IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

# Improvement of an Effective Data Emplacement and Redistribution Algorithm among Nodes in Cloud Based Environment

**S. Annapoorani[1*], B. Srinivasan[2]**

[1]Dept. of Computer Science, Gobi Arts & Science College, Bharathiar University, Gobi, India

[*]*Corresponding Author: sannapooranisathy@gmail.com, Tel.: +91-85085-38504*

**Available online at: www.ijcseonline.org**

Accepted: 14/Nov/2018, Published: 30/Nov/2018

***Abstract***— This paper is concerned with the study and analysis of Data Emplacement and Redistribution (DER) in large set of databases called Big Data and proposes a model for improving the efficiency of data processing and storage utilization for dynamic load imbalance among nodes in a heterogeneous cloud environment. With the era of explosive information and data receiving, more and more fields need to deal with massive, large scale of data. A method has been proposed with an improved Data Placement algorithm called Effective Data Emplacement and Redistribution approach (EDER) with computing capacity of each node as a predominant factor that promotes and improves the efficiency in data processing in a short duration time from large set of data. The proposed solution improves the performance of the heterogeneous cluster environment by effectively distributing data based on the performance oriented sampling as the experimental results made with word count applications.

***Keywords***— Cloud Computing, Big Data, HDFS, MapReduce, Data emplacement, MapReduce applications

## I. INTRODUCTION

Cloud computing is emerging as powerful paradigm for dealing with Big Data, which is developed from distributed computing. Big Data is a complex, un-structured or very large size of data. Hadoop is a tool or environment that is used to process Big Data in parallel processing node. HDFS enables Hadoop MapReduce Applications to transfer processing operations towards nodes storing application data to be processed by the operation.

The current HDFS block placement policy cannot evenly and fairly distribute data blocks across the heterogeneous cluster and results in an unbalanced cluster. By particularly, it is hard to utilize the power of each node correctly this may cause performance degradation. From the expected approach it is to develop a method that can dynamically adapt and balanced data stored on each node based on node load status in a heterogeneous cloud environment.

Rest of the paper is organized as follows, Section I contains the introduction of introduction of cloud computing, Section II contain the related works of literature survey, Section III contain the some measures of motivation of research, Section IV contain the architecture and design of proposed system, section V describes results and discussion, Section VI contain the recommendation of conclusion.

## II. RELATED WORK

The data reorganization and distribution algorithms implemented in HDFS can be used to solve skew problem due to dynamic insertions and deletions [1]. The heterogeneity – aware data distribution and rebalance method is used for data splitting phase which priors to job execution and to solve the problem during task execution that some straggler task occurs [2].

The Data placement algorithm is based for map tasks of data locality to allocate data blocks, and the data required for performing a task is often non local [5]. A data placement that focuses on the hardware generation of the Datanodes during placing blocks. Their observation leads to strategy that placing more data blocks into newer hardware generation nodes, the performance of an application is expected to increase [6]. A strategy tends to select nodes in the high network load group with the larger disk space, which realizes load balance as far as possible because the focuses is on load balancing by selecting optimal node to place replica by the balancer procedure [4].

In the consideration of processing speed of the data nodes in the cluster by assigning the data blocks into its speed analyser and Data Distribution Components for handling straggler node [8]. The heterogeneity – based data placement based on the computing capacities and file formats of data

© 2018, IJCSE All Rights Reserved 

for allocating data blocks in the cluster [7]. The data placement policy tries to explore the consequence of migrating the data block to a faster node in the heterogeneous cluster. The CRBalancer is responsible for migrating the data from one to another [9].

The information from the reviews made over the existing methods shows that there exists hard evidence on the lack of performance evaluation of data placement in practice showing. There is still a need for the method that provides better performance over the data processing methods.

## III. MOTIVATION

On determining the data locality with data imbalance over the task execution time and response time in the data processing systems, the earlier research have been concentrated as much on the reduction of processing time along with the data retrieval. Although it is realistic need to reduce the time factor for achieving better performance, to place and distribute the data based on the computing capacity and storage utilization of each node is also the most indispensible in making timely decision in various aspects of data processing methods. There is a need to improve the performance of cluster nodes in the heterogeneous cloud environment. In this regard, when the amount of transferred data due to load sharing is very large, the overhead of moving unprocessed data from slow nodes to fast nodes becomes a critical issue which affects the Big Data Cluster performance. This research aims to minimize the data movement between slow and fast nodes. This goal can be achieved by an effective data emplacement scheme that distribute and store data across multiple heterogeneous nodes based on their computing capacities and storage utilization.

## IV. ARCHITECTURE AND DESIGN OF THE PROPOSED SYSTEM

The Effective Data Emplacement and Redistribution approach has been proposed to obtain the efficiency in the dynamic load stability and data processing from the large set of databases which has follows two distinct phases. In the first phase, name node allocates data blocks based on each node computing capacity ratios in the Ratio table. Therefore, the computing capacity adopts the average time required to complete one task. To measure the heterogeneity of those computing resources, which defines the storage capacity and propose assessment method based on historical job execution log for a given task and computing node.

In the second phase, name node calculates each node appropriate data block numbers which is more compatible with node load status based on the storage utilization parameters of each node. Finally, the data should be processed based on the benchmark applications to analysis the performance of the cluster in the cloud environment. Thus proposed system is designed with the developed pseudo code and algorithm in the form of three phases as initial data emplacement, Redistribution for dynamic data imbalance and data processing to improve better performance of overall system with a minimum duration of time. Figure 1 shows that the dynamic data emplacement algorithm among nodes.

```
Step 1:  When a data is written into HDFS
Step 2:  Input ←Data and JoType
Step 3:  Set result=0;
Step 4:  for each record in the Ratio Table do
Step 5:      if compare jobType with record are the same
             then
Step 6:      result=1
Step 7:          if compare data volume of the record in the
                 antiquity table are same then
Step 8:              Allocate block number data blocks to
                     the node
                 Goto step 15
Step 9:  if result=0 then
Step 10:     Add JobType with the computing capacity ratio
             to ratio table

Step 11:     Calculate total number of nodes
Step 12:     Calculate storage weight of each data node
Step 13:     for each data node in the cluster do
Step 14:         NodeCapacity=1
Step 15:             Calculate block number for each
                     node
```

**Figure 1 : Dynamic Data Emplacement Algorithm**

## V. RESULTS AND DISCUSSION

The proposed system has been developed for improving the performance of clusters in the data processing with the large set of databases with a less amount of time using Hadoop as a software tool in the Cloudera package. Word Count is a type of benchmark job run to evaluate the performance of the proposed algorithm in the heterogeneous cluster environment.

First step is to create Hadoop cluster and fine every node processing of the cluster. MapReduce applications have been executed on the system extended with the proposed Data Distribution Technique. The behaviour of the word count MapReduce applications is analyzed for the data redistribution using different data sizes. Fig 2 presents the execution time of each node taken by the word count application for a data size. Fig 3 shows that comparison between the execution time of the whole cluster in each round for running job.

© 2018, IJCSE All Rights Reserved 

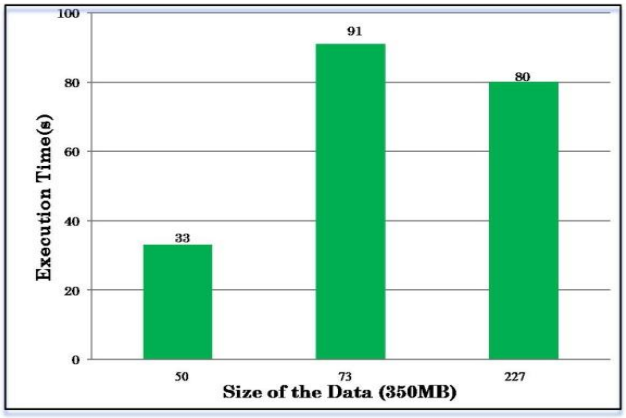


Fig 2 : Execution time of each node

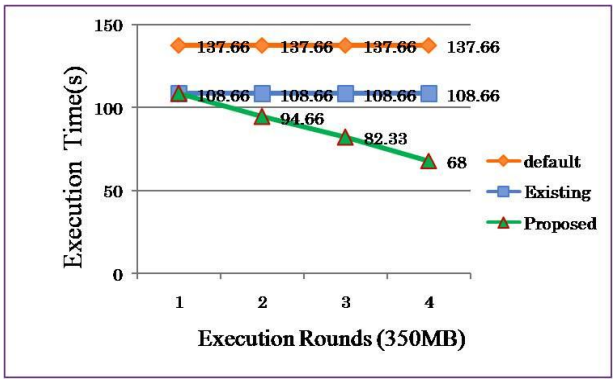


Fig 3 : Comparison between the Execution time

## VI. CONCLUSION

In this paper, it is proposed for improving the better performance in processing the large set of databases with the recent issues in the research. The proposed system gets improves in the execution and response time of the data retrieval with efficiency and also reduces dynamic load stability of each nodes in the cluster. The adaptability can be enhanced by dynamically distribute data on the nodes by using computing capacity and storage utilization with the various components to improve the overall performance of a system with a less amount of time.

## REFERENCES

[1] Jiong Xie, Shu Yin, Xiaojun Ruan, Zhiyang Ding, Yun Tian, " Improving MapReduce Performance through Data Placement in Heterogeneous Hadoop Clusters", 19th International Heterogeneity in Computing Workshop, Atlanta, Georgia, April 2010.

[2] Yuanquan Fan, Weiguo Wu, Haijun Cao, Huo Zhu, Xu Zhao, Wei Wei, "A heterogeneity-aware data distribution and rebalance method in Hadoop cluster", Seventh ChinaGrid Annual Conference, 2012.

[3] Mahesh Maurya, Sunita Mahajan "Performance analysis of MapReduce Programs on Hadoop Cluster" IEEE World Congress on Information and Communication technologies,2012.

[4] Wentao Zhao, Lingjun Meng, Jiangfeng Sun, Yang Ding, "An Improved Data Placement Strategy in a Heterogeneous Hadoop Cluster", The Open Cybernetics & Systemics Journal, 2014.

[5] Chia-Wei Lee, Kuang-Yu Hsieh, Sun-Yuan Hsieh, Hung-Chang Hsiao , "A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments", Big Data Research, 2014.

[6] Dipayan Dev, Ripon Patgiri "Performance Evaluation of HDFS in Big Data Management", International Comference on High Performance Computing and Applications (ICHPCA), 2014.

[7] Suhas V. Ambade, Priya R. Deshpande, "Heterogenity-based files placement in Big Data Cluster", International Conference on Computational Intelligence and Communication Networks, 2015.

[8] Vrushali Ubarhande, "Novel Data-Distribution Technique for Hadoop in Heterogeneous Cloud Environments", IEEE Transactions 2015.

[9] Ch. Bhaskar VishnuVardhan and Pallav Kumar Baruah, "Improving the Performance of Heterogeneous Hadoop Cluster", Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2016.

[10] Anton Spivak and Denis Nasonov "Data Preloading and Data Placement for MapReduce Performance Inproving" Procedia Computer Science 101, 2016.

[11] Ramchandani Hema Megharajbhai, Viral Parmar, "Heterogeneity based Fairly Data Distribution in Cluster Environment", International Journal of Advance Engineering and Research Development, 2018.

[12] S. Annapoorani, Dr. B. Srinivasan, "Initial Dynamic Data Allocation for Heterogeneous hadoop clusters" International Journal of Scientific Research in Computer Science Applications and Management Studies, Volume 7, Issue 3, 2018.

[13] S. Annapoorani, Dr. B. Srinivasan, "Improving performance of data in Hadoop clusters using dynamic data replication" International Journal of Engineering sciences & Research Technology, Feb, 2018.

## Authors Profile

*S. Annapoorani* pursed Bachelor of Science from Gobi Arts & Science College, Bharathiyar University, Coimbatore in 2011 and Master of Computer Application from Gobi Arts & Science College, Bharathiyar University, Coimbatore in 2014. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science, Gobi Arts & Science College since 2014. She has published more than 10 research papers in reputed national and international journals and Conferences including IEEE and it's also available online. Her main research work focuses on Cloud Security and Privacy, Big Data Analytics. She has 4 years of teaching experience and 2 years of Research Experience.

*Dr. B. Srinivasan* pursed Master of Computer Application in Gobi Arts & Science College, Bharathiyar University, Coimbatore and Ph.D in Computer Science at Vinayaka Missions University, Salem. He is working as Associate Professor in Department of Computer Science, Gobi Arts & Science College. He has published more than 70 research papers in reputed national and international journals and conferences including IEEE and it's also available online. His main research work focuses on Network Security and Automata theory. He has 25 years of teaching experience and 18 years of Research Experience.

© 2018, IJCSE All Rights Reserved 