2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.9315977

Propositional Aspect between Apache Spark and Hadoop Map-Reduce for Stock Market Data

Yogesh Kumar Gupta[1]
[1]Assistant Professor, Department of Computer Science
Banasthali Vidyapith
Niwai, India
gyogesh@banasthali.in

Nidhi Sharma [2]
[2]M. Tech (CS) Research Scholar
Banasthali Vidyapith
Niwai, India
nidhish21021999@gmail.com

***Abstract*— Big data analytics is becoming tremendously popular in every field today. Everyday lots of data are being generated and analyzed using big data analytics tools and technique. Here the technology used is apache spark and language used is Scala. So, in this paper study is being done on the behalf of research done in stock market data using apache spark technique. Here the nifty-50 data is taken to analyze the impact due to covid-19. As it is being seen that Covid-19 has affected almost everything around the globe, so the purpose is to analyze its effect on stock market. Thereafter comparison is done between the techniques used to analyze that massive volume of stock exchange data. Here the comparative analysis between Hadoop maps-reduce and apache spark on the behalf of some important parameter is being done. That concludes which technique is better for the analysis of the stock exchange data.**

Keywords*— *Apache spark, Big data analytics, Covid-19, Hadoop map-reduce, Stock market.

I. INTRODUCTION

Big data analytics is a process of analyzing the large volume of data. Now the data is generated in each and every field that's why tera bytes and peta bytes of data is being produced per day. So, to analyze, process and store the large amount of data which is structured, unstructured and semi structured various tools and techniques are needed. Many modern organizations are managing a huge amount of data that's why they need different techniques of data analytics. The data is being produced today at high speed which has also increased the need of data analytics.

Several numbers of tools and techniques are used in data processing. By using the big data analytics there are many benefits like customer services get improved due to which customer satisfaction is also there. Data storage get easier and the data stored is secure and can be analyzed any time. While using big data analytics, data collection and interpretation become easier this will change the working of small businessmen. New technology, innovations and idea are getting involved which improve the organizations overall development. There are different challenges also which are faced by big data analytics like there is shortage of the professionals who knows big data analytics concepts very well. There is data quality also due to which missing data and duplicate data can happen, which all result in data quality challenges which get affected. Privacy and secrecy of the data is also an important factor to consider.

There are some features of big data analytics like data predictive analytics, reporting, security, technical support, and row data processing. It is being proved and believed that more accurate data means more accurate results, and obviously more accurate analysis will be performed to ensure the timeliest and confident decision making which will also lead to better judgment and good decision making therefore cost and risk factor will be reduced.

So, the summary is that data analysis is the process to find knowledge from the large amount of data. Big data analytics includes 7 V's they are volume, velocity, variety, veracity, variability and visualization.

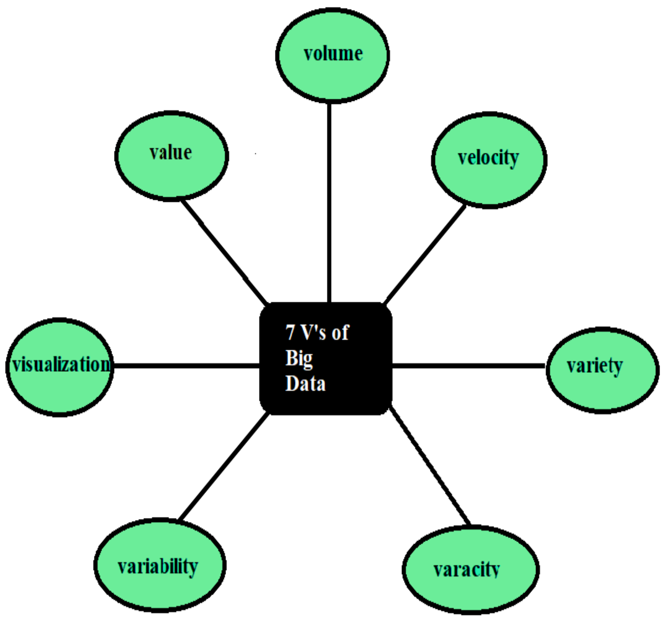

Fig 1. 7 V's of big data analytics.

Here volume is the Data quantity. Velocity is the speed. Variety means it can be structured, unstructured, & semi structured all three types of data. Veracity means that the data is perfect or not. Variability and variety are not same they have a little difference. Visualization is done using charts and graph to show the results of data analysis in the form of graphs. In value this is seen that the organization is having value or not.

Covid-19 has affected almost all areas of whole world if it is business, education, healthcare, economy, development, stock market and many more. Due to covid-19 so many people are dead and many are still positive and are being treated. As there is no vaccine yet the treatment is also not possible, covid-19 has affected the lives too much almost everything has been disturbed.

There is huge effect due to covid-19 on stock market also so I decided to analyze the impact of covid-19 on stock market Nifty 50 data. In nifty-50 there are 50 most popular large-cap company's stocks are included. Nifty-50 is popular one stock indices used in India other one is BSE SENSEX.

The stock market is known to be a non-stable and complicated system. In financial investment field also the stock market prediction is considered as a great factor. For prediction of stock market trends, only historical data is not considered as a main factor. There are many other factors also which can affect the stock market trends like politics, natural events which also affects social media like tweeter and Facebook which results to change in data that is provided for data analysis.

In this research work the impact of covid-19 on nifty 50 data is being analyzed. I have got the data from kaggle.com. The data includes the date, symbol, type of security, previous day close price of the day, open price of day, highest price of the day, lowest price of the day, last traded price of the day, closing price, volume weighted average price. These are the attributes which are there in the data with the help of these data analysis will take place; the data includes 50 different firms for which the data is provided. Apache spark technique will be used for data analysis. Hadoop is very highlighted data analysis framework as there is simple programming model which is map reduce. Hadoop is having limitations that it is not having real time data processing, it is not easy to use, code is lengthy in Hadoop and processing speed of Hadoop is slow. Then to enhance the processing speed and faster run time apache spark is introduced for processing larger data sets in less time.

In big data analytics apache spark and Hadoop map reduce both have played a very vital role. In spark applications can be written very quickly. Spark includes over 80 high level build-in operators. In spark the applications can be quickly written in java, Scala, python etc. Hadoop map reduce is one of the popular and powerful computational model. Hadoop have flexible scalability and ease of use. Hadoop also includes failover properties.

Apache spark is a cluster-based computing technology. The main purpose to design apache spark is for fast computation. In-memory cluster computing is the main features of apache spark that is used to increase the processing speed of an application in use. The designing of the spark technique is done in the way that it is having the capacity to take heavy loads such as iterative algorithms, batch applications, streaming etc. Apache spark having some features like speed; it supports multiple languages and advanced features to analyze the data. On the other side apache spark is also having some limitations like it doesn't include file management, less no. of algorithms are there and the throughput of apache spark is less due to its high latency.

Recently there are lots of multinational companies that already going to work in this field for the analysis of the massive volume of data. There are many sectors in stock exchange market, in which the prices of stock are fluctuate rapidly at very high speed in a sort interval. So the analytical tools apache spark is the most important tool that plays a very big role for the analysis of the stock data and finds the movement in stock prices. It facilitates the analyst and researchers to understand the scenario of the stock market in a very limited period of time.

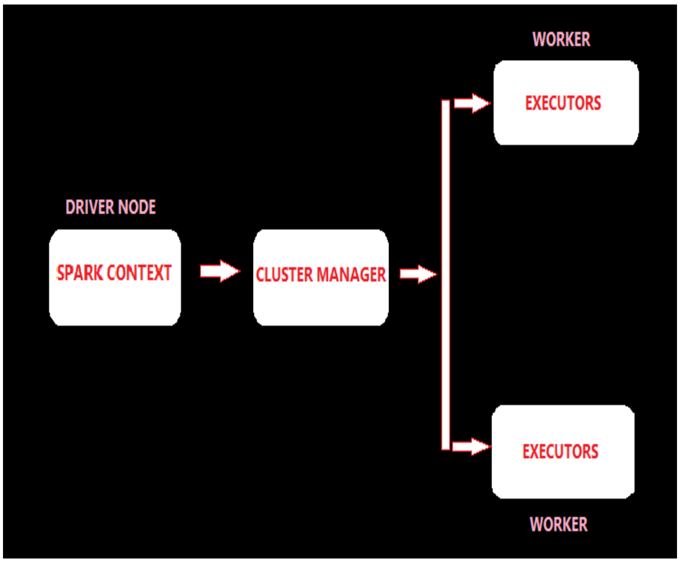

Fig 2. Apache Spark Architecture

Apache spark is having layered architecture. In Apache spark master worker architecture is used where in master node there is driver program to drive the applications. Cluster manager is an external service which is used to acquire services on the cluster. In worker node executors are there which acts as a process to launch the application that is used to run task and keep data in memory across them. For all applications are they are provided their own executors.

II. LITERATURE SURVEY

There are many research works done on stock market & many other topics using apache spark. All such studies are shown below in table 1.

Table 1. Review of Literature.

Sr. No.	Author Name	Algorithm	Observation
[1]	Y Xu, Hi LIU, Z Long (2020)	DCPELM method.	A hybrid distributed computing strategy is used.
[2]	GP Gupta, M Kularia (2016)	Chi-squared feature selection. Correlation based feature selection.	Aim is to use apache spark for designing a framework. Scalability and accuracy are also determined for intrusion detection in cyber security. Chi-squared is having the limitation that for large values it gives the error.
[3]	P Zečević, CT Slater, MC juric, AJ Connolly (2019)	Distributed Zones algorithm, cross match algorithm, extension of the lomb-scargle algorithm.	Astronomy extension about spark is shown here. In distributed zones there is no global clock and shared memory.
[4]	Abdul Ghaffar Shoro, Tariq Rahim Soomro (2015)	Twitter Stream API Apache spark technique.	Querying data is much faster in spark as compared to Hadoop. Spark is very competitive in terms of stream processing. Apache spark is having a disadvantage that it is not having the file management system and it is expensive.
[5]	H Lee, E Kweon, M Kim, S Chai (2017)	Analytical techniques, visualization technologies.	It focuses on the things which should be done before investing in big data analytics. In analytical techniques Measurement errors come.
[6]	M. Balaanand, CB Siva parthipan (2018)	KNN algorithm, machine learning.	The objective is to alert the guardians about the impact of amblyopia on youngsters through prescient examination.
[7]	J Archenaa, Anita EAM (2016)	Machine learning algorithm, k-clustering algorithms, graph theory algorithms.	For improvement in data quality, for healthcare management use of data analytics is important. K clustering algorithms can handle only numeric data.
[8]	A palve, RD Sonawane, AD Potgantwar (2017)	Map-Reduce algorithm.	The real time tweets will be analysed and sentiment analysis is done by showing results in the form of graph.
[9]	RC Maheshwar, D haritha. (2016)	Graph algorithms, machine learning algorithm, and optimization algorithm.	It is being evaluated that time series analysis in apache spark is more efficient over Map Reduce.
[10]	MM Seif, EMR hamed. (2018)	Machine learning algorithms, Least Square Algorithm, Sigmoid Algorithm, applied linear regression algorithms.	Sentiment analysis is done to build the hybrid model to handle big data generated through various sources.
[11]	DR budhathoki, D Das Gupta, P jain (2018)	Algorithms used here are data mining algorithm, FP growth algorithm.	For conducting an experiment on the top of Hadoop a framework of spark is used. FP tree may be expensive.
[12]	D Andrešić, P Šaloun (2017)	Partitioning algorithm, broadcast hash join algorithm, k-modes algorithm.	Correlation of stock market data and Sentiment analysis of tweets is being done.
[13]	MF Aljunid, D H manjaiah (2019)	Collaborative Filtering algorithm, alternating least squares (ALS) algorithm, nearest neighbour algorithm, matrix factorization algorithm.	Conclusion is done on the basis of parameter selection which is for ATS algorithm affected by recommender engine to be used. Collaborative filtering algorithms cannot handle fresh items
[14]	R Shyam, BG HB, S Kumar, P Poornachandran (2015)	Machine learning algorithms, Pattern matching, classification and clustering algorithms.	This paper aims to perform BDA operations on smart grid application like automatic demand response & real time pricing.
[15]	L Agnihotri, S Mojarad, N Lewkow, A Essa (2016)	Learning algorithms, Map Reduce algorithms.	The purpose is to know the fundamentals of EDA & learning the use of python. In MapReduce algorithms there is performance degradation.
[16]	A Wijayanto, E Winarco (2016)	Multi-criteria collaborative filtering algorithm, Perceptron	As compared to sequential counterpart spark cluster is having faster running time. Perceptron only work with

		algorithm.	linearly seperable set of vectors.
[17]	J Shanahan, L Dai (2017)	Graph processing algorithm, gradient descent algorithm.	Here techniques used are apache spark and machine learning.
[18]	DEL Berron, DV K Yarlagadda, P Rao (2018)	Deep learning algorithms.	The system is being shown for fast retrieval and scalable storage of the tiles.
[19]	S Salloum, R Dautov, X Chen, PX Peng (2016)	big data algorithms, scalable data algorithms, machine learning algorithms.	Review of the features of spark for BDA is shown in this paper.
[20]	Y Yan, L Huang, L Yi (2015)	Seismic data processing algorithms.	SAC is developed and estimated to be a better applicant for processing of seismic data.
[21-27]	Gupta, Y., et. al. (2016-20)	Apache Pig.	Authors Analyzed various datasets such as stock exchange Data, Crime rates of India, Population of India, and Healthcare data using apache Pig. Also elaborated various tools and techniques used to analyze massive volume of data i.e. big data in hadoop distributed file system of cluster of commodity hardware. Authors also describe various image processing techniques.
[28]	Smys Smys, Jennifer S Raj (2019).	Big data analytics, internet of things.	Author has explained the concept of internet of things and big data application on the behalf of that. Also the importance of BDA and IOT in rural areas to improve the facilities for patient.
[29]	Iwin Thanakumar Joseph Swamidason (2019).	Data mining algorithms.	Here author has done analysis on data mining algorithms and intelligent computing system. Conclusion of the survey along with the limitations is being examined.
[30]	SM Idrees, MA Alam, P Agarwal (2019).	ARIMA model, time series analysis.	The objective of this paper is to show the stock market prediction and there difference using time series data. it is concluded that ARIMA approach is good for handling time series data.

RESEARCH GAP

Many papers are analyzed they have done work on other things, but nobody has done on apache spark. Data analysis of stock market nifty 50 will be done on the behalf of covid-19. The data will be analyzed so that the effect of covid-19 can be evaluated on stock market nifty 50. Apache spark technique is being used for analysis work. Till now no work has been done on nifty 50 data for finding the impact of covid-19 using apache spark.

III. COMPARATIVE STUDY BETWEEN APACHE SPARK AND MAP-REDUCE.

After going through the exhaustive study corresponding to the apache spark and hadoop Mapreduce in both the aspect like practical and theoretical, some differences between them are found that are further elaborated on the behalf of some parameters in the table 2 shown below.

Table 2. Comparison between apache spark and Map-Reduce.

S.NO	PARAMETERS	APACHE SPARK	MAP-REDUCE
1.	performance	faster	Comparatively slower
2.	Ease of use	Easy to use	Difficult to program
3.	cost	More cost effective	cheaper
4.	Data processing	Real time as well as batch processing	Batch processing
5.	Security	Less secure	More secure
6.	Fault tolerance	Recover loses after it occurs.	Highly fault tolerant
7.	Category	Data analytics engine	Data processing engine
8.	Latency	Much faster	Very high latency
9.	Complexity	Easy to write and debug	Difficult to write and debug
10.	coding	More lines to code	Lesser lines to code

IV. DISCUSSION

Covid-19 has affected many life's and almost everything is affected so as stock market. Stock market data is a kind of data which changes very frequently that's why prediction in this field is not an easy task. After doing literature review of many papers, it has been decided to work on apache spark using Scala. It is necessary to use data analytics technique apache spark which overcome the limitation of map reduce on the behalf of time and performance. Execution time of apache spark is 10 minutes whereas in map reduce it is taking 50 minutes. Here comparative analysis is done between apache spark and Hadoop map-reduce and it is observed that apache spark plays better role as compared to Hadoop map-reduce.

V. CONCLUSION

In this paper, it is concluded that big data analytics acts as an important part in data processing and data analysis of stock market data so, after reading so many papers, comparative analysis is done in between apache spark and Hadoop map-reduce. Apache spark is having very fast processing and due to its processing capabilities only, it is very much known today. So, finally on the behalf of comparative analysis done it is being identified that apache spark is much better then Hadoop map-reduce to analyze the stock market nifty-50 data.

REFERENCES

[1] Xu Y, Liu H, Long Z. A distributed computing framework for wind speed big data forecasting on Apache Spark. Sustain Energy Technology, pp. 1-14 2020.

[2] Gupta GP, Kulariya M A framework for fast and efficient cyber security network intrusion detection using apache spark. Procedia Comput Sci, pp. 824–831, 2016.

[3] Zečević, P., Slater, C. T., Jurić, M., et al. AXS: A Framework for Fast Astronomical Data Processing Based on Apache Spark, pp. 1-14, 2019.

[4] A. Ghaffar, S. & Tariq, R. Soomro, A. G. Shoro, and & Tariq, "Big Data Analysis: Ap Spark Perspective," Glob. J. Comput. Sci. Technol. Glob. Journals Inc. Glob. J. Comput. Sci. Technol., pp. 7–14, 2015.

[5] Lee, H., Kweon, E., Kim, M., & Chai, S. Does Implementation of Big Data Analytics Improve Firms" Market Value? Investors" Reaction in Stock Market. Sustainability, pp. 1057-1073 2017.

[6] BalaAnand M, Sivaparthipan CB, Karthikeyan N, Karthik S Early Detection and Prediction Of Amblyopia By Predictive Analytics Using Apache Spark. International Journal of Pure and Applied Mathematics (IJPAM), pp. 3159–3171, 2018.

[7] J. Archenaa and E. A. M. Anita, "Interactive big data management in healthcare using spark", Proc. 3rd Int. Symp. Big Data Cloud Comput. Challenges, pp. 265-272, 2016.

[8] Amit Palve, Rohini D.Sonawane, Amol D. Potgantwar, "Sentiment Analysis of Twitter Streaming Data for Recommendation using Apache Spark", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.3, pp.99-103, 2017.

[9] R. C. Maheshwar and D. Haritha, "Survey on high performance analytics of bigdata with Apache Spark", Int. Conf. on Advanced Communication Control and Computing Technologies, pp. 721-725, 2016.

[10] M. M. Seif, E. M. R. Hamed and A. E. F. A. G. Hegazy, "Stock market real time recommender model using apache spark framework", International Conference on Advanced Machine Learning Technologies and Applications, pp. 671-683, 2018.

[11] Budhathoki, D., Dasgupta, D., & Jain, P., "Big data framework for finding patterns in multi-market trading data" Springer, 2018.

[12] D. andresic, P., Saloun & I. Anagnostopoulos., "Efficient big data analysis on a single machine using apache spark and self-organizing map libraries" 12th international workshop on semantic and social media adaptation and personalization, pp. 1-5, 2017.

[13] Aljunid MF, Manjaiah DH., "Movie Recommender System Based on Collaborative Filtering Using Apache Spark. InData Management, Analytics and Innovation Springer, Singapore pp. 283-295, 2019.

[14] R Shyam, Sachin Kumar, Prabaharan Poornachandran, and KP Soman. Apache spark a big data analytics platform for smart grid. Procedia Technology, pp. 171– 178, 2015.

[15] AGNIHOTRI, L., S. MOJARAD, N. LEWKOW et A. ESSA., "Educational Data Mining with Python and Apache Spark: A Hands-on Tutorial". In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. LAK "16. Edinburgh, United Kingdom: ACM, pp. 507–508,2016.

[16] Wijayanto, A., & Winarko, E., "Implementation of multi-criteria collaborative filtering on cluster using Apache Spark". In Science and Technology-Computer (ICST), International Conference on IEEE, pp. 177-181, 2016.

[17] Shanahan, J.; Dai, L., "Large Scale Distributed Data Science from scratch using Apache Spark 2.0". In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia; International World Wide Web Conferences Steering Committee: Geneva, Switzerland; pp. 955–957,2017.

[18] Barron, D. E. L., Yarlagadda, D. V. K., Rao, P., Tawfik, O., and Rao, D., "Scalable storage of whole slide images and fast retrieval of tiles using apache spark," in [Medical Imaging 2018: Digital Pathology], 10581, 1058113, International Society for Optics and Photonics 2018.

[19] S. Salloum., "Big data analytics on Apache Spark, Int. J. Data Sci. Anal., vol. 1, no. 3, pp. 145–164, 2016.

[20] Y. Yan et al., "Is apache spark scalable to seismic data analytics and computations?" in IEEE Int.Conf. Big Data, pp.2036-2045,2015.

[21] Gupta, Y. K. & Sharma, S. (2019). Impact of Big Data to Analyze Stock Exchange Data Using Apache PIG. International Journal of Innovative Technology and Exploring Engineering. ISSN: 2278-3075, 8(7), Pp. 1428-1433.

[22] Gupta, Y. K. & Sharma, S. (2019). Empirical Aspect to Analyze Stock Exchange Banking Data Using Apache PIG in HDFS Environment. Proceedings of the Third International Conference on Intelligent Computing and Control Systems (ICICCS 2019).

[23] Gupta, Y. K. & Gunjan B. (2019). Analysis of Crime Rates of Different States in India Using Apache Pig in HDFS Environment. Recent Patents on Engineering. Print ISSN: 1872-2121, Online ISSN: 2212-4047, 13:1. https://doi.org/10.2174/1872212113666190227162314. site:http://www.eurekaselect .com/node/170260/article.

[24] Gupta, Y. K.* & Choudhary, S. (2020). A Study of Big Data Analytics with Two Fatal Diseases Using Apache Spark Framework. International Journal of Advanced Science and Technology (IJAST), Vol. 29, No. 5, pp. 2840 - 2851.

[25] Gupta, Y. K.*, Kamboj, S. & Kumar, A. (2020). Proportional Exploration of Stock Exchange Data Corresponding to Various Sectors Using Apache Pig. International Journal of Advanced Science and Technology (IJAST), Vol. 29, No. 5, pp. 2858 - 2867.

[26] Gupta, Y. K.* & Mittal, T. (2020). Empirical Aspects to Analyze Population of India using Apache Pig in Evolutionary of Big Data Environment, International Journal of Scientific & Technology Research (IJSTR). ISSN 2277-8616, 9(1), Pp. 238-242.

[27] Gupta, Y. K. & Jha, C. K.(2016). A Review on the Study of Big Data with Comparison of Various Storage and Computing Tools and their Relative Capabilities. International Journal of Invocation in engineering & technology (IJIET). ISSN: 2319-1058, 7(1), Pp. 470-477.

[28] Smys, S., and Jennifer S. Raj. "Internet of things and big data analytics for health care with cloud computing." J. Inf. Technol 1,pp. 9-18,2019.

[29] Joseph, S. Iwin Thanakumar, and Iwin Thanakumar. "Survey of data mining algorithm's for intelligent computing system." Journal of trends in Computer Science and Smart technology (TCSST) 1,pp. 14-24,2019.

[30] S. M. Idrees, M. A. Alam, and P. Agarwal, "A prediction approach for stock market volatility based on time series data," IEEE Access, vol. 7, pp. 17 287–17 298, 2019.