Stocks Analysis and Prediction of Indian oil Trading Using Big Data Analytics

M Sunil Kumar[1], K. Gokula Chandra[2], Keshav Kumar Gupta[3]

[1] Professor, Department of Computer Science and Engineering,

Sree Vidyanikethan Engineering College (Autonomous), Tirupati, AP, India.

[2,3] UG Scholars, Department of Computer Science and Engineering,

Sree Vidyanikethan Engineering College (Autonomous), Tirupati, AP, India.

Abstract:

Big data analytics is in high demand in a variety of sectors (such as human sciences, business, and others) since it allows for exact forecasting and analysis of large data sets. We are able to extract essential information by embracing big and multiple data sources that would otherwise be inaccessible. It is presented in this paper how a comprehensive strategy based on the Cloudera and Hadoop frameworks may be used to accomplish analysis for any sort of data. For example, Indian oil trading stocks are studied to anticipate profits using real-time data from the Yahoo Finance API on a daily basis. It's feasible to store and handle massive datasets in a distributed fashion using Apache Hadoop big-data framework and modules like PySpark for machine learning. Stocks are selected from the Indian stock market and disseminated into training data set and test data set in order to forecast stocks that are likely to make large profits.

Key word: Big data, stock Analysis, PySpark, Hadoop framework,

I. INTRODUCTION

Big data facilitates us to perceive various customer related patterns and trends. It is exercised in various field including targeted advertising, education, healthcare and in many other real-life scenarios. Furthermore, big data is of significant relevance in the fields of information sciences, technology, and the cloud computing sector, among other things [3-4].

Every second, the amount of data generated in the banking sector skyrockets. According to GDC's forecast [1], this data is expected to increase by 700% by 2023. When it comes to venture credit risk management, money laundering, and the misuse of debit and credit cards (among other things), the analysis of big data is invaluable [14]. Assists firms in working with their data more efficiently and identifying new opportunities using data. With the help of various tools, such as strategies and algorithms, companies are able to make informed decisions that lead to better business decisions and more profitability[5].

Examining real-time data on the price of IND oil stocks allows us to better understand how that stock price is affected by the IND Oil index. To make the IND Oil Stocks Trading community successful, it can also assist us forecast the profitable stocks [6].

II. PRELIMINARY KNOWLEDGE

There are a number of elements that can be used to better estimate demand, stock value, or anything else that persists in data, such as Hadoop framework. In this section, some of the Hadoop framework's essentials are overlooked [13].

Hadoop framework:

Apache Hadoop is an open-source software framework that operates on data processing applications in a distributed computing environment [7]. It is developed by the Apache Software Foundation. This framework is written in java to improve performance constraints such as throughput, latency, etc. [12],

The Hadoop ecosystem consists of

->Distributed File System (HDFS)

->MapReduce, at its core.

As a result, the Hadoop MapReduce framework [2] is frequently used to refer to it. Using distributed computing and parallel computing, this framework allows a program/model written in Java to be processed by numerous computers. The prime benefit for using MapReduce framework is that it is easy to scale enormous data over multiple computing modes [8].

This model comprises of primitives namely mappers and reducers (at times is nontrivial).

In a word, the aim is to distribute the workload among a number of clusters that are compatible with one another's file systems. At the heart of the Hadoop Ecosystem is the Hadoop Distributed File System (HDFS), which works in conjunction with the MapReduce methodology [9] (As shown in Fig 1)

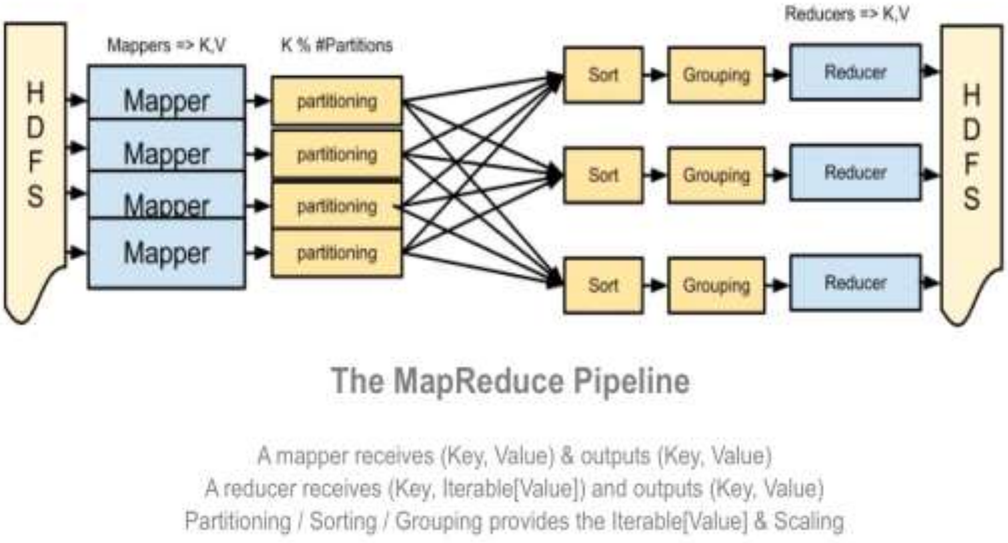


Fig 1: Architecture Pipeline:

Based on Google File System (GFS), Hadoop's Distributed File System offers a distributed file system capability that runs on commodity hardware. Deployed on low-cost hardware with strong fault-tolerance, it also reveals the differences between several distributed file systems.[10]

In this paper our prime issue is on reducing the overall hardware and software cost, hence

For capacity planning and management, Hadoop enables telecoms companies better evaluate switching performance and network frequency use. Telcos can pinpoint the optimal locations for new cell towers and more quickly fix network issues by looking at how mobile services are used and the available bandwidth in different regions[11][15].

III. A HADOOP BASED DATA PIPELINE APPROACH

There are normally five phases to implementing the approach: Data Acquisition and Characterization, Data Transfer, Storage, Pre-processing, and Machine Learning (as shown in Fig 2)

A. Data Characterization:

Our data collection comprises 17 oil stocks from the S&P BSE 500, all of which are publicly traded on Yahoo Finance. From April 2018 until the present, we will train the model on the values of these stocks. The date was chosen because the S&P BSE (IND Oil Fund) became available for trading oil stocks in India on that day. For float, we have 13 different numerical variables (data type). There were 4705 rows and 11 columns in the data we processed. The dataset, which contains no null values, is also accessible in CSV format for local analysis. After normalizing the data with 100, the results were obtained.

B. Data Transfer:

Flume was used to transfer the log data to HDFS. Components in the same hierarchical namespace must meet the following requirements:

->Source

->Sink

->Channels

The source was accessed using the tail command on the shell, and the sink was DFS. Configuring the local database to HDFS was required for data injection. The configuration that is used is as follows show in Fig 2.

```
# list the sources, sinks and channels for the agent
<Agent>.sources = <Source>
<Agent>.sinks = <Sink>
<Agent>.channels = <Channel1><Channel2>
# set channel for source
<Agent>.sources.<Source>.channels = <Channel1><Channel2>
# set channel for sink
<Agent>.sinks.<Sink>.channel = <Channel1>
```

Fig 2: Namespaces and their Syntaxes*:*

C. Storage:

Intuitively, tightly-coupled HDFS data is stored via subsequent steps namely moving file to cluster by employing Cloudera. HDFS on Cloudera was utilized to store the data. In order to generate dependable and efficient data processing, data replication is done utilizing a single node. Data is always kept in many sequential files, with each entry only being represented once (hence no duplication). Data was also available from yahoo finance via the directory HDFS/Cloudera/flume/events/.

D. Pre-processing:

Python API Data is pre-processed using PySpark. The sequence file is read into the spark and then transformed to an RDD (Resilient Distributed Datasets) using spark context. Converting it to a data frame provides the requested input with a suitable schema and data type. The Dense Vector function is used to combine the input features into a single feature named features. The data frame contains two columns: label and features, both of which are ready to be input into the machine learning model in fig 3.

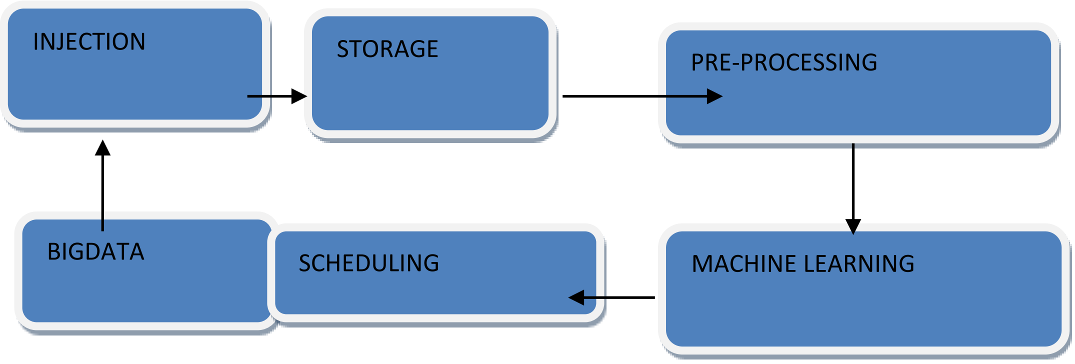
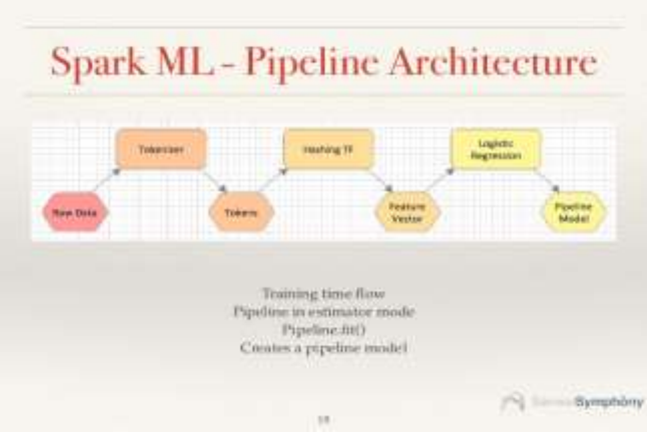


Fig 3: Preprocessing Pipeline*:*

E. Machine Learning:

The data is separated into training and testing data sets and learning is performed in Spark using the Mlib function. The training dataset is used to create the linear regression function. The returns of the S&P BSE 500 index are predicted for the test set of data. Computed mean squared error, for example (as shown in Fig 4).



Spark ML - Pipeline Architecture

Training time flow
Pipeline in estimator mode
Pipeline.fit()
Creates a pipeline model

Fig 4: Spark Architecture:

IV. RESULTS AND DISCUSSION

In Figure 5, we can see how the model learned from the training data and then used the regression model coefficients to correlate stock prices with forecasts. Figure 5: Learning from Training Data. The R squared value and the Mean Average Error are utilized to gather additional information for the investigations.

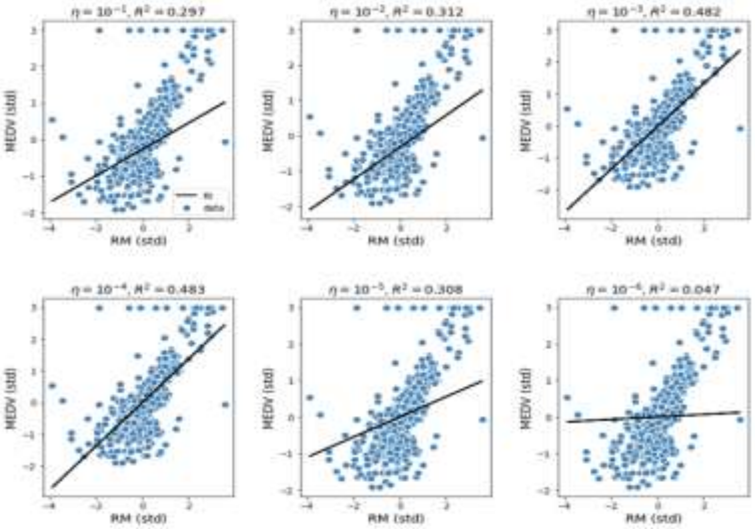


Fig 5: Charts describing the detailed plot b/w R-value and Error values*:*

(1) Coefficients Regression Model

Based on Yahoo Finance, the problem of the prediction of stock prices like "[0.0817, 0.0, 086, 0.0, 0.0.0.0682, 0.0675, 0.11, 0.0438, 0.0046]" and the reversal value of the regression model like – 0.097938396250894261 might assist us to take accurate data-driven decisions.

(2) Evaluation Metrics:

R-squared and MAE values are two important metrics (as shown in Fig 8). Because there are no positive coefficient values in the regression model, we may conclude that inventory prices are not related to the other IND oil stocks that were used as predictors in the study. The model is generated using the parameter regularization (~=0.3). And the R-squared value is determined by utilizing the regression evaluator package evaluation function of the machine learning module's machine learning module. (Namely Spark, etc.).

This explains 6 percent of the changes in S&P BSE stock prices. The MAE will decide whether or not the model is appropriate to estimate inventory return margins from data of high dimensionality.

V. CONCLUSION AND FUTURE WORKS

In this study, big data analysis is used to do efficient stock market analysis and forecasting, which is a first in the field. Generally speaking, the stock market is an unpredictably volatile environment in which a failure to accurately estimate the value of a company's stock can result in severe financial losses for investors. The results of our investigation allowed us to provide an approach that can help us find companies with positive daily return margins, which can subsequently be selected as potential stocks for increasing trading. So, Hadoop may learn from previous data and decide on streaming updates which IND inventories are beneficial for trading by using this method as an intermediary. We're also looking to the future to see how we can improve our studies. We plan to use the programming module to automate the analytic activities and then receive regular IND inventory trading advise to continue our investigation. In addition, rather of utilizing linear regression, we'd like to try some model-based learning using Neural Networks to reliably estimate the values of IND stock.

Reference:

1. ZhihaoPeng, "Stocks Analysis and Prediction Using Big Data Analytics", 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE Xplore: 21 March 2019, DOI: 10.1109/ICITBS.2019.00081, ISBN:978-1-7281-1307-4.
2. Xingguo Cheng, Nanfeng Xiao, Faliang Huang, “Research on HDFS-based web server cluster", 2011 International Conference on E-Business and E-Government (ICEE), IEEE Xplore, 16 June 2011,ISBN:978-1-4244-8691-5.
3. Sushama, C., Kumar, M. S., & Neelima, P. (2021). Privacy and security issues in the future: A social media. Materials Today: Proceedings.
4. Kumar, M. Sunil, and K. Jyothi Prakash. "Internet of things: IETF protocols, algorithms and applications." Int. J. Innov. Technol. Explor. Eng 8, no. 11 (2019): 2853-2857.
5. Kumar MS, Harshitha D. Process innovation methods on business process Reengineering. Int. J. Innov. Technol. Explor. Eng. 2019.

6. Peneti, Subhashini, M. Sunil Kumar, Suresh Kallam, Rizwan Patan, Vidhyacharan Bhaskar, and Manikandan Ramachandran. "BDN-GWMNN: Internet of Things (IoT) Enabled Secure Smart City Applications." Wireless Personal Communications (2021): 1-17.
7. Kumar, M. Sunil, and A. Rama Mohan Reddy. "An Efficient Approach for Evolution of Functional Requirements to Improve the Quality of Software Architecture." In Artificial Intelligence and Evolutionary Computations in Engineering Systems, pp. 775-792. Springer, New Delhi, 2016.
8. Harika, A., M. Sunil Kumar, V. Anantha Natarajan, and Suresh Kallam. "Business Process Reengineering: Issues and Challenges." In Proceedings of Second International Conference on Smart Energy and Communication, pp. 363-382. Springer, Singapore, 2021.
9. MS Kumar, ARM Reddy, AV Sriharsha, " Pragmatic Applications of Architectural Knowledge", Int. J. of Recent Trends in Engineering and Technology, 2009.
10. MS Kumar, A Ramamohan reddy, "A survey on user-interface architectures and ADLs", Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012), 2012 p. 229 – 232.
11. B Rupesh, MS Kumar, "Predicting the Hard Keyword Queries over Relational Databases", International Journal of Applied Engineering Research, 2015.
12. V Guna, MS Kumar,"A Survey on Software Code Clone Detection to Improve the Maintenance Effort and Maintenance Cost of the Software", Volume-6, Special Issue-3, April 2018.
13. D Ganesh, MS Kumar, VVR Prasad, "IMPROVING NETWORK PERFORMANCE IN WIRELESS SENSOR NETWORKS", Integrated Intelligent Research (IIR) , International Journal of Web Technology Volume: 05 Issue: 01, June 2016, Pages: 58-61.
14. MS Kumar, A Harika, "Extraction and classification of Non-Functional Requirements from Text Files: A Supervised Learning Approach", Psychology and Education Journal Volume 57 No. 9 (2020).
15. V Anantha Natarajan, D. Ganesh, Macha Babitha and M. Sunil Kumar, "Machine Learning Based Identification of Covid-19 From Lung Segmented CT Images Using Radiomics Features, Biosc.Biotech.Res.Comm. Special Issue Vol 14 No 07 (2021)