Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.S1, 2019, pp.87-88

DOI: https://doi.org/10.51983/ajcst-2019.8.S1.1944

Implementation of Effective Data Emplacement Algorithm in Heterogeneous Cloud Environment

S. Annapoorani[1] and B. Srinivasan[2]
[1]Assistant Professor, [2]Associate Professor
[1&2]Department of Computer Science, Gobi Arts & Science College
E-Mail: sannapooranisathy@gmail.com, srinivasan_gasc@gmail.com
(Received 17 December 2018; Revised 20 January 2019; Accepted 12 February 2019; Available online 15 February 2019)

Abstract **- This paper is concerned with the study and implementation of effective Data Emplacement Algorithm 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 Algorithm 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 adaptability of the proposed model can be obtained by minimizing the time with processing efficiency through the computing capacity of each node in the cluster. 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: **Big Data, Cloud computing, Hadoop, Data Emplacement, Cluster**

I. INTRODUCTION

Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Distributed computing is a computing concept that, in its most general sense, refers to multiple computer systems working on a single problem. In distributed computing, a single problem is divided into many parts, and each part is solved by different computers. As long as the computers are networked, they can communicate with each other to solve the problem [1]. If done properly, the computers perform like a single entity. Cloud Computing, which simply involves hosted services made available to users from a remote location, may be considered a type of distributed computing. Emerging distributed file systems in production systems strongly depend on a central node for chunk reallocation. This dependence is clearly inadequate in a large-scale, failure-prone environment because the central load balancer is put under considerable workload that is linearly scaled with the system size, and may thus become the performance bottleneck and the single point of failure.

A distributed system consists of multiple autonomous computers, each having its own private memory, communicating through a computer network. Information exchange in a distributed system is accomplished through message passing. The cloud applies parallel or distributed computing, or both. Clouds can be built with physical or virtualized resources over large data centers that are centralized or distributed. Cloud computing is the emerging trend in the field of distributed computing.

II. CLOUD COMPUTING

Cloud computing is a robust technology, which facilitate to resolve many parallel distributed computing issues in the modern Big Data environment. Cloud Computing is a recent trend in IT that moves computing and data away from desktop and portable PCs into large data centers [3]. It refers to applications delivered as services over the Internet as well as to the actual cloud infrastructure- namely, the hardware and systems software in data centers that provide these services.

In a cloud computing environment, the traditional role of service provider is divided into two: the infrastructure providers who manage cloud platforms and lease resources according to a usage-based pricing model, and service providers, who rent resources from one or many infrastructure providers to serve the end users. Date becomes a great concern when it is outsourced to cloud. Hence, the most active domain of research in cloud computing is concentrating on the data placement with the security and privacy of cloud data.

III. DEFAULT DATA PLACEMENT ALGORITHM

By default, the input data of a job was divided into equal size of data blocks that are available on the data nodes. This data placement policy is applicable only in homogeneous cluster but it is not suitable in heterogeneous clusters because of in cluster each node computing capacity is different. [4]

In order to measure the computing and storage capability, it is to propose an assessment method based on historical task execution time for a given set of task and computing node. By default, data placement framework does not consider the node load state in the distribution of input data blocks; this may cause insufficient overhead and less performance in the

cluster. It assumes that each node computing capacity and storage capacity are equal in the homogeneous cluster environment. Each and every node is assigned by the same workload in all the clusters. But in a heterogeneous environment, each nodes having different computing capacity, this may leads a load imbalance [2]. So that the result would be moving a high amount of data from one another is unnecessary.

IV. DESIGN OF PROPOSED ALGORITHM

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.

V. EXPERIMENTAL RESULTS

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 1 presents the execution time of each node taken by the word count application for a data size. Fig 2 shows that comparison between the execution time of the whole cluster in each round for running job.

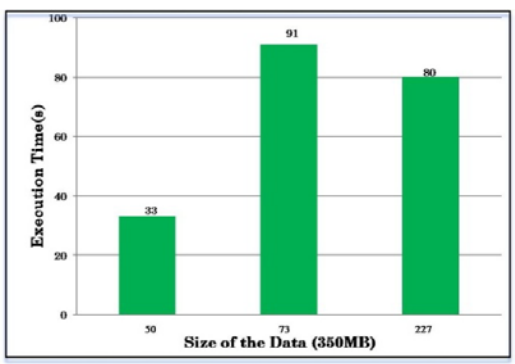

Fig. 1 Execution time of each node

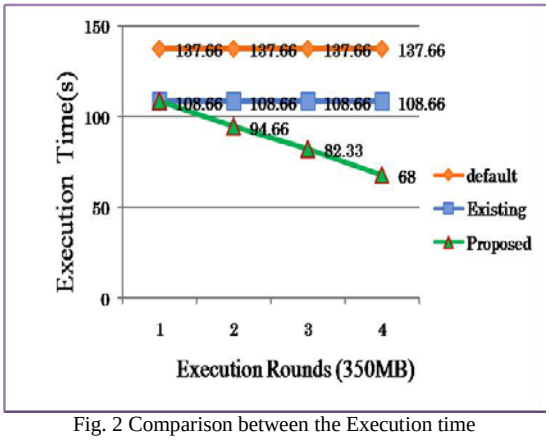

Fig. 2 Comparison between the Execution time

VI. CONCLUSION

An Effective Data Emplacement Algorithm has been proposed for improving the better performance in processing the large set of databases with the recent issues in the research. The proposed system improves in the execution and response time of the data retrieval with efficiency and also reduces dynamic load stability of each node in the cluster. The adaptability can be enhanced by dynamically distribute data on the nodes by using computing capacity and storage utilization with various components to improve the overall performance of a system with highly secure in a less amount of time.

REFERENCES

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

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

[3] Mahesh Maurya and 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 and 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 and Hung-Chang Hsiao, "A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments", *Big Data Research*, 2014.

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

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

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