.85,2356200,81.85
2015-12-15,81.50,81.80,79.45,80.35,1860200,80.35
2015-12-14,76.80,81.35,76.25,80.90,3307200,80.90
2015-12-11,78.45,79.30,75.10,76.15,1636800,76.15
2015-12-10,76.40,78.80,75.60,78.40,1958700,78.40
2015-12-09,79.00,79.40,74.80,76.00,2643700,76.00
2015-12-08,80.50,81.65,78.35,78.75,1544000,78.75
2015-12-07,82.30,82.30,78.90,80.90,3728200,80.90
2015-12-04,83.70,86.60,78.95,81.05,9437100,81.05
2015-12-03,78.40,85.90,77.75,83.65,6688900,83.65
2015-12-02,77.50,78.90,76.50,78.50,2016300,78.50
2015-12-01,76.10,77.60,75.10,77.00,2139000,77.00
2015-11-30,74.50,78.20,73.50,75.90,2489300,75.90
2015-11-27,75.35,75.65,73.20,74.10,1926000,74.10
2015-11-26,72.90,75.20,72.70,74.45,3196300,74.45
2015-11-25,71.15,71.15,71.15,71.15,000,71.15
2015-11-24,65.20,72.20,65.00,71.15,3858900,71.15
2015-11-23,66.10,66.80,64.75,65.10,900900,65.10
2015-11-20,65.75,67.15,65.05,65.95,1241000,65.95
2015-11-19,65.40,66.10,64.40,65.75,963800,65.75
2015-11-18,66.15,67.20,64.60,64.95,1515500,64.95
2015-11-17,67.70,67.75,65.50,65.85,1208600,65.85
2015-11-16,66.50,68.15,65.45,67.05,1819100,67.05
2015-11-13,71.10,71.90,63.75,66.15,3844100,66.15
2015-11-12,70.15,70.15,70.15,70.15,000,70.15
```

Fig 5.1: Dataset

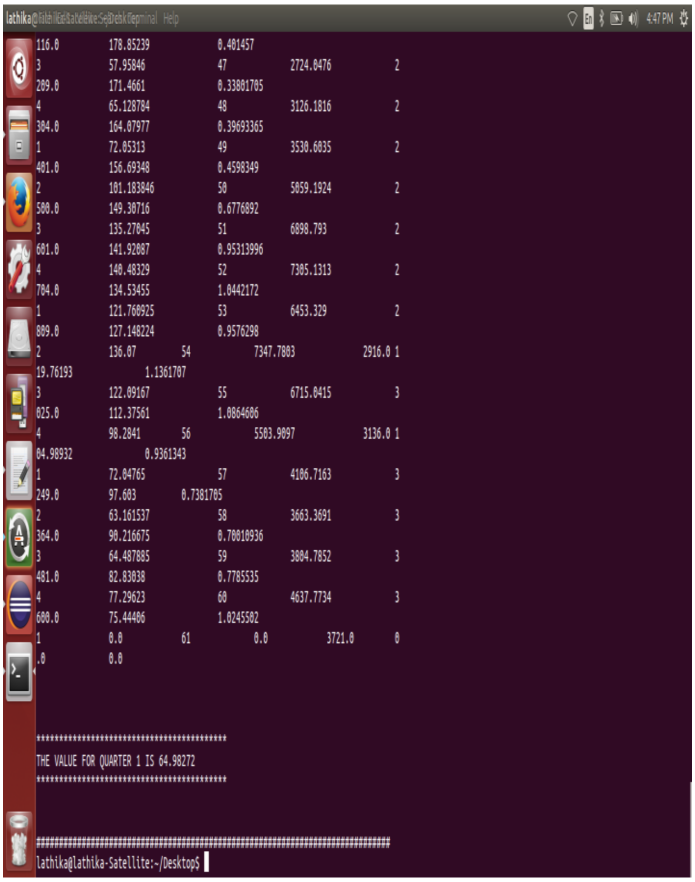
Figure 5.2 shows the output of the reducer



```
########################################################################
lathika@lathika-Satellite:~/Desktop$ hdfs dfs -cat /usr/local/hadoop/output005/p
art-00000
16/04/20 17:08:54 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
2001-1  397.6921
2001-2  360.55
2001-3  306.3664
2001-4  282.53308
2002-1  310.58853
2002-2  283.43787
2002-3  254.45
2002-4  267.68582
2003-1  286.58124
2003-2  289.36545
2003-3  374.06973
2003-4  484.88113
2004-1  565.0845
2004-2  492.14542
2004-3  471.02573
2004-4  529.42285
2005-1  542.4399
2005-2  558.4015
2005-3  710.166
2005-4  813.15076
2006-1  623.7233
2006-2  276.64017
2006-3  288.36093
2006-4  407.76965
2007-1  437.25162
2007-2  478.34274
2007-3  544.6125
2007-4  719.5993
2008-1  621.7621
2008-2  538.85254
2008-3  416.68747
2008-4  227.45084
2009-1  170.56746
2009-2  267.00235
2009-3  277.01065
2009-4  196.92654
```

Fig 5.2: Prediction Model

Figure 5.3 shows the output of the prediction algorithm on the test data set.



```
116.0           178.85239               0.401457
3               57.95846                47              2724.0476               2
209.0           171.4661                0.33801705
4               65.128784               48              3126.1816               2
304.0           164.07977               0.39693365
1               72.05313                49              3530.6035               2
401.0           156.69348               0.4598349
2               101.183846              50              5059.1924               2
500.0           149.30716               0.6776892
3               135.27045               51              6898.793                2
601.0           141.92087               0.95313996
4               140.48329               52              7305.1313               2
704.0           134.53455               1.0442172
1               121.760925              53              6453.329                2
809.0           127.148224              0.9576298
2               136.07          54              7347.7803               2916.0 1
19.76193                1.1361707
3               122.09167               55              6715.0415               3
025.0           112.37561               1.0864606
4               98.2841         56              5503.9097               3136.0 1
04.98932                0.9361343
1               72.04765                57              4106.7163               3
249.0           97.603          0.7381705
2               63.161537               58              3663.3691               3
364.0           90.216675               0.70010936
3               64.487885               59              3804.7852               3
481.0           82.83038                0.7785535
4               77.29623                60              4637.7734               3
600.0           75.44406                1.0245502
1               0.0             61              0.0             3721.0          0
.0              0.0



*************************************
THE VALUE FOR QUARTER 1 IS 64.98272
*************************************


########################################################################
lathika@lathika-Satellite:~/Desktop$
```

Fig 5.3: Prediction

5.2 Results: Closing Prices

The table below shows the value when comparison is made with the predicted average and actual average using the proposed system . The actual average is the closing price average for the 2016's first quarter.

Table 5.1 Results

	Actual average	Predicted average
Reliance	60	64.98
Google	750	829.91
General electric	29	32.43

In average, reduction is dataset size after MapReduce execution is approximately 98.5%.

5.3 Discussion

Using the proposed system, the value of next quarter can be predicted. This will help the shareholders and the company people to make better decisions to attain higher profits. Also enables to quickly learn patterns within large data sets and make intelligent predictions. It is possible to utilise multiple systems to work together for the output.

6. CONCLUSION AND FUTURE WORK

6.1Conclusion

Using the proposed system, we will be able to:
-Help the shareholders and the company people to make better decisions to attain higher profits.
-Quickly learn patterns within large data sets and make intelligent predictions.
-Utilise multiple systems to work together for the output.

6.2Future work

-Make use of the current news for better prediction of stock prices.
-Use a prediction algorithm which can produce better predictions for a broader choice of datasets.

REFERENCES

[1] Raj Kumar,Anil Balara, "Time Series Forecasting Of Nifty Stock Market Using Weka" , 2010, JRPS

[2] Gabriel Fiol-Roig, Margaret Miro-Julia, and Andreu Pere Isern-Deya, “Applying Data Mining Techniques to Stock Market Analysis”,2010, Springer
[3] Binoy.B.Nair, V.P Mohandas, N. R. Sakthivel , “A Decision Tree- Rough Set Hybrid System for Stock Market Trend Prediction”, 2010, IJCA
[4] Thawronwong, “The use of data mining and neural networks for forecasting stock market returns”, 2012, Journal of Computers
[5] Yuqing Dai, Yuning Zhang , “Machine Learning in Stock price trend forecasting” , 2012
[6] Qasem a. Al-radaideh, Adel abu assaf ,Eman, “ Predicting stock prices using data mining techniques” ,2013,ACIT
[7] Khalid Alkhatib Hassan Najadat Ismail Hmeidi Mohammed K. Ali Shatnawi, “Stock Price Prediction Using *K*-Nearest Neighbor (*k*NN) Algorithm”, 2013 , IJBHT
[8] Abhishek Gupta, Dr.Samidha D Sharma, “ Clustering-Classification Based Prediction of Stock Market Future Prediction”, 2014, IJSCIT
[9] Mustain Billah, Sajjad Waheed,Abu Hanifa, “Predicting Closing Stock Price using Artificial Neural Network and Adaptive Neuro Fuzzy Inference System(ANFIS)”, 2015 , IJCA
[10] https://www.youtube.com/watch?v=2S5kf73Cdws, Time series analysis video
[11] Apache Hadoop, http://Hadoop.apache.org.
[12] Tom White. Hadoop: The Definitive Guide. O’Reilly, Sebastopol, California, 2009.
[13] http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/