IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-8 E-ISSN: 2347-2693

Performance Improvement of Heterogeneous Hadoop Clusters Using MapReduce For Big Data

P. Dadheech[1*], D. Goyal[2], S. Srivastava[3]

[1*] Dept. of CSE, Suresh Gyan Vihar University, Jaipur, India
[2] Principal, Suresh Gyan Vihar University, Jaipur, India
[3] Dept. of ICT, Manipal University, Jaipur, India

[*]*Corresponding Author: pankajdadheech777@gmail.com*

Available online at: www.ijcseonline.org

Received: 05/Jul/2017, Revised: 17/Jul/2017, Accepted: 19/Aug/2017, Published: 30/Aug/2017

***Abstract*—** The problem that has occurred as a result of the increased connection between the device and the system is creating information at an exponential rate that it is becoming increasingly difficult for a possible solution for processing. Therefore, creating a platform for such advanced level data processing, which increase the level of hardware and software with bright data. In order to improve the efficiency of the Hadoop Cluster in large data collection and analysis, we have proposed an algorithm system that meets the needs of protected discrimination data in Hadoop Clusters and improves performance and efficiency. The proposed paper aims to find out the effectiveness of the new algorithm, compare, consultation, and find out the best solution for improving the big data scenario is a competitive approach. The map reduction techniques from Hadoop will help maintain a close watch on the underlying or discriminatory Hadoop clusters with insights of results as expected from the luminosity.

***Keywords*—** Big data, hadoop, heterogeneous clusters, map reduce, throughput, latency.

I. INTRODUCTION

This study has shown so far that about 30% of the information displayed from the digital world can be useful for various purposes for analysis accurately. However, only 0.5% of the data is used in the existing data, which clearly indicates the inefficiency of large data fields due to limited efficiency in the existing ways. The current analysis of unorganized data in servers or clusters shows that the information on the system has isolated effects. The map reduction model can provide high performance, job distribution or less fairness in Venezuela. But the speed at which it has been done is to improve and to improve in order to deal with unorganized data. Another thing that will be taken care of, effective time, effective clustering strategies, algorithms for formatting information, structure and retrieve information with high savings and fewer visitors. Hadoop, the open source framework for saving and analyzing data from low-cost hardware, has created a major disturbance between the developers and the organization. Hadoop entertains data storage, analyzes, and recovery using the operating node's cluster under it to reveal the chances of screening big data sets that are silent for the relational database. It is designed and supported in such a way that a single framework is sufficient for thousands of servers quickly performed by local computers and storage. This feature became effective in the rapid development, that it could monitor the irregular and secure information that dominated the Internet by using the constraints of the map. Big data that is so large and complex that unable to save the traditional database, or it includes information on the analysis. Large data is understood by four V, such as variations, speeds, values, and volume.

Large data is predicted to achieve values from a very large database. Analysis and processing involved can determine how much value an information or frame can provide. It prints with information that can provide some useful and meaningful content by saving, sharing, searching, imagining or questioning. Business intelligence and statistics are deeply linked to it. All related to language processing, artificial intelligence, astronomy, genomics, weather or other relative fields, like any other area, which contain bright amount of information. Data is made with today's exponential quantity to deal with these databases; Big data involves prediction, some complicated calculations and algorithms involved in and involved in the value available. The upcoming years will be huge in terms of large scale measurement and processing applications. There are many applications and frameworks that use large data, but many point losses are encountered. There is an algorithm that delivers data based on their own principals that provide a smooth utility for clusters. This way, the algorithms used provide a way to access all data accounts and connect them with the time of map loads. Objectives Identify existing research with Hadoop, Big Data, and the Map Blaster Framework and work with commonly encountered problems, deal with large sets of statistical or

unorganized data and review them for research work, perform comparisons. Existing algorithms study that is used to use clustering data for supernatural Hadoop clusters, and one of them best designs for future trends, designs a new algorithm that errors in existing algorithms provide a better effect than earlier ones and reducing the map for larger data to use. Model business Hadoop cluster and the algorithm to improve the performance of the discriminatory child show results. The proposed work identifies and improves such challenges to create a better and useful version of the community than the previous ones. With great data, anyone can improve the new world of sales, marketing, new drug discovery, exit strategies and a new job. Hadoop stands out most of the problems in managing large data from unlabelled objects which are difficult to deal with in clusters.

In this Paper, Section I contains the introduction of Hadoop Clusters Using MapReduce For Big Data, Section II contain the related work of Hadoop, Big Data, MapReduce & different Job Scheduling Algorithm used in Hadoop, Section III contain the proposed methodology to improve the performance of heterogeneous clusters, Section IV contain the results and discussion of the proposed methodology, Section V contain the conclusion of the proposed research work with future directions.

II. RELATED WORK

Internet is expanding at the exponential rate; Assess the fact that the present situation of the scientists is a matter of few years time, the information will exceed thousands of petabytes of data, in order to burst information such as information or special data, only a good system needs to recover data and not only a good system but it is able to question and analyze may be. Big data scenarios themselves offer challenges that generally store data, which are very contradictory. Primarily that information is important for today's scientific, retail, or business-based. So to raise some profits or simply to publish data to understand patterns, one must deeply immerse one and conclude a meaningful explanation. There are many platforms to store information. Traditionally, relational databases release data by preserving and analyzing different types of data, but the type of data limits and types of data exceed current trends, it becomes very important to save unwanted or unorganized data. Here the unorganized data represents any data in raw form and is only a collection of a set where one part may differ from the other.

Jolyu said that 33% of digital data can be proven valuable if we analyze the right tools and methods but it only uses up to 0.5%. These types of observations are very efficient, measurable, and flexible for large-scale data analysis. The map has proved their importance as a skilled, scalable, and fault-tolerant data analysis programming model by reducing algorithms in a short time, but there is a lack of query-level terminology in the work schedule, which can be used to reduce the amount of system resources, prolonged exposure time and low query troupe . The map reduces task-level scheduling through guided Acyclic graph (DAG) and sets support languages with data warehousing capabilities. The scalability of Hadoop greater cluster is usually contradictory. Consider where the content is collected on the continents of a website. Generally, it will include different platforms users' language and format according to their own. There should be discrepancy with the variance of the website and it is a big task to save such information on clusters. Include clustering preference will be attached to those data with enticed value so that the cluster can easily be used to identify and query the cluster configuration as noted by [jingban jou, et al] as parallel data in the Hadoop data parallel effects. In contrast to the clusters of clusters, the challenge is to produce and manage the infrastructure for compositioning not only on existing computing platforms, but also for comparatively good results with low cost equipment. Computing is basically essential tools to deal with structured and unorganized data. Hadoop's contradiction is the lack of cluster performance, which further processes implementation of the problem. For example, according to the proverbial kametkar, the effectiveness has decreased, hardware configurations, logical connectivity and appropriate underlying structure may be facing problems of heterogeneous hopper clusters. To avoid a note failure of the clients, replication techniques, rack maintenance, a minimum number of clusters of a cluster can be operated by following the minimum requirements. As mentioned by [Jiong GE], there is clearly a difference where performance clusters in less local data transfer are exceeded by high performance clusters. Delivering loads of high and low performance contrast clusters, reducing a significant improvement in performance on the map.

III. METHODOLOGY

The experimental work follows a set of phases to improve the performance of heterogeneous clusters, which improves the data input / output queries, improves the routing of the Herodotus cluster algorithm and then improves the efficiency of the query processing. They are easily connected to the right part of the execution in perfect time without increasing the cost of the computing level. The proposed work follows a series of steps to follow these clauses which are described in the following sections.

A series of steps or steps is added to improve the efficiency of contrast clusters. Each step is explained below:

A. Query Improvisation: When questions are brought in parser and semantic analyzer, they evaluate dependencies. But once this purse query was sent MapReduce of Hadoop is lost in the transition to run dependency that is calculated for the hip query. Depending on the dependency, they can be

used for the Hive QL processor's Cementers Extensions. In the second step, we can use these dependencies such as logical prediction and input tables, depending on the reliability of different signals being closely linked to each other while processing. When the interval between Honey Chandrima and Hadoop is bound by these intermediate steps, they can be easily used for querying similar to clustering at the query level, and therefore by minute steps to improve the performance of the query work.

B. Query Scheduling Algorithm: Hadoop uses First In First Out (FIFO) scheduler for job distribution by default. But there are other schedulers as well like, Hadoop Fair Scheduler (HFS) and Hadoop Capacity Scheduler (HCS) that are customizations for Hadoop ecosystem. To have a good fairness among different jobs, the HCS and HFS are sufficient. In the proposed methodology, both the schedulers are used one by one to find the outcome which can help in maintaining order among different capacity clusters.

C. Clustering Algorithm: The algorithm proposed under this dissertation work is to simplify the need to categorize big data that arrives in dynamic fashion. If the clusters of data could be improved effectively.

The algorithm that is used for this proposed work is as following:

ALGORITHM

Input: Datasets cleansed of garbled values.
Tq = Read_queries from Query.DB();
fq=Frequency of item dataset
Ci= item in Cluster i
ds = Similarity matrix from Clustering through K- Means Algorithm.

COMPUTE FQ

Output: Clustered data of importance.

1. Step 1:- We will read the log transition.
2. Step 2:- We store the transition value in Qi, Where i=1, 2, 3……..n
3. Where n is equal to the log transition value.
4. Step 3:- We store the query which is request from client and store in array by get Query () method.
5. IQ = I get query. Where I is number of query request and .We store the query in array to create
6. cluster. By method table – put (null, array, parameter1, parameter 2 ….n);
7. To convert the array in object they are
8. Qi= Table – get query ();
9. Step 4:- Merge the pair of most similar queries (qi,qj) that does not convert the same queries. If
10. (Qi is not in Ci) then store the frequency of the item and the increase the value.
11. Step 5:- Compute the simulating Matrix if Qi in Ci.
12. Step 6:- If Qi is not Ci then compute new cluster IQ =New (Ci).
13. Step 7:- Go to Step 3.
14. Step 8:- Step go to Step 2.

Pseudo code for query mapping

Class Mapper
Method Map (null, record [value f, categories [g1, g2…]])
For all category g in [g1, g2...]
Emit (record [g, f], count 1)
Class Reducer
Method Reduce (record [g, f], counts [n1, n2, n3, n4 …])
Emit (record [g, f], null)

Pseudo code for query mapping optimization

Class Mapper
Method Map (record [f, g], null)
Emit (value g, count 1)
Class Reducer
Method Reduce (value g, counts [n1, n2...])
Emit (value g, update ([n1, n2...]))

The algorithm finds the frequency of an item that is of importance and is associated to a similarity matrix.

IV. RESULTS AND DISCUSSION

The following results have been established by the conductance of the proposed work, the graphs are depicted on the formula,

$$\text{Throughput } (Nq)=\sum^{Q}_{Q=0} \textbf{FILESIZE(X)} / \sum^{Q}_{Q=0} \textbf{TIME(X)}$$

The calculation of time taken by a query is calculated. The following graphs illustrate the skill picture based on their repeaters. As well as keeping them on track to increase the data volume, as well as to address the appropriate information to use the benefits of the expenditure, the solutions face challenges that are needed. The two-level query processing technique can be used to improve the efficiency of such problems and to classify the data with the help of large data on the big data.

The test time passed by all scheduled schedules is as follows: Creating a new algorithm for commercial and also non-commercial use solutions to help solve these issues can help in community development. The proposed algorithm can help to improve the placement of data categorization algorithms MapReduce in the delicate Hadoop clusters.

V. CONCLUSION AND LIMITATIONS

This research work have emulated quite surprising results, some of them being the choice of schedulers to schedule jobs, placement of data in similarity matrix, clustering before

scheduling queries and moreover, iterative, mapping and reducing and binding the internal dependencies together to avoid query stalling and execution times. The Calculating query dependencies which analyzed in the parsers. Scheduling the jobs that process such data so that the clusters with high-performance can deal with big data sets and clusters with slow-ends can help in other processing. The iterative mapping and reducing in it proves to be fruitful applicant of processing and finding out meaningful insights about data. The iterative processing can define the input of similarity matrix and hence simplify the processing and could complete jobs in relevantly shorter time spans.

Large data and hadoop clusters must be the number of n structures, the query output partly depends on i/o processing time and the name of the data node is the node server.

REFERENCES

[1] Zhuo Liu, *"Efficient Storage Design and Query Scheduling for Improving Big Data Retrieval and Analytics"*, Dissertation, Auburn University, Alabama 2015.

[2] Zongben Xu, Yong Shi, *"Exploring Big Data Analysis: Fundamental Scientific Problems"*, Springer Ann. Data. Sci., Vol. 2, Issue. 4, pp 363–372, December 2015.

[3] F.G. Tinetti, I. Real, R. Jaramillo, and D. Barry, *"Hadoop Scalability and Performance Testing in Heterogeneous Clusters"*, In the Proceedings of the 2015 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA-2015), Part of WORLDCOMP'15 pp.441-446, 2015.

[4] K. Kamtekar, under the guidance of R. Jain *"Performance Modeling of Big Data"*, Washington University in St. Louis, pp. 1-9, June 2015.

[5] F.H. Liu, Y.R. Liou, H.F. Lo, K.C. Chang and W.T. Lee, *"The Comprehensive Performance Rating for Hadoop Clusters on Cloud Computing Platform"*, International Journal of Information and Electronics Engineering, Vol. 4, No. 6, pp.480-484, November 2014.

[6] T.K. Das, P.M. Kumar, *"BIG Data Analytics: A Framework for Unstructured Data Analysis"*, International Journal of Engineering and Technology (IJET), ISSN: 0975-4024, Vol. 5 No. 1, pp.153-156, Feb-Mar 2013.

[7] F. Novacescu, *"Big Data in High Performance Scientific Computing"*, International Journal of Analele Universităţii "Eftimie Murgu", published by the "Eftimie Murgu" University of Resita, ANUL XX, NR. 1, pp.207-216, 2013, ISSN 1453 - 7397.

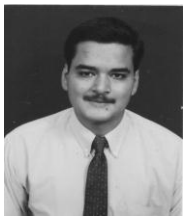
[8] B.T. Rao, N.V. Sridevi, V.K. Reddy, L.S.S. Reddy, *"Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing"*, Global Journal of Computer Science and Technology, Volume XI, Issue VIII, May 2011.

[9] J. Xie, S. Yin, X. Ruan, Z. Ding, Y. Tian, J. Majors, A. Manzanares, and X. Qin, *"Improving MapReduce Performance through Data Placement in Heterogeneous Hadoop Clusters"*, Proceedings of the 19th International Heterogeneity in Computing Workshop, Atlanta, Georgia, pp.1-9, April 2010.

Authors Profile

Mr. Pankaj Dadheech received his M.Tech degree in Computer Science & Engineering from Rajasthan Technical University, Kota and he has received his B.E. in Computer Science & Engineering from University of Rajasthan, Jaipur. He has more than 10 years of experience in teaching. He is currently working as an Associate Professor in the Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan (SKIT), Jaipur, Rajasthan, India. He has presented 25 papers in various National & International conferences. He has 12 publications in various International & National Journals. He is a member of many Professional Organizations like the IEEE Computer Society, CSI, ACM & ISTE. He has also guided M.Tech Research Scholars. His area of interest includes High Performance Computing, Data Mining, Big Data Analytics.

Prof. (Dr.) Dinesh Goyal received his PhD degree in Computer Science & Engineering from Suresh Gyan Vihar University, Jaipur, Rajasthan, India and he has received his M.Tech degree in Computer Science & Engineering from Rajasthan Technical University, Kota, Rajasthan, India. He has more than 16 years of experience in teaching. He is currently working as a Principal, Suresh Gyan Vihar University (SGVU), Jaipur, Rajasthan, India. He is a President, Computer Society of India, Jaipur Chapter since April 2015. He has been Vice President, CSI Jaipur Chapter since 2009 to April 2015. He is a member of the IEEE New Delhi Chapter. He has Submitted Project worth Rs. 15, 00, 000/- under MODROBS to AICTE. To his credit, he has more than 30 publishing in the proceedings of the reputed National & International conferences. He has 84 publications in various International & National Journals. He has also guided M.Tech & Ph.D. Research Scholars. His area of interest includes Information Security, Mathematical Computing, Cloud Computing, Big Data Analytics, High Performance Computing.



Dr. Sumit Srivastava received his PhD degree from University of Rajasthan, Jaipur, Rajasthan, India. He has received his M.Tech from Rajasthan Vidyapeeth, Jaipur and M.C.A. from Birla Institute of Technology, Mesra, Ranchi. He has more than 17 years of experience in teaching. He is currently working as a Professor in the Department of Information & Communication Technology, Manipal University, Jaipur. To his credit, he has more than 35 publishing in the proceedings of the reputed National & International conferences. He has 15 publications in various International & National Journals. He has also guided M.Tech & Ph.D. Research Scholars. His research interest includes Statistical method in Data Mining, Grid & Cluster Computing and Network Security, High Performance Computing, Cloud Computing and Data Mining.

