IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 10, October 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 10, October 2020 ||

Big Data and Cluster based Innovation for Map Reduction

M. Mahima[1], D.R.Anitha Sofia Liz[2], Venkatesh N[3], Elakkiya T[4], A.George Amal Jose[5]

Department of Information Technology, New Prince Shri Bhavani College of Engineering and Technology, Chennai, India[1,3]

Department of Computer Science and Engineering, New Prince Shri Bhavani College of Engineering and Technology, Chennai, India[2,4]

Foreview Technologies Pvt. Ltd., Chennai, India[5]

ABSTRACT: Map-Reduce is programming structure that permits specific sort of parallelizable or distributable issues including cumbersome data sets to be settle utilizing processing clusters. This paper presents a hybrid Map-Reduce system that accumulates calculations assets from various clusters and runs Map-Reduce occupations across them. The instrument is acknowledged utilizing DBSCAN clustering algorithms among Map-Reduce, equal handling structure over clusters. Nonetheless, the moment achievement of algorithms goes through from productivity issue for higher execution time as well as lacking memory. This paper presents an effective DBSCAN clustering technique for mining huge datasets with apache Hadoop and Map-Reduce that will reduce season of open algorithms and dataset from the disparate area will work at the same time from single hub and track down the suitable result in conveyed climate.

KEYWORDS: Big Data, Hadoop, Map-Reduce, Data Mining and Data Clustering.

I. INTRODUCTION

Big data comes with four features [2]: volume, velocity, variety and value, which build scalable and fault tolerance more and more significant for data organization. MapReduce systems such as Hadoop are considered more suitable to these new requirements. However, as a novel computing engine, MapReduce systems are not perfect; the key problem is the efficiency. To improve hadoop's efficiency, lots of work has been done such as [3][5][8][9]. However, scheduling, the most important, complicated and challengeable problem, has not extensively and deeply studied decisions and add worth to their business.

Data mining is a combination of three main factors: Data, Information and knowledge. Data are the most elementary description of the things, events or the activity and transactions. Information is organized data which have some valuable meaning or some useful data. Knowledge is a concept of understanding information based on the recognized pattern or algorithms that provide the information. Data Mining is a technique of finding valuable knowledge from the large amount of dataset [1][2][3]. Main techniques for data mining are classification and prediction, clustering, outlier analysis, association analysis, evolution analysis.

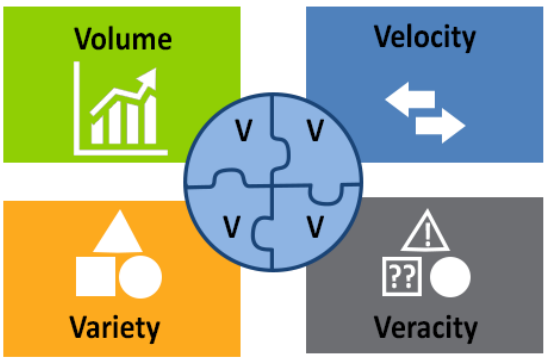

Figure.1 Four V's of Big Data

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 10, October 2020 ||

Clustering is one of the techniques applied on the unsupervised dataset. Different types of clustering methods are hierarchical, partition, Density Based method and Grid based method. DBSCAN is one of the density based method. DBSCAN can find arbitrary shape. This algorithm is also sufficient for the spatial dataset and also for the large dataset. There are many different algorithm invented from the original DBSCAN algorithm [4]. Performing DBSCAN algorithm in the real world application is challenging due to mainly two reasons. First is increasing large amount of dataset rapidly so single machine user cannot handle it or having trouble to handle using single user. Second is cost of DBSCAN algorithm i.e much higher computation time with respect to other clustering algorithms. Thus many existing studies try to improve efficiency of the DBSCAN algorithm. For example FDBSCAN [7] can find the arbitrary shape same as DBSCAN algorithm but time complexity of this algorithm is less than the original DBSCAN algorithm. ODBSCAN [8] is also find arbitrary shape but the main different is identical circles are predefine in this method. VDBSCAN [9] is also find arbitrary shape and also can find cluster in varied density. ST-DBSCAN [10] is also find cluster in arbitrary shape and all this methods are more efficient than DBSCAN algorithm.

II. LITERATURE REVIEW

Density Based algorithm is one of the approach for the clustering algorithm. It is mainly based on the core points, border points, and density reachable points. In this approach data which are cover in the denser region will be grouped and form one cluster. They use some threshold value to determine denser region. DBSCAN is one of the density based clustering algorithm with noise. DBSCAN can find cluster in arbitrary shape and filter noise. DBSCAN is not effective in massive and varied dataset. It has lower time complexity. To achieve the problem of the lower time complexity new method invented that is IDBSCAN [11]. Apache-Hadoop [12] is one of the open-source platforms to perform distributed data mining for the big data. Hadoop software document is a framework that permits for the dispersed processing of the large dataset across cluster of computers using single programming models. It is designed as way so that from the single servers, thousands of machine's data can be mine in the scalable manner. Each local machine contribution calculation and storage space, quite rely on hardware to carry high-availability, the library itself is intended to detect and manage failures on the application layer, thus delivering a extremely-available service over a cluster of computers, both of which might be prone to failures. Hadoop is combination of two main components. First is a HDFS that is Hadoop distributed File System and second is the Map/Reduce. Minimal map reduce [12] algorithm is used to sort the data and joining it very efficiently. We combine both DBSCAN and Minimal map reduce algorithm to make it hybrid and very efficient. Tera sort won the annual general purpose terabyte sort benchmark in 2008 and 2009.

III. PROPOSED METHODOLOGY

The proposed scheme will work with DBSCAN and Tera sort algorithms in distributed environment using Hadoop. In proposed system weather datasets are used. Below fig2 shows the architecture of proposed system which contains input data sets of weather data; dataset split operation, task tracker and job tracker are the inbuilt function of Hadoop MapReduce, mapper is to map the splitted data before applying hybrid mechanism on splitted data with respect to the dimension or attribute of the dataset. After successful mapping of splitted data hybrid mechanism will apply on that mapped data, in hybrid mechanism we perform DBSCAN clustering as well as tera sort sorting and Reducer reduced the query with respect to Distributed Fie System (DFS). DBSCAN algorithm will be apply on different site and create local cluster and global cluster will be the master node which having multiple local cluster.

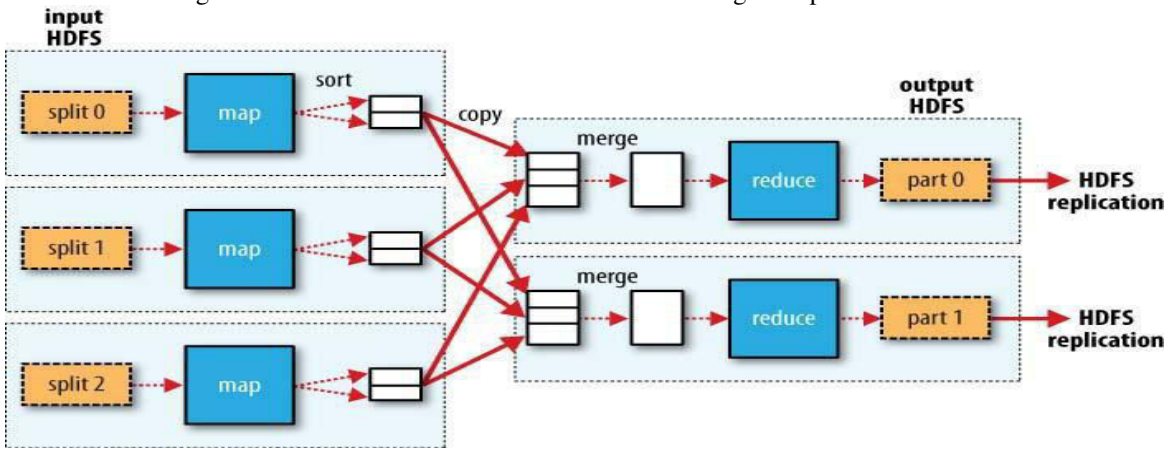

Figure.2 Proposed Architecture

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 10, October 2020 ||

IV. RESULTS AND DISCUSSION

The hardware configuration of our test systems are physical machines with Intel 2.4 Quad core i7 processor, 1 TB HDD with 8 GB RAM. The software of Hadoop 2.0.0 cluster configurations are 1024 MB Hadoop heap size. The virtual node configurations are 256 MB memory. In the proposed system dataset are distributed among the different sites that means data are spatial dataset. In the initial phase clusters will be generated at different site using DBSCAN algorithm. After that all that clusters will be send to the Master Node of the Hadoop system. Noise in the dataset will remove at individual site only in the initial phase and store in .csv file. The hybrid mechanism is implemented in JAVA language using net beans IDE. Proposed mechanism tested on Windows 8.1 64 bit using Hadoop setup. Here for the algorithm basic input parameter Minpts is taken as 6 and Epsilon is taken as 0.9.

V. CONCLUSIONS

Contrast with focal data mining clustering strategies, circulated data mining is more effective, versatile and execution is superior to the focal data mining procedures. We propose Hybrid component which works on the presentation by taking benefits of data region. We use DBSCAN clustering and Tera sort arranging method. Hybrid instrument can give better execution in disseminated climate as far as run time intricacy utilizing Hadoop platform, we can reduce execution assessment time.

REFERENCES

1. Katarina Grolinger, Michael Hayes, Wilson A. Higashino, Alexandra L'Heureux David S. Allison, Miriam A.M. Capretz, “Challenges for MapReduce in Big Data”, IEEE 10th World Congress on Services, 2014.
2. F. Ohlhorst, “Big Data Analytics: Turning Big Data into Big Money”, Hoboken, N.J, USA: Wiley, 2013.
3. Maitry Noticewala, Dinesh Vaghela, “MR-IDBSCAN: Efficient Parallel Incremental DBSCAN Algorithm using MapReduce”, International Journal of Computer Applications (0975 – 8887) Volume 93 – No 4, May 2014.
4. Han, P.N., Kamber, M., “Data Mining: Concepts and Techniques”, (2006).
5. Tan, P.N., Steinbach, M., Kumar, V, “Introduction to Data Mining”, (2006).
6. Shou Shui-geng, zhou Ao-ying Jin Wen, Fan ye and Qian Wei-ning "A Fast DBSCAN Algorithm", 20006.
7. J. Hencil Peter, A. Antonysamy “An Optimised Density Based Clustering Algorithm” International Journal of Computer Applications (0975 – 8887) Volume 6– No.9, September 2010.
8. Wei Wang, Shuang Zhou, Bingfei Ren, Suoju He”improved vdbscan with global optimum k” ISBN: 978-0-9891305-0-9 ©2013 SDIWC.
9. Derya Birant, and Alp Kut, “ST-DBSCAN: An algorithm for clustering spatial-temporal data”, Data Knowl. Eng. (January 2007).
10. Cheng T. Chu, Sang K. Kim, Yi A. Lin, Yuanyuan Yu, Gary R. Bradski, Andrew Y. Ng, and Kunle Olukotun, “Map-Reduce for Machine Learning on Multicore”, NIPS, page 281--288. MIT Press, 2006.
11. Navneet Goyal, Poonam Goyal, K Venkatramaiah, Deepak P C, and Sanoop P S, “An Efficient Density Based Incremental Clustering Algorithm in Data Warehousing Environment” 2009 International Conference on Computer Engineering and Applications IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore.
12. Yufei Tao, Wenqing Lin, Xiaokui Xiao, “Minimal MapReduce Algorithms”, SIGMOD13, June 22-27, 2013, New York, USA.
13. http://hadoop.apache.org/
14. http://www01.ibm.com/software/data/infosphere/hadoop/
15. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, “Data Mining with Big Data”, IEEE transaction on knowledge and data enigineering, vol. 26. No. 1, Jan 2014.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details