LOAD BALANCING AND DUPLICATION AVOIDANCE USING A HEURISTIC ALGORITHM

Dhanalakshmi. G 1, Chandhra Pradhiksha. N2, Kunapriya. P3, Mansha Devi. J4

[1]Associate Professor, [2],[3],[4]UG SCHOLAR

Department of Information Technology, Panimalar Institute of Technology ,Chennai

---***---

Abstract—In this paper we propose a solution to select Virtual Machine task scheduling with load balancing clustering using optimization techniques has been implemented in Twitter api formation of dataset. Our solution belongs to showing the correctness and the efficiency of the proposed reputation management system using analytical and experimental analysis. To provide datacenter which forms the resources with computational cost to the virtual network for low trust on the temporary resources with their computational resources protects the users to reduce the cost computational resources are shared. Our method Robust reputation management mechanism that encourages the CPs(cloud provider) federated cloud to differentiate between good and malicious users and assign resources in such a way that do not share resources.

I. INTRODUCTION

Big data is a field that treats ways to analyze, systematically extract information from, or in any case manage informational collections that are excessively huge or complex to be managed by usual information handling application programming. Big data analysis challenges incorporate capturing data, data storage, information investigation, search, sharing, allocation, perception, questioning, refreshing, data protection, and information source. Information is useless if data that can be utilized for further reasoning can't be gathered from it. Cluster analysis, in light of certain models, shares information into significant, practicality or the two categories (clusters) based on shared common characteristics. Load balancing of computational tasks and data plays a critical role in big data systems. Load balancing attempts to optimize the overall computation or system. This mainly focuses IAAS paradigm of cloud computing environment with computational data center for resources which is access in the Virtual Machine in the form of cloudlets. We show the correctness and the efficiency of the proposed reputation management system using analytical and experimental analysis. We analyse the challenging issues in the data-driven model and also in the Big data revolution. It is the realization of greater business intelligence by storing and analyzing data that was previously ignored due to the limitations of traditional data management Technologies

II. RELATED WORKS

A. Big Data Processing Workflows Oriented Real-Time Scheduling Algorithm using Task-Duplication in Geo-Distributed Clouds (2020)

Scheduling big data processing work processes includes both huge scope task and transmission of monstrous middle information among task, accordingly improving their finish time and expense turns into a difficult issue. The real-time scheduling algorithm using task-duplication, RTSATD, with the end goal that limiting both the finish time and financial expense of preparing huge information work processes in clouds.

B. Study on Load Balancing of Intermittent Energy Big Data Cloud Platform

This paper proposes a system that the cloud stage is incorporated with the discontinuous fuel sources information and the heap adjusting of multifaceted prescient cloud stage. Initially, conveying the general cycle of the discontinuous energy information preparing on another information handling stage, and afterward running the multifaceted prescient cloud stage load adjusting on the preparing stage.

C. A Survey of Data Partitioning and Sampling Methods to Support Big Data Analysis

Computer clusters with the common nothing design are the significant figuring stages for big data processing and analysis. In cluster computing, data partitioning and sampling are two fundamental strategies to accelerate the calculation of large information and increment versatility. The basic methods of data partitioning are then discussed including three classical horizontal partitioning schemes: range, hash, and random partitioning. The traditional techniques for data sampling are then researched, including straightforward irregular testing, separated examining, and supply testing.

D. Load balancing in cloud computing using dynamic load management algorithm

The Load balancing issue of distributed computing is a significant issue and basic part for sufficient tasks in distributed computing framework and it can likewise forestall the fast advancement of distributed computing. Numerous clients from one side of the planet to the other are requesting the different administrations at fast rate in the new time. Although different load balancing calculations have been planned that are proficient in demand designation by the choice of right virtual machines.

III. EXISTING SYSTEM

The traditional exact methods such as divide and conquer, branch and bound, dynamic programming, and Linear Programming gives the global optimum, but it is a lot of time consuming for solving typical real-world problems. Optimal task scheduling to minimize the sum of task running and communication costs using the branch-and-bound technique and by using simulation methods, can calculate computational complexity of this method. Conventional methods can apply in optimization are deterministic, fast, and give exact answers but often tends to get stuck on local optima. n this situation heuristic approach Plays significant role. Heuristic means serving to find out. It gives most possible (if not optimal) solutions, which are good enough from practical point of view

IV. PROPOSED SYSTEM

Our proposed system, suggest that the Energy consumption with cost budget mechanism that encourages the PM (Physical Machines) in a heterogeneous cloud to differentiate between energy scheduling and energy efficient in lower bound users which assign resources in such a way that they do not share resources. A robust reputation management mechanism that encourages the CPs in a federated cloud to differentiate between good and malicious users and assign resources in such a way that they do not share resources. Service deployment requests from customers is place to the service portal, which forward the requests to the request management and processing component to validate the requests with the help of SLA.

Data Preprocessing:
Preprocessing is a necessary procedure to improve the quality of raw data, such as filling hole, noise removal etc. With the appropriate signal preprocessing procedure, the undesired information is eliminated from the raw information. Hence, it has few effects on the quality of the feature extraction, leading to an improvement in the identification, accuracy rate.

Twitter API:
The Twitter API allows high-throughput near real-time access to various subsets of public Twitter data.We download trending topics and definitions from Twitter on whichever topic is trending. All the tweets containing a trending topic constitutes into a complete document.

Clustering:
Clustering is a method of gathering items or documents based on some similar characteristics among them. It performs categorization of data items exclusively based on similarity among them.We use clustering instead of classification here because it is hard to find data set for new topics if classification is used.

V. METHODOLOGY

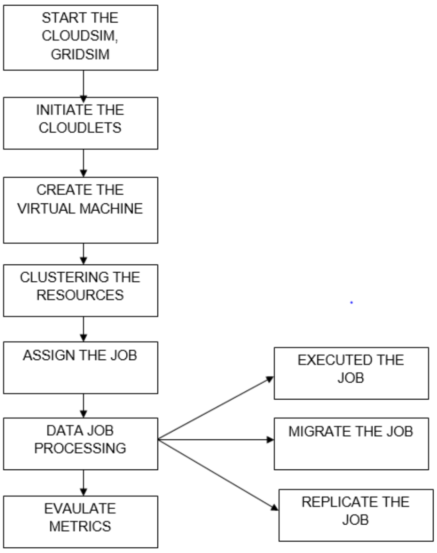
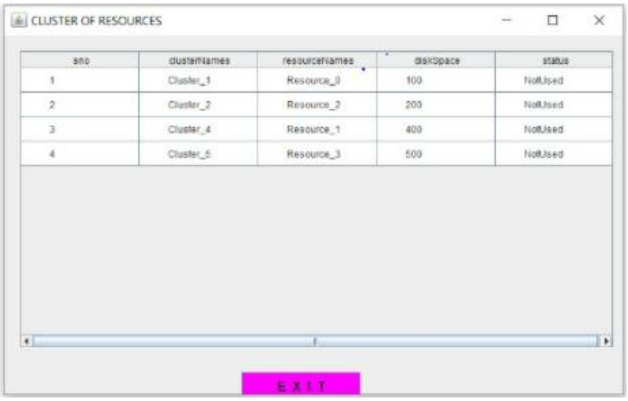


Fig.1 Block Diagram

Step1: We are collecting data
Step2: Starting the CloudSim and Gridsm which is an opensource cloud simulator platform which can process large datasets, Gridsm is a software platform through which we are going to test our new algorithm respectively
Step3:Cloudlets are initiated which is a small center provide quick cloud computing services to our device
Step4:Virtual Machine is created.
Step5:We are Clustering the resources to increase the performance and availability.
Step6:After clustering the jobs are assigned to clusters.
Step7:Data job processing works with databases which we went through so far and start to execute the job.
Execution, Migration and Replication of jobs are done.
Step8:At the end in Evaluate metrics the algorithm is tested to get the desired output.

VI. RESULT

The Data are collected by which resources and disk space are allocated to retrieve the resource successfully. The jobs are allocated without data duplication. The Force directed resource assignment (FRA) Heuristic algorithm with existing algorithm executes the execution time and accuracy of results. Hence it deals with throughput and process value along with process and execution time. The online data prediction for resource normal process is elongated using clusters under process, throughput and execution time.



sno	clusterNames	resourceNames	diskSpace	status
1	Cluster_1	Resource_0	100	NotUsed
2	Cluster_2	Resource_2	200	NotUsed
3	Cluster_4	Resource_1	400	NotUsed
4	Cluster_5	Resource_3	500	NotUsed

Fig.2 Clustered Resources

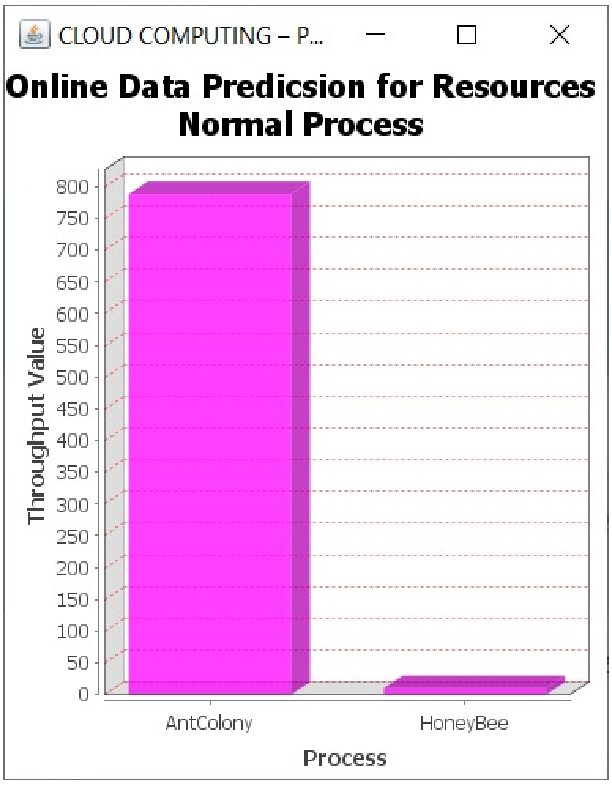


Fig.3 Throughput Value

Task manager is responsible for task status management (start, stop, cancel…), determining the scheduling sequence and resource assignment for the requests and allocating suitable resources to each job under the help of the scheduling algorithm. Resource state component plays the role of managing the available resources, monitoring the performances of resources, dynamic optimization of scheduling strategy and error notification. In data access optimization cluster name, resources name, disk space and status are uploaded. Followed by job allocation for cloudlets clustering.

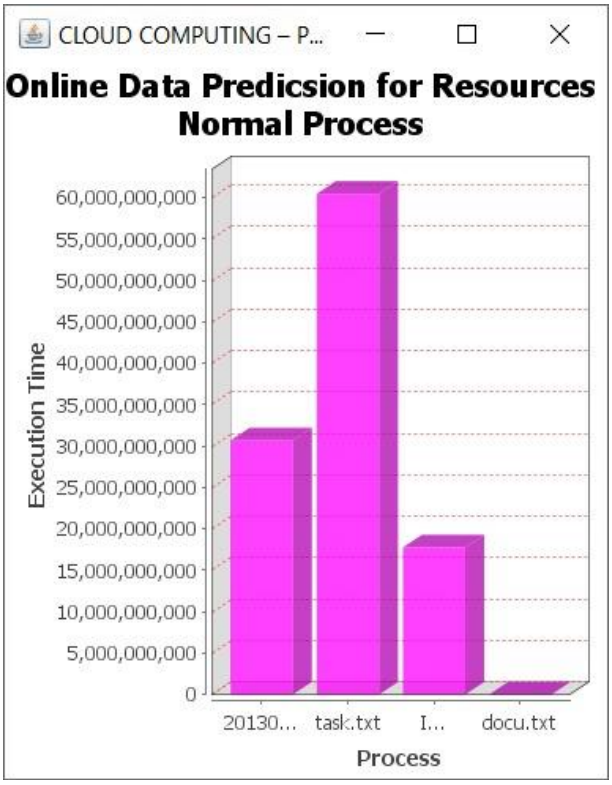


Fig.4 Execution Time

VII. CONCLUSION

We compare the force-directed resource assignment (FRA) heuristic algorithm with existing algorithms on the bases of execution time and accuracy of results. The system is partitioned into a number of clusters and each cluster has a load balancing technique. Therefore, it includes the computing, storage and communication requirements for computing, arrival law and concurrent conditions, security and privacy requirements, QoS of the service and so on.

VIII.REFERENCES

[1] Huangke Chen,Jinming Wen, WitoldPedrycz, and Guohua Wu, "Big Data Processing Workflows Oriented Real-Time Scheduling Algorithm using Task-Duplication in Geo-Distributed Clouds",IEEE Transactions on Big Data, Vol. 6, No. 1, January-March 2020.

[2] Tao Lin, Pengfei Zhao, Jing Zhao,Kang Du, "Study on Load Balancing of Intermittent Energy Big Data Cloud Platform", DOI 10.1109/ICITBS.2018.00101.

[3] Mohammad Sultan Mahmud, Joshua Zhexue

Huang, Salman Salloum, Tamer Z. Emara, and KuanishbaySadatdiynov, "A Survey of Data Partitioning and Sampling Methods to Support Big Data Analysis", Big Data Mining and Analytics, Vol. 3, No. 2, June 2020

[4] F. Messina, G.Pappalardo,D.Rosaci,C.Santoro, and G. M. L. Sarn. A Trust Model for Competitive Cloud Federations. In Complex, Intelligent and Software Intensive Systems (CISIS), 2018 Eighth International Conference On, Pages:469–474, July 2018

[5] LI and J. DU. Adaptive and Attribute-Based TrustModel for Service Level Agreement Guarentee In Cloud Computing. IET INFORMATION SECURITY, 7(1):39–50, MARCH 2019

[6] Z.-J. ZHA, L. YANG, T. MEI, M. WANG, AND Z. WANG. Visual Query Suggestion in Cloud Computing for Task Scheduling Mechanism. 2020.

[7] HongxunYao;WeiLiu;sun; QIitian.Task-dependent Visual-Codebook Compression.2020.

[8] Rajani Sharma and Rajendar Kumar Trivedi Literature review: CLOUD COMPUTING–SECURITY ISSUES, SOLUTION AND TECHNOLOGIES. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH ISSN, PAGES 2319–6890, 2013.

[9] ReenaPanwar,BhawnaMallick"Load balancing in cloud computing using dynamic load management algorithm",2015 International Conference on Green Computing and Internet of Things (ICGCIoT).

[10] D. Wunsch,Rui Xu,"Survey of clustering algorithms",IEEE Transactions on Neural Networks2005

[11] Jyoti Malhotra, JagdishBakal,"A survey and comparative study of data deduplication techniques", 2015 International Conference on Pervasive Computing (ICPC)

[12] Kiran A Jadhav,Mohammed MoinMulla,D. G Narayan,"An Efficient Load Balancing Mechanism in Software Defined Networks",2020 12th International Conference on Computational Intelligence and Communication Networks