Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Stock Exchange Trend Analysis using Hadoop Server: A Survey

Krishna Patel
Department of Computer Science & Engineering, Medicaps Institute of Technology and Management, Indore, Madhya Pradesh, India

Abstract— The growth of technology increases the role of network in applications and services. A stock market is a place where equity or share can buy or sale for economic transaction purpose through online system. Online transaction of economic interest not only helps to increase security and flexibility of operations but also gives a platform to analyze the market trend by mining the history of the previous market trading. Stock market is an organization where millions of treads happen for the duration of an event of time. Hereby, an exclusive big data analysis is required to observe and conclude the trend of market and getting an interest of user for a particular share or organization. Hadoop server gives a wide platform for parallel execution and performs mining observations on huge data collection. Clustering algorithms are used to form the group of large datasets while Hadoop server helps to partition the large datasets for efficient execution. This paper attempts to explore the detail about techniques and framework to explore the knowledge from large data set.

Keywords— Stock Exchange Market, Clustering, Hadoop, Map-Reduce

I. INTRODUCTION

A rapid growth in society changes the responsibility and role of technology. Network plays a very important role for atomization of work. Thus it becomes the spine of system. Rapid development of applications and services raise online usage and it becomes more complex and emergences. Therefore, Stock Market Analysis and Control is required to manage stock market trends and classification. Most methods for stock market analysis are operated on a single server environment. But if the amount of market data is increased, the existing infrastructure required increased infrastructure, memory speed and storage drives. A large quantity of data is known as Big Data, required specialized methodology for efficient classification.

Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. In the recent year, a rapid hike in the large quantity of data is observed and it becomes the most useful technology tool for business.

A heavy processing and large amount of communications produces large amount of data which may be structured or unstructured from the different sources. The need of big data comes from the Big Companies like yahoo, Google, face book etc for the reason of study of big quantity of data which is in shapeless outline. The study report conclude that Google, Face book, tweeter, WHO etc. have large amount of data and required special technique to process them.

Here, an example of Big-Data might be in petabytes or exabytes. The data may be unstructured or structured some time may be incomplete or inaccessible. It can be categorized by 3Vs may be Volume, Velocity and Variety.
A brief classification of 3Vs is cited below and drawn is figure 1.1.

- Volume
- Velocity
- Variety

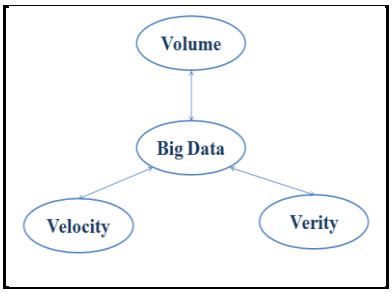

Fig 1 V Relation with Big Data

Big Data term is rapidly getting a huge hike in technical aspects. A wide scale of utility it becomes centric interest of various researchers. Large size of data creates complex and difficult environment for processing, storage and transfer of information. Traditional algorithms and mining approach is not suitable for large data set and required a separate standards for parallel processing and distributed storage along with computation which help to reduce overhead and increase execution performance. Various tools are analysed and discussed in this paper which is listed below.

A. Hadoop

This is an Open source tool provides a reliable, scalable and distributed environment for creating data partition from inexpensive servers. Here, Map-Reduce framework can be used to process large scale of data with minimum overhead. This paper investigates certain situations where Hadoop can be useful.

1. Complex information processing is needed.
2. Unstructured data needs to be turned into structured data.
3. Queries can't be reasonably expressed using SQL.
4. Heavily recursive algorithms.
5. Complex but parallelizable algorithms needed, such as geo-spatial analysis or genome sequencing.
6. Machine learning.
7. Data sets are too large to fit into database RAM, discs, or require too many cores (10's of TB up to PB).
8. Data value does not justify expense of constant real-time availability, such as archives or special attention information, which can be stimulated to Hadoop and remain available at lower charge.
9. Results are not needed in real time.
10. Fault tolerance is critical.
11. Significant custom coding would be required to handle job scheduling.

B. HDFS

The Hadoop Distributed File System (HDFS) is the file system component of the Hadoop framework. HDFS is designed to play down the storage overhead and mining the large amount of data on distributed fashion hardware.

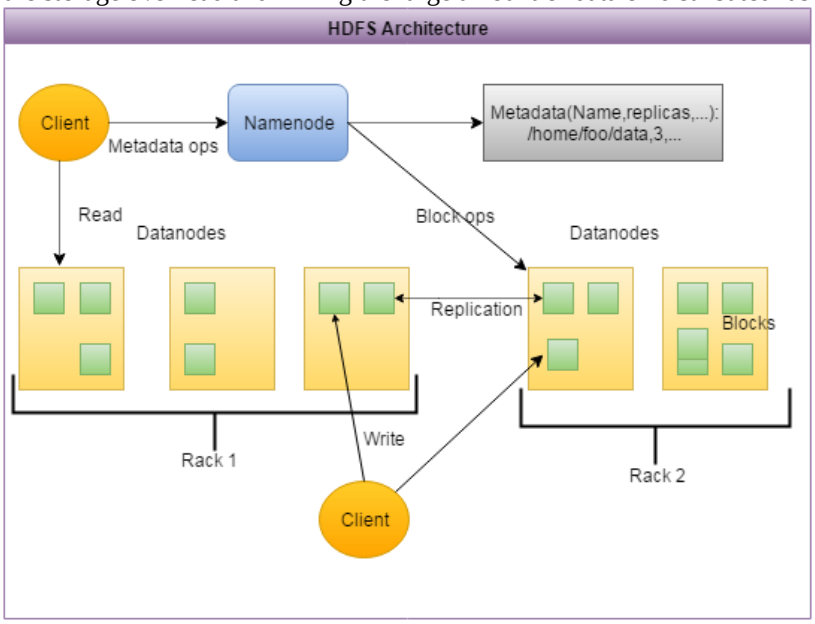

Fig 2 Block Representation of HDFS

C. PIG

PIG is the important component of Hadoop server like Map-Reduce and HDFS. Pig is made up of two components: Block representation of PIG is shown in figure 3.

1. The first is the language itself, which is called Pig Latin (people naming various Hadoop projects for relation with naming conventions.)
2. The second is a runtime environment where Pig Latin programs are executed.

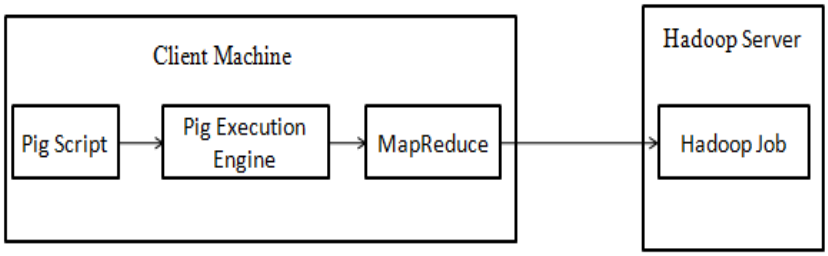

Fig 3 Block Representation of PIG

D. Map-Reduce

Map-Reduce is a framework for developing tools and source code for large data processing. It partitions the large data set into multiple parts to make processing easy and convenient. It simplifies the processing by auto making cluster according to name node based on machine.

Map-Reduce algorithm examines the dissimilar clusters and counsels the client for the common set of services used by the other users for the similar type of task. This will reduce the complexity and ambiguity of user to analysis the services provided by the cloud

The Map-Reduce Framework have two main function named with Map and Reduce. Here Map function is used to map large data into clusters and Reduce function is responsible to join the result into single unit. Proposed solution required that relative frequency of frequent service can be identified from different data set items. A block representation of Map-Reduce Framework is shown in figure 4.

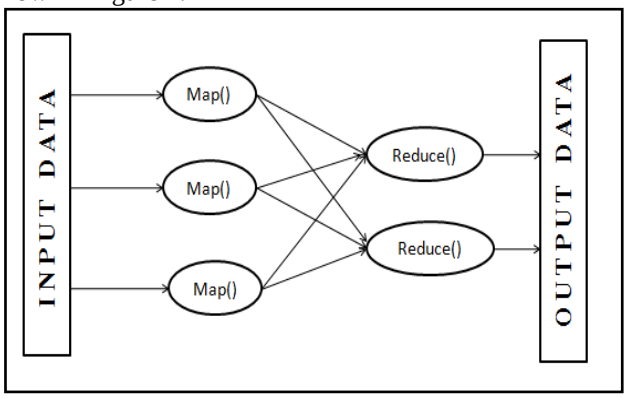

Fig 4 Block Representation of Map-Reduce

Analysis of Big Data through Map-Reduce function obtains the six phases which are;

1. Input reader
2. Map function
3. Partition function
4. Comparison function
5. Reduce function
6. Output writer

II. RELATED WORKS

G. Xu [1] In this paper author discussed about Cloud Stack, Map-Reduce and their implementation technique known as Hadoop along with virtualization technology Xen and KVM. They also express the advantages and disadvantage of Map-Reduce. They address that Map-Reduce can help to better utilize the system resources, cost-effectively process for vast amounts of data, reliability of services, processing and equation of power consumption. The only disadvantage of this method is poor load handling strategy.

R. Sun[2] Map-Reduce function obtains are an important model for Big Data processing and large data programming. Proposed solution approaches to overcome drawbacks of traditional Hadoop cluster deployment framework. Proposed solution improves data locality to improve Map-Reduce performance. In this paper they deploy Hadoop server with task scheduling algorithm. It uses three main modules which are Job Profiler, Task Profiler Monitor, and Migration Controller. Proposed solution evaluated on efficiency based on virtualized cluster with deployment model using computing nodes and storage nodes. The performance observation concludes that 86% of performance was benchmark and hike is observed on proposed solution.

D. Pandove[3] This research paper gives brief detail about various clustering algorithms and also explores a comparative study among them. This paper gives a basic overview of clustering algorithm and suggest some suitable approached to use them.

A. Method

Based on fundamental differences, clustering strategies can be divided into two parts:

1. Hierarchical or Agglomerative Algorithms
2. Point Assignment Algorithms

Hendahewa, C. [4] in this paper author observe the stock market data to perform financial analysis. Online stock market is the passage to perform various financial operations through remote computer terminal and give flexibility of location as well as feasibility of maneuver. Study of stock market observes that a huge probability of instability is possible and it may make drastic changes on financial situation. Thus analysis of stock market trend is mandatory to make positive operations. Here, authors proposed a sparse, smooth regularized regression model to develop a stable market situation using separate accounting of dependencies among companies. They consider real stock market data and develop a model for dynamic time varying system. Proposed solution also attempt to investigate fluctuation during instability and attempt to observe financial crisis on different time period.

Zhang, M. et. al.[5] In this paper author consider computational verb theory (CVT) to observe stock market data and develop relative cluster for same. They consider dataset of shanghai stock exchange of March 2010. Here, they perform pre-process of stock data and consist curve smoothing and low pass filtering approach for normalization. A K-means algorithm has been used to develop cluster for same.

Dubey, A, [6] In this paper author proposed a stock market prediction algorithm for large amount of data. Stock exchange is the most sensitive marketplace whereas millions of operations happen for the period of an occasion of time. Thus, terabyte data is generated which is tough to observe and process.

A separate and advance technique is used for processing and data analysis. In this paper author uses Map-Reduce Framework for large data partitioning using cluster and Hadoop server. A HDFS system has been consider for high bandwidth data partitioning. They observe that stock market is high risk place and required a keen prediction to avoid collapse condition. Here, they uses artificial neural network (ANN) to observe market trend. The complete work is observed on basis of computation time and memory consumption on one machine and various machines.

Shim[7] et. al. state that this research work proposed an application traffic classification in Hadoop Distributed Computing Environment. Rapid development of technology increased the network traffic which demand strong traffic classification for enhanced performance.

However, methods for network traffic analysis are not proposed to grab up the trend of increasing practice on the network. Developed methods are evaluated on single server and not feasible for distributed environment and Big Data. In this paper a payload signature classification method has been used in Hadoop Distributed Server.

Study of this research work concludes that traffic phenomena of current network have been changes and conventional traffic analysis method are not adequate. Here, dynamic changes and large traffic data sets are major challenges for analysis mechanism. Conventional mechanism such as signature based classification, traffic correlation based classification or machine learning based classifications are outdated and not suitable for large data set analysis. This paper proposed a new way of traffic categorization based on signature method in Hadoop distributed system.

In the proposed section, authors consider packet units of traffic from campus network. Collected packets are converted into Flow format through the flow generator. The flow is defined by 5–tuple analysis. It implement the complete phenomena in Map-Reduce Framework of Hadoop Server and observe the performance on basis of processing speed by traffic acquisition time.

Proposed method performs well in term of computation speed in comparison with computation speed of single node. On the other hand, it has certain drawbacks which are;

1. Adoption of Classification technique rather than clustering.
2. Low analysis rate.

Wang, R. [8] Stock exchange market place is the most sensitive area in financial field. Proposed solution aims to develop clustering algorithms to observe the trend of stock exchange. This may give the classifications of data on basis of corporate and market gain. Clustering algorithms are implemented to classify the stock and observe the classified results. The stock selection and recommendation is based on the success rate of investment decisions.

A comparative study on existing work is also done and concludes and shown in table 1.

III. PROBLEM DOMAIN

Stock Exchange classification is important parts in the fields of Stock exchange marketplace analysis. Among them, method for stock exchange classification based on payload signature promises high classification accuracy rate. But there is disadvantages in payload signature based analysis that it requires high processing load comparing to other analysis method. Therefore, as the exchange data volume is increased the application stock data exchange classification using payload signature suffers from the burden in memory, processing speed, storage space. The complete analysis observes that it need to implement exclusive grouping mechanism to handle large data set and generate more accurate traffic analysis output.

Classification is a technique to learn to assign to predefined classes where clustering is task of grouping related data points together without labelling them.

Proposed application traffic classification technique follows the payload signature based classification which slows down the grouping mechanism and reduce the performance. The complete phenomena generate a need to develop exclusive mechanism for large data set of stock exchange data for Hadoop Distributed Server.

IV. SOLUTION DOMAIN

The data mining is a technique of software application that is used to analyse the huge amount of data. This approach becomes more crucial when data source found too large. A clustering based traffic analysis is proposed instead of signature based classification to observe the packet data set.

Clustering is a technique used for exploratory data analysis. It is an unsupervised technique used to classify large datasets in to correlative groups. No predefined class label is live for data instances. Clustering approach integrates the content into single unit and help for the classification of data. Such groups can help to retrieve knowledge from large amount of data for useful processing. This makes a convenient and simplifies processing with intellectual output.

Table 1: Comparison of various existing work and problem examination

Title	Problem Investigation	Proposed Solution
Analysis of Causality in Stock Market Data	Stochastic volatility in financial time series using traditional econometric methods	Sparse, smooth regularized regression model to detect causality in stock exchange

Application of Computational Verb Theory to Analysis of Stock Market Data	Classification of intra-day time series of stock prices.	Computational verb theory(CVT) using K-means clustering
Stock Market Prediction using Hadoop Map-Reduce Ecosystem	Processing overhead for large data input on single computation node	Proposed solution implies Map-Reduce framework to partition the large data into multiple parts and process map function to map and classify large data set into computation node and reduce function merge the output into single unit.
Stock Selection Based on Data Clustering Method	Stock selection and trend observation is major problem due to volatile nature of stock market.	Clustering technique is implemented to analyze and observe financial data and volatile stock exchange data. Classification of stock provides abstract view of data.

Data clustering is related with the separation of a data set into several groups such that the resemblance within a collection is better than that among groups. The complete system implies that a clustering approach can help to develop knowledge based on groups of large data. These groups will be compressed data set for knowledge retrieval queries and can help to recued overhead and simplify to meet the objective with fast growth.

A DBSCAN clustering algorithm is suggested to implement stock exchange traffic classification in Hadoop server. Expected flow of solution is cited below:

1. Collection of stock exchange data (Sample Packet).
2. Conversion of stock exchange Information into flow format.
3. Extract and Combine exchange information into structured format.
4. Analyze the stock exchange according to market priority.
5. Process of Map Reduce Function
6. Result observations.

For the purpose of processing the large amount of data, the big data requires outstanding technologies. Several techniques have been developed for manipulating observing and analysing the data.

Among all the available solutions HDFS (Hadoop Distributed File Server) is one of most appropriate solution. It is designed and optimized to store data over a large amount at low-cost hardware in a distributed fashion. So, HDFS on SE Linux and Stock Exchange Traffic Sample will be used to simulate and implement the proposed solution.

V. CONCLUSIONS

The study of complete work concludes that stock market have large amount of transactions and very frequent market trends. Analysis of stock market data can help to drawn market trends and help financial advisors to recommend about share broking and company investment prediction. Hadoop server and Map-Reduce Framework can help to process large amount of data with minimum overhead and better performance to perform stock market data analysis.

Clustering techniques can help to perform mining and data observations. Thus the complete work proposed to develop a solution for stock exchange transaction observation for large data and retrieve knowledge for market trend observations.

REFERENCES

[1] Guanghui Xu, Feng Xu*, Hongxu Ma, "*Deploying and Researching Hadoop in Virtual Machines*" published in International Conference on Automation and Logistics Zhengzhou,2012,pp.395-399.

[2] Ruiqi Sun, Jie Yang, Zhan Gao, Zhiqiang He, "A *Virtual Machine Based Task Scheduling approach to improve the data locality for virtualized Hadoop* " Published in IEEE ICIS-2014.

[3] Divya Pandove, Dr. Shivani Goel, "*A Comprehensive Study On Clustering Approaches for Bigdata Mining*"" published in 2nd International Conference on Electronics and communication system,2015,pp.1333-1338.

[4] Chathra Hendahewa, Vladimir Pavlovic, "*Analysis of Causality in Stock Market Data*" published in 11th International Conference on Machine Learning and Applications, 2012, pp. 288-293.

[5] Mengfan Zhang and Tao Yang, "Application of Computational verb theory to analysis of stock market data"" published in IEEE conference ,2010, pp. 261-264.

[6] Arun Kumar Dubey, Vanita Jain, A.P Mittal, "Stock Market Prediction using Hadoop Map-Reduce Ecosystem." published in IEEE conference ,2015, pp. 616-621.

[7] Ruizhong Wang, "Stock Selection Based On Datta Clustering Method" published in Seventh International Conference on Computational Intelligence and Security,2011,pp. 1542-1545.

[8] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, "Survey Paper On Bigdata" published in Vibhavari Chavan et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014,pp.7932-7939.

[9] Quang Tran, Hiroyuki Sato, "A Solution For Privacy Protection In MapReduce" published in 36th International Conference on Computer Software and Applications,2012,pp.515-520.

[10] Mr.S.Ramamoorthy, Dr.S.Rajalakshmi. "Optimized Data Analysis in Cloud Using Big Data Analytics Techniques" published in 4th ICCNT-2013.

[11] Krishna Mohan Pd Shrivastva, M A Rizvi, Shailendra Singh. "Big Data Privacy Based On Differential Privacy a Hope for Big Data" published in Sixth International Conference on Computational Intelligence and Communication Networks,2014,pp.776-781.

[12] Kyu-Seok Shim, Su-Kang Lee and Myung-Sup Kim, "Application Traffic Classification in Hadoop Distributed Computing Environment" published in Asia-Pacific Network Operation and Management Symposium, 2014.

[13] Carson Kai-Sang Leung, Richard Kyle MacKinnon, Fan Jiang, "Reducing the Searching Space for Big Data Mining for Interesting Patterns from Uncertain Data" published in IEEE International Congress on Big Data, 2014,pp.315-322.

ABOUT AUTHOR

Krishna Patel is currently pursuing M.E. (computer science and engineering) from Medicaps Institute of Technology And Management, Indore. He is a research scholar in the institute. He has completed B.E. (computer science and engineering) from SKSITS, Indore.

