ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Performance Analysis of Data Placement Approach in Heterogeneous Hadoop Clusters

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

Research Scholar, Department of Computer Science, Gobi Arts & Science College, Gobi, Tamilnadu, India[1]

Associate Professor, Department of Computer Science, Gobi Arts & Science College, Gobi, Tamilnadu, India[2]

ABSTRACT: In this work, we model an Effective Data Emplacement and Redistribution approach for large scale streaming data application in the heterogeneous cloud environment. It acts as efficient to provide scalability to the data in the shortest duration. It also has proven to be an effective platform to process a large set of unstructured data. The approach employs the distributed computing at scale usually involving hundreds to thousands of machines to facilitate the large scale data processing. It offers a more cost-effective model to implement data processing and data placement. The difference in processing capabilities on MapReduce nodes results in load imbalance which requires separate mechanism to done to make task scheduling and load balancing as heterogeneity aware model. MapReduce model uses different constraints and configuration based on the resources to balanced the data towards increasing the efficiency of the resource clusters in the cloud environment

KEYWORDS: Data placement, distribution, heterogeneous clusters, scheduling

I. INTRODUCTION

Distributed computing is a model in which elements of a 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 [1].
Cloud Computing, which simply involves hosted services made available to users from a remote location, may be considered a type of distributed computing. In reality, examining complicate problems includes division of given problem into sub tasks and which can be solved by one or more computing nodes that communicate with one another by message passing. The current technology towards Big Data Analytics has lead to such compute well intensive tasks.
The default data block placement policy distributes the blocks across the multiple cluster nodes. In this, the algorithms are used to create server hotspot that make unnecessary node loads into the cluster nodes and improves only less performance of the cluster. However, the default block placement policy does not view the various hardware configurations that affect results in under utilization of resources.[2] To achieve this, block placement mechanism is to distribute the large number of input data blocks to each nodes based on the computing capacity of each node.
The workflow engines need to improvise substitutes, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability. A uniform data management system for scientific workflows running across geographically distributed sites, aiming to reap economic benefits from this geo-diversity. It is environment-aware, as it monitors and models the global cloud infrastructure, offering high and predictable data handling performance for transfer cost and time, within and across sites.

Paper is organized as follows. Section II describes the related work in data placement schemes. The flow diagram represents the step of the proposed algorithm discussed in Section III. Section IV presents experimental results showing results based on the various parameters tested. Finally, Section V presents conclusion.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

II. RELATED WORK

With the increasing utilization and popularity of the cloud infrastructure, more and more data are moved to the cloud storage systems. This makes the availability of cloud storage services critically important, particularly given the fact that outages of cloud storage services have indeed happened from time to time. Thus, solely depending on a single cloud storage provider for storage services can risk violating the service-level agreement (SLA) due to the weakening of service availability. This has led to the notion of Cloud-of-Clouds, where data redundancy is introduced to distribute data among multiple independent cloud storage providers, to address the problem. The key in the effectiveness of the Cloud-of-Clouds approaches lies in how the data redundancy is incorporated and distributed among the clouds[3].

However, the existing sample based load balancing approaches utilize either replication or erasure codes to redundantly distribute data across multiple clouds, thus incurring either high space or high performance overheads. In this work, the author proposed a hybrid redundant data distribution approach, called HyRD has been considered to improve the cloud storage availability in Cloud-of-Clouds by exploiting the workload characteristics and the diversity of cloud providers. [4]. In HyRD, large files are distributed in multiple cost-efficient cloud storage providers with erasure-coded data redundancy while small files and file system metadata are replicated on multiple high-performance cloud storage providers.

III. EFFECTIVE DATA EMPLACEMENT AND REDISTRIBUTION

The configurations on heterogeneous nodes inevitably lead to sub-optimal performance. The data redistribution configuration optimizes the storage and processing capabilities on the heterogeneous resource clusters. A large number of performance critical parameter can have complex interplays on data placement and data classification. However, it is difficult to construct models to connect parameter setting for optimal solution on the MapReduce performance towards data processing. The figure 1 depicts the architecture of effective data emplacement and redistribution.
Different configurations are needed for different execution phases on the dynamic imbalance data [11]. The data classification configurations should also be changed in response to the changes in the heterogeneous cluster resource performance.

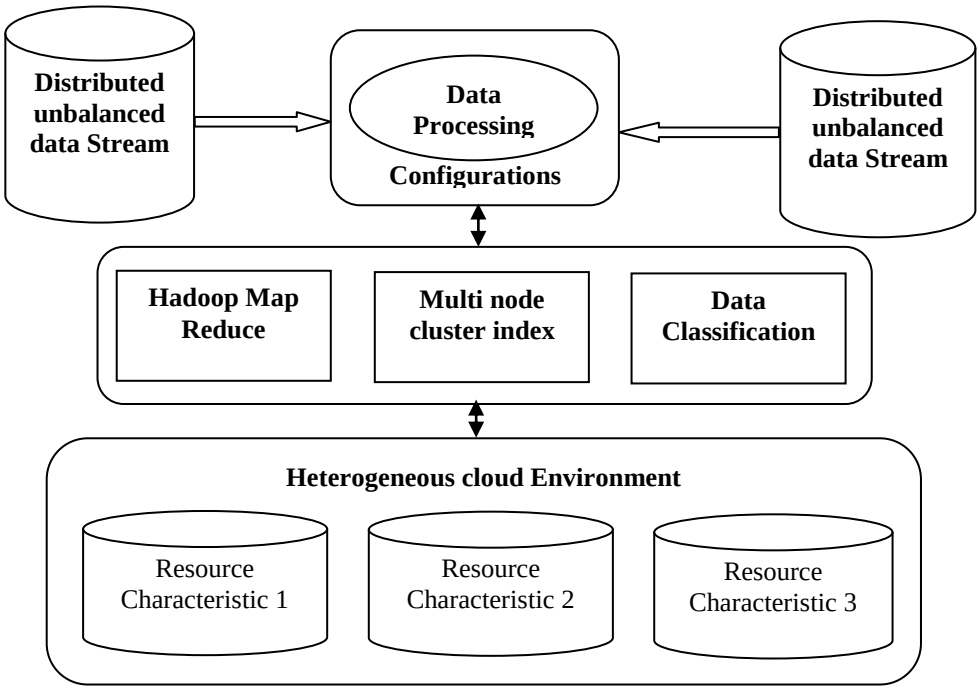

Figure 1: Architecture of Effective data Emplacement and Redistribution

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Task-level parameters control the behaviour of task execution, which is critical to the Hadoop. Task-level parameters which have significant performance impact the resource mapping. The Task distribution to Map is given as

$$TD = \sum_{n=0}^{x} D^k K$$

Where D^k is the document instance or data instance on the cluster

K is the cluster

$$TR = T + K \sum_{n=1}^{\infty} \left(\text{rand} \frac{T}{l} + \log \frac{T}{l} \right)$$

Task on reduce function is computed randomly as well logarithmically at different task configurations. Scheduling on the heterogeneous cluster is to minimize job transfer time, task execution time and VM utilization based on the algorithm 3.1.

Algorithm 3.1:Dynamic imbalance data

Compute the fitness for each job
If (job== fitness value)
Then
Determine slot availability
Select two configuration candidates as parents;
Do
Crossover and Mutation operations;
Use the obtained new generation Cnew to assign task
end if
Until
The running job is completed.

IV. RESULTS & DISCUSSION

4.1 Performance Evaluation Metric

In this section, various performances metric employed to the compute the performance of the proposed architecture towards deadline requirement, secure transaction processing and dynamic task execution has been derived and its definition is as follows

- **Job Transfer Time**
  We evaluate the performance changes with respect to different workflow sizes when the resources for each workflow are increased from the lower bound to the upper bound of the task type (particle)

  Job Transfer time = Task execution time in the Resource -Task Submission Time to the Scheduler

- **Task Execution Time**
  The Task execution time computes difference of time against the task completion in the resource to task submitted to the resource

  Task Execution time = Task Completion time in the resource – Task Submitted time to the resource

- **VM Utilization**
  VM utilization is defined as aggregation of memory, bandwidth and Ram utilization for the processing of the entire task on the virtual machine slot.

  Vm= $\sum_{i=1}^{\infty}(a_n + b_n)$
  VM utilization = (ram Utilization * memory Utilization * bandwidth Utilization) / 3

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

4.2 Performance Evaluation

In this proposed approach, data processing model has been applied with datasets of 50GB is used for calculating the data retrieval cost of a job. Table 1 provides the performance values of the Effective Data Emplacement and Redistribution approach.

Technique	Job Transfer Time in ms	Task Execution time in ms	VM utilization in mb
Sampling based Load balancing –Existing	25.32ms	45.25ms	12.56
Effective Data Emplacement and Redistribution approach – Proposed Methodology 1	21.32ms	28.98ms	9.56

Table 1: Performance values of the Effective Data Emplacement and Redistribution approach

This model estimates the available slots for serving the job to meet job deadline. Further constraints utilize the slot heterogeneous performance to assign each task to the best suitable slot. Storage is management in the cluster to serve better on the task evolutions. Data placement constraint is used to determine the hidden features of the job which hide the unobserved features of the dataset which can be even eliminated as non valuable feature on estimating the hardware properties of the computing platform. It can be considered as resource sampling technique for the effective task redistribution. The figure 2 represents the job transfer time performance of the proposed approach.

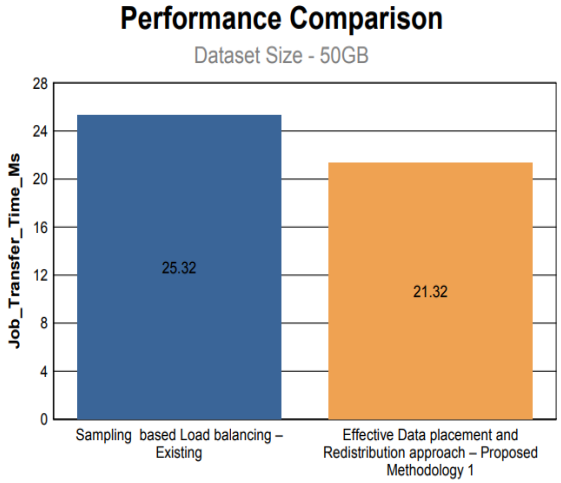

Figure 2: Performance Evaluation of the EDER approach against SLB approach in terms of Job Transfer time

The constraint is needed to calculate the number of slots required for each job on the particular VM class. The figure 3 represents the performance of the proposed approach on the VM utilization.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

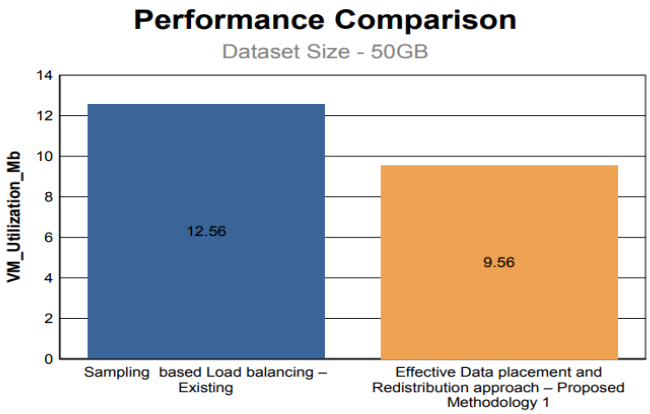

Figure 3: Performance Evaluation of the EDER approach against SLB approach in terms of VM utilization

The VM feature transition probabilities from familiar to unobserved VM may be carried out using markov property analysis against slot computation time. The figure 4 provides the performance of the proposed model in terms of execution time.

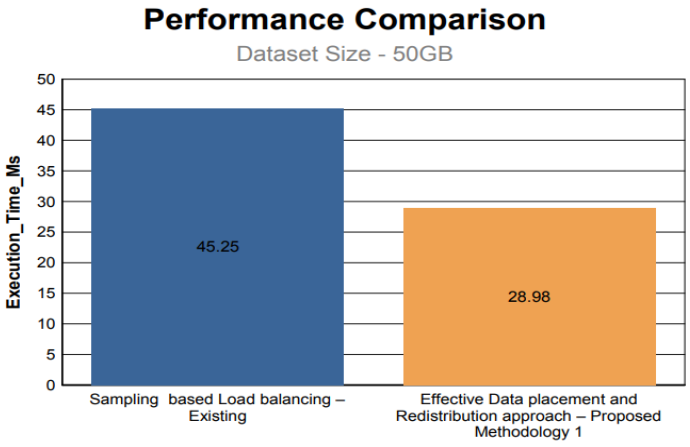

Figure 6.3: Performance Evaluation of the EDER approach against SLB approach in terms of Execution time

V. CONCLUSION

We have implemented an effective data emplacement and redistribution approach for dynamic load stability in heterogeneous resource clusters. The performance analysis of the proposed framework is discussed against various datasizes and different performance metrics. It indicates proposed mechanism provides the better results on processing the large set of imbalance data. We have applied our algorithm on the datasets and found that it successfully reduce the minimum cost of the each parameter.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

REFERENCES

[1] 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.
[2] M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica, "Improving mapreduce performance in heterogeneous environments," in Proc. of USENIX Symposium on Operating System Design and Implementation (OSDI), 2008.
[3] 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.
[4] N. Tiwari, S. Sarkar, U. Bellur, and M. Indrawan, "Classification Framework of MapReduce Scheduling Algorithms," ACM Computing Surveys, vol. 47, no. 3, pp. 49:1–49:38, Apr. 2015.
[5] Mahesh Maurya, Sunita Mahajan "Performance analysis of MapReduce Programs on Hadoop Cluster" IEEE World Congress on Information and Communication technologies,2012.
[6] Wentao Zhao, Lingjun Meng, Jiangfeng Sun, Yang Ding, "An Improved Data Placement Strategy in a Heterogeneous Hadoop Cluster", The Open Cybernetics & Systemics Journal, 2014.
[7] Dipayan Dev, Ripon Patgiri "Performance Evaluation of HDFS in Big Data Management", International Comference on High Performance Computing and Applications (ICHPCA), 2014.
[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.