ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Enhancing Map-Reduce Mechanism for Big Data with Density-Based Clustering

Vinod S. Bawane, Deepti P. Theng

Student M.E.(IV Sem), Dept. of CSE, G. H. Raisoni College of Engineering, Nagpur, India

Assistant Professor, Dept. of CSE, G. H. Raisoni College of Engineering, Nagpur, India

ABSTRACT: Map-Reduce is software framework that allows certain type of parallelizable or distributable problems involving bulky data sets to be solve using computing clusters. This paper presents a hybrid Map-Reduce framework that gathers computations resources from different clusters and runs Map-Reduce jobs across them. The mechanism is realized using DBSCAN clustering algorithms among Map-Reduce, parallel processing framework over clusters. However, the instant accomplishment of algorithms undergoes from efficiency problem for higher execution time as well as inadequate memory. This paper presents an efficient DBSCAN clustering method for mining large datasets with apache Hadoop and Map-Reduce that will reduce time of accessible algorithms and dataset from the dissimilar location will work simultaneously from single node and find the appropriate outcome in distributed environment.

KEYWORDS: Data mining, Data clustering, DBSCAN, Hadoop, Map-Reduce

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.

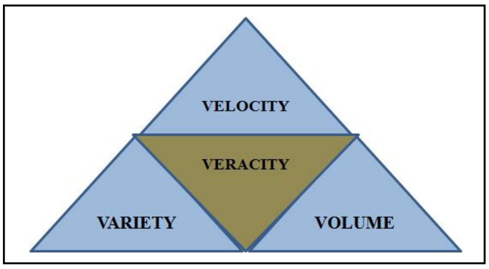


Fig1: 4 V's of Big Data

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.

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

10.15680/ijircce.2015.0304115

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.

Remain paper is organized as follows. In section 2 we discuss several clustering methods and as well as several related to DBSCAN method. In section 3 we introduce proposed mechanism that is hybrid mechanism, section 4 is related to experimental set up and result analysis, and conclusion and future work is given in section 5 and 6.

II. BACKGROUND AND RELATED WORK

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.

A. *DBSCAN:*

IDBSCAN algorithm is capable of adding points in the bulk to existing set of clusters. In this algorithm data points are added to the first cluster using DBSCAN algorithm and after that new clusters are merged with existing cluster to come up with the modified set of the clusters. In this algorithm clusters are added incrementally rather than adding points incrementally. R*- tree data structure is use in this algorithm. In this algorithm new data points which intersect with old data points are determine. For each intersection point, new dataset use incremental DBSCAN algorithm to determine new cluster membership. Cluster memberships of the remaining points are then updated.

Here existing clusters are referring as old cluster and cluster points added are referring as new clusters. By adding the new data points following transitions are possible. In this case Eps and Minpts will be same.

It may possible that old Noise points may become border point or core point in the new cluster formation. Border point in the old cluster may become core point in the new cluster and core point of old cluster may become core point of the new cluster.

Case 1: The affected point to see if any point is density reachable from core point of old cluster, cluster membership is changed and it become core point of the new cluster.

Case 2: If the affected points are density reachable from the core point of the old cluster than these two clusters will be merged.

Case 3: If border point of old cluster becomes core point of new cluster than two clusters will merge. If the border point is not becoming a core point, then it retains its cluster membership. In all the above cases, if the new point is not an intersection point, then its cluster membership will not change.

Case 4: If any point becomes a border point of a new cluster, then it will be absorbed in the cluster. If it becomes a core point, then formation of a new cluster or merging of clusters may happen.

B. *MINIMAL MAP REDUCE*

Minimal MapReduce Algorithms denote by S the set of input objects for the underlying problem. Let n, the problem cardinality, be the number of objects in S, and t be the number of machines used in the system. Define m = n/t, namely, m is the number of objects per machine when S is evenly distributed across the machines. Consider an algorithm for solving a problem on S.

We say that the algorithm is minimal if it has all of the following properties.

- **Minimum footprint:** at all times, each machine uses only O(m) space of storage.
- **Bounded net-traffic:** in each round, every machine sends and receives at most O(m) words of information over the network.
- **Constant round:** the algorithm must terminate after a constant number of rounds.
- **Optimal computation:** every machine performs only O(Tseq/t) amount of computation *in total* (i.e., summing over all rounds), where Tseq is the time needed to solve the same problem on a single sequential machine. Namely, the algorithm should achieve a speedup of t by using t machines in parallel.

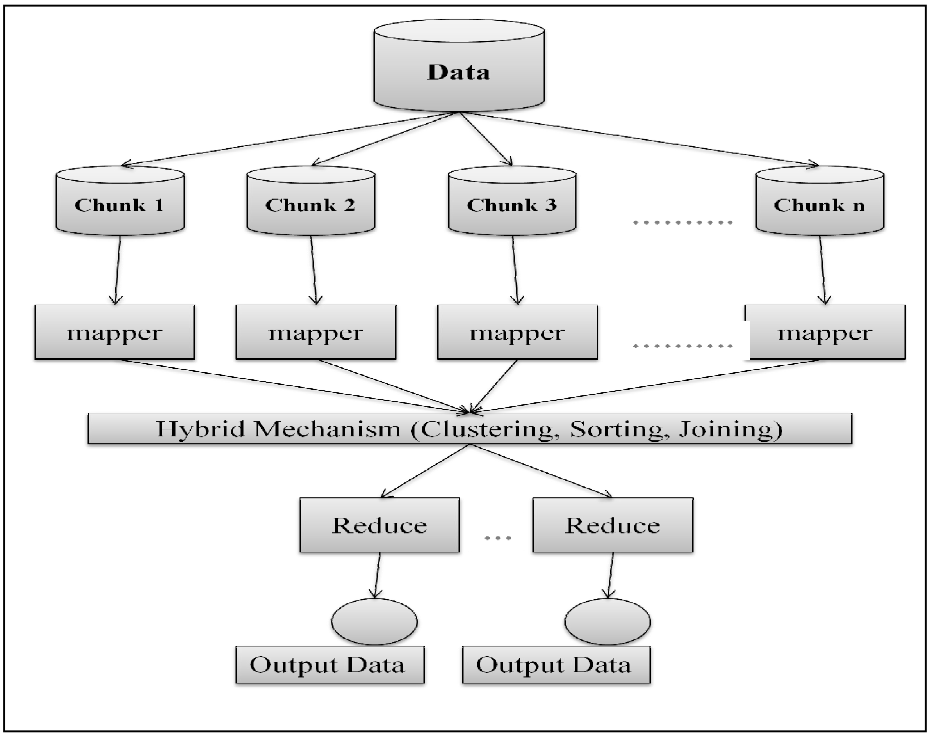
The core of this work comprises of neat minimal algorithms for two problems one is Sorting and other Sliding Aggregation.

III. PROPOSED MECHANISM

A. *Overview of Proposed work*

In propose work I am planning to 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.



IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

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.

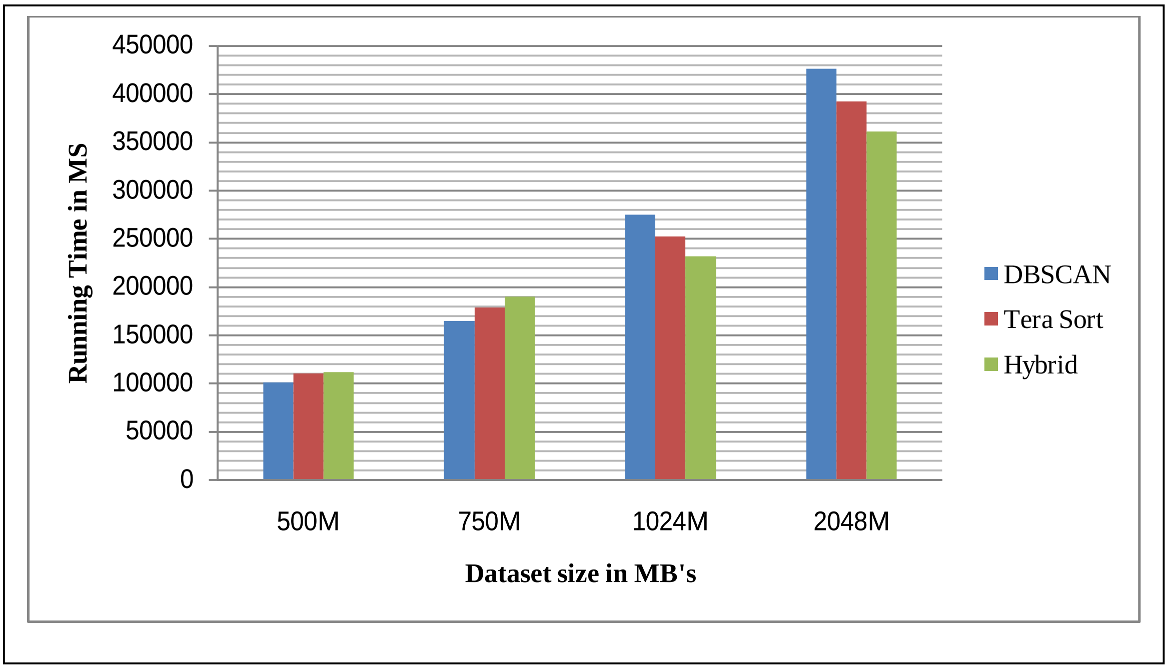
In the proposed system dataset are distributed among the different sites that means data are spatial dataset. In the initial phase clusters will be generated at different site using DBSCAN algorithm. After that all that clusters will be send to the Master Node of the Hadoop system. Noise in the dataset 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.

Table1. Running time of different mechanism in ms

Sr. No	Algorithm	500M	750M	1024M	2048M
1	DBSCAN	101520	165000	274900	426100
2	Tera Sort	110500	179000	252500	392500
3	Hybrid Mechanism	112200	190102	232000	361000

In table above results comparison is done between running time and different dataset sizes of 500M, 750M, 1024M and 2048M. Here by the use of charts it is shown that the time taken by the Hybrid mechanism is comparatively less than the DBSCAN and Tera sort methods.



V. CONCLUSION

Compare to central data mining clustering techniques, distributed data mining is more efficient, scalable and performance is better than the central data mining techniques. We propose Hybrid mechanism which improves the performance by taking advantages of data locality. We use DBSCAN clustering and Tera sort sorting technique. Hybrid mechanism can give better performance in distributed environment in terms of run time complexity using Hadoop platform, we can reduce performance evaluation time.

VI. FUTURE WORK

Future work will include experimenting with other variations of DBSCAN and sorting methods for much better speed. Also may involve efforts to reduce the duplicate work by using better partitioning scheme and achieve higher speedups.

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

BIOGRAPHY

Vinod S Bawane is a PG student at G. H. Raisoni College of Engineering, Nagpur. He is from Mobile Technology in Computer Science and Engineering Department. His research interest includes data mining, big data, and data visualization.

Deepti P. Theng is an Assistant Professor at G. H. Raisoni College of Engineering, Nagpur. She is from Computer Science and Engineering department. Her research interest includes distributed computing; cloud computing, operating system, and high performance computing architecture**.**