Cloud Data Partitioning For Distributed Load Balancing With Map Reduce

Nutan. N
PG student
Dept of CSE,CIT GubbiTumkur

Girish. L
Assistant professor
Dept of CSE, CIT GubbiTumkur

***Abstract*-Cloud computing is an internet based computing. The computing concept has improvedthe usage of a network in whichcapacity of one node can be utilized by other node. Cloud will provide the services on demand tothe distributive resources such as database,software's, infrastructure,servers etc..Load balancing in a cloud computing environment is an important factor which affects the performance.Presently , the usage of internet and related resources have been increasing rapidly. Because of this there is an huge increase in workload. Hence there is an irregular distribution of this workload which results in a server overloading and might crash. In such systems, resources are not optimally been used.Because of this performance will degrade and efficiency reduces. Load balancing is the mechanism that shares the dynamic workload for all nodes in the whole cloud. Here Load balancing introduces a better load balance for public cloud based on partitioning of cloud concept.If the load is increased in cloud ,it will be analysed by using hadoop.The use ofMap Reduce in Hadoop ,will divide the into logical chunks and every chunk may firstly processed in parallel, by a map job.Different load balancing algorithms have been proposed in order to manage the resources of service provider in an well organised format and effectively. This paper presents a comparison of various policies utilized for load balancing.**

***Keywords*-cloud computing;loadbalancing model; public cloud; cloud partition; Main controller; Balancer; Load balancing algorithm; Hadoop.**

I. INTRODUCTION

In a present days cloud computing is very popular and it is still evolving standard. It is fast growing area in a computing research and industry today. Cloud computing is on demand service where information, shared resources, software and other devices are given according to client requirements at specific time. Cloud is a term which is used generally in case of internet. In this environment users need not to own the infrastructure for different computing services here user can access their data from any computers and from any part of world.

Several services and models make cloud computing advantageous and accessible to the end user. It has two models they are: Deployment model and service model. Here deployment model has three categories i.e., public, private, and hybrid cloud. Here private cloud is also called as internal cloud. Private cloud is used in an organization that needs more control over their data compared to that is provided by the third party organization. Public cloud gives services to anyone through the internet. Here in public cloud business rents capability and they will pay for what they use. Private and public cloud together forms the hybrid cloud.

Service model provides three types of services i.e., Infrastructure-as-a-service (Iaas), Platform-as-a-service (paas) and Software-as-a-service(saas)[1]. Here in Iaas only network is provided where as in Paas ,network and operating system will be provided and in Saas the software's and network will be provided . cloud service is popular because it reduces the convolution of network and users need not buy software licenses and information in cloud will not be misused.Above fig shows the architecture of cloud in which particular users will be connected to cloud from their own personal computer over the internet. For these particular users, the cloud will be visible as a single application. Hardware in the cloud and also operating system that maintains the hardware connection will not be visible

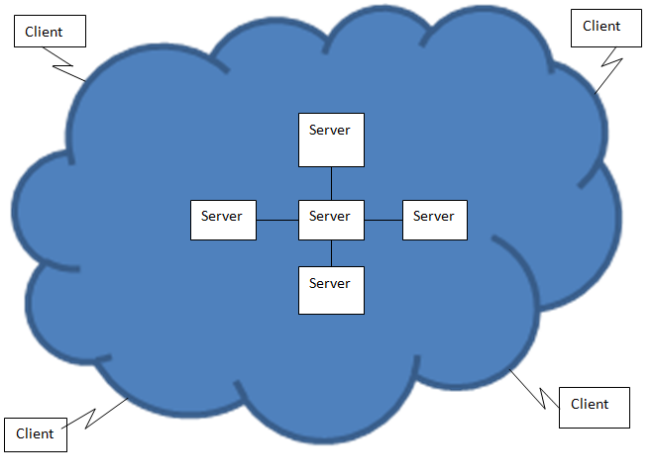

Figure 1: Architecture of cloud computing.

Above fig shows the architecture of cloud here particular users will be connectedto cloud from their own personal computer over the internet. For these particular users, the cloud will be visible as a single application. Hardware in the cloud and also operating system that maintains the hardware connection is not visible.

ISSN: 2278 – 1323

Big data is a set of huge datasets that cannot be processed by traditional computing technique. Big data is not merely a data; relatively it has become an complete subject, which involves different techniques, tools and frameworks. Big data technologies are important in providing more accurate analysis, which may lead to further actual decision-making resulting in a greater operational efficiency, reduced risk for business and cost reduction. To connect the power of the big data, you require an infrastructure that can be managed and process large volumes of a structured and unstructured data in a real-time and can protect information privacy and safety.There are several technologies in a market from various vendors like IBM, Amazon, Microsoft, etc., in order to handle big data.

Hadoopprovides an open source construction for cloud computing, as well as a distributed file system. Hadoop uses the Map Reduce model. HDFS is a file system that uses the Map Reduce to perform the tasks in which it reads the input in huge chunks, process it, and write potentially huge chunks of output. HDFS does not handle arbitrary access well. For reliability, file data is simply a mirror to several storage nodes. This is referred to as areplication in Hadoop community. As long as at least one replica of a data chunk is available, the consumer of that data will not know of storage server failures.

HDFS service is provided by two processes:Name Node handles the management of file system of the metadata and provides control services and management. Data Node provides retrieval services and block storage. There will be one Name Node process in an HDFS file system, and this is a single point of failure. Hadoop Core provide recovery and automatic backup of the Name Node, but there is no failover services. There will be several Data Node processes within the cluster, with typically one Data Node process per storage node in a cluster.

Although cloud computing has a lot of advantages it is still lacking in some tasks such as load balancing. Load balancing means when node has been overloaded that load has been shared over other nodes. It is a procedure that will share load to the various nodes in which it gives a good resource fabrication, when the nodes are overloaded. A good load balancing makes cloud environment well organized and hence improves user's fulfilment. Load balancing will be helpful for the users to perform the several operations. Load can be examined in a network load, cpuload . Load balancing aims in providing the good quality or tries to provide better size and performance to improve efficiency and when system fails it should have the backup plans, shouldincrease the user fulfilment, improve the resource fabrication, increase the accessibility, minimize waiting time of task in queue as well as decrease task execution time [5] [6].

Cloud computing supports the researches such as virtual machine as service on demand. Assigning well organized VM on demand can be achievedby load balancing algorithm. VM is a software which accomplishesthe computing environment, in which operating system or program can be implemented and it can run. Load balancing algorithm plays a vital role in which VM is to be assigned on demand for the user .While giving services it is achievable to have a number of requests at a time and due to some requests it has to remain in the queue even though requestor has the ability to send request for another service provider. Hence with the help of load balancing algorithm users are able to identify whether they can stay in queue or they might get service from other service provider[4]. There are Number of algorithm for load balancing in cloud computing is accessible for assigning the well organized VM. Here a study is achieved on various algorithm subsist for load balancing in a cloud computing.

II. RELATED WORK

Cloud computing is a recent technology in an IT industry. Confidential information towards researches advanced in a number of domains. There are number of studies on load balancing for cloud environment. Load balancing in a cloud computing has been described in the white paper [1] that was composed by Alder [2] who used tools and techniques frequently for load balancing in cloud. Chaczko et al.[3] described the role that load balancing plays to improve the performance and maintaining stability. Randles et al.[10] gave a compared analysis of few algorithms in cloud computing by examining the performance time and cost. In [11], Rodrigo presents an analysis for minimum amount of time and memory requirement to initialize an experiment while the hosts in the datacenter increases.

There are several load balancing algorithm namely round robin, equally spread current execution algorithm. Round robin algorithm is used here because it is simple and it is easy to implement. load balancing algorithm proposed in [12] can possibly improve the response time with respect to number of VMs in Datacenter.

III. PROPOSED SYSTEM

Load balancing model has been considered at public cloud which has several nodes with dispersed computing resources in various different geographic locations. Hence load balancing model split public cloud into various cloud partitions. When environment is very huge and difficult, these divisions make simpler to load balancing. Cloud has main controller that will choose proper partition for arrive task while balance for every cloud partition choose most excellent balancing tactic.

Public cloud based on paradigm cloud computing model with service gives by a service provider. A huge public cloud will contain lots of nodes and node in disparate geographic location. Cloud partition is a subzone of public cloud by means of division base on geographic location. After create cloud partition load balancer then start. While a work arrives at system, by the way main controller decides which cloud partition must obtain work. Partition load balancer decides how to allocate the work to nodes. As the load position of cloud partition is normal, the partition will be able to execute locally. If cloud partition is not normal this work would be migrate to another partition.

ISSN: 2278 – 1323

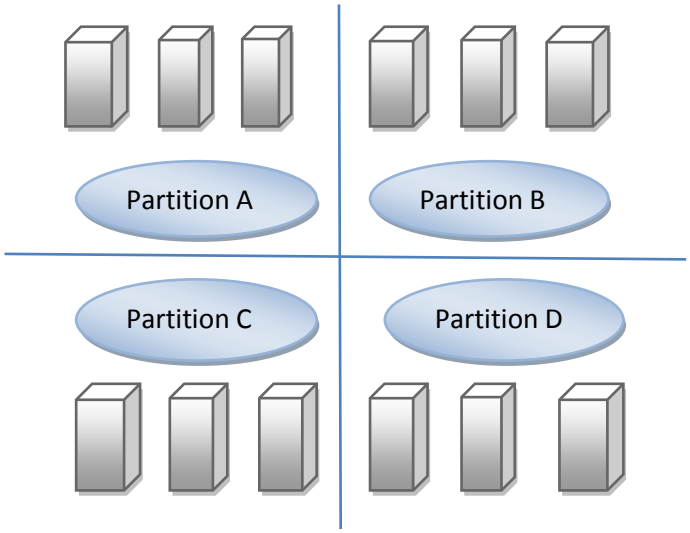

Figure. 2: Cloud partitions

Main controller and Balancer:

A public cloud has numerous nodes placed at a different physical location. A tiny part of this cloud is called a partition. Here system has a main controller, balancer and multiple servers and nodes. Main controller helps to choose a partition. Partition has been selected by best partition strategy..Load balancer solution has been made by main controller and balancer. Main controller first allocates work to suitable cloud partition and after that it communicates with balance in every partition to retrieve the status information. Main controller deals with information for every partition. Here balancer in every partition assembles the status information from each node and after that chooses suitable strategy to share job. Here best partition strategy help to choose to which partition request has to be allocate and hence status information is then checked a request has been allocate to the server which is having a least load.

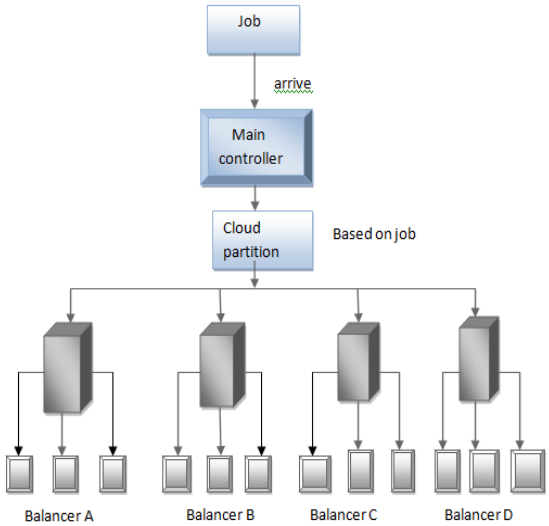

Figure. 3: Architecture diagram for proposed system.

IV. IMPLEMENTATION

In this paper we are implementing two vm scheduling algorithms that is Round robin and Throttled algorithm on Intel core i3 machine with 500 GB Hard disk and 4 GB RAM on Windows 7 operating system, Eclipse with Java version 1.6.

A. Round Robin Algorithm

It is a static load balancing algorithm, which does not consider the earlier load state of a node at the time of allocating tasks. It makes use of round robin scheduling algorithm for assigning jobs. It select the first node randomly and then, assigning tasks to all other nodes in a round robin manner. This algorithm works on arbitrary selection of the virtual machine. Thedata enter controller assigns the request to a list of VMs on a rotator basis. The first request is assigned to a VM chosen arbitrarily from the group and then Data Centre controller allocates the requests in anspherical order. Once the VM is selected the request, the VM is shift to the end of the list.

Round-robin is a simplest algorithm accessible to distribute load between nodes. Since for this reason it is frequently the first choice when implementing simple scheduler. One of the reasons for it being so simple is that the only information required is list of nodes.

The round robin algorithm is as follows:

Step1: Round Robin VM load Balancer maintain an index of a VMs. At begin all VM's contain zero allocation.
Step2**:**
a. The data centre controller receives user request/cloudlet.
b. The request is assigned to VMs in spherical way.
c. The round robin VM load balancer will assign the time quantum for a user request execution.
Step3**:** After the execution of cloudlet, VMs are de-allocated by Round Robin VM Load balancer.
Step4**:** The data centre controller check for new /pending/waiting requests in queue.
Step5: Continue from step-2.

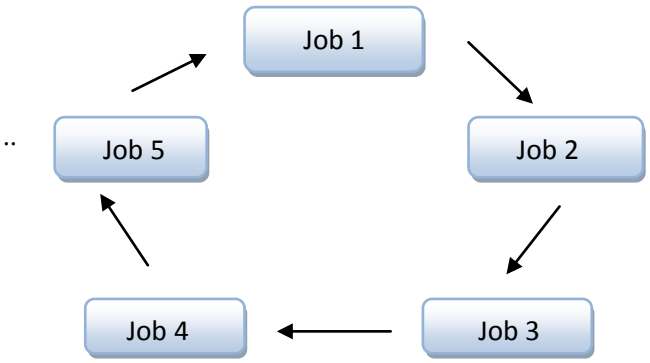

Figure. 4: Round-Robin processing

B. Throttled Algorithm

Throttled load balancer is a dynamic load balancing algorithm. In this throttled algorithm client first request a load balancer to find a suitable Virtual Machine to perform required operation. The process is first start by maintaining a list of VMs. Every row is separately indexed to accelerate

ISSN: 2278 – 1323

lookup process. If a match has been found on the basis of size and accessibility of machine, then load balancer accept the request of client and allocate that VM to client. On other hand, if there is no VM accessible that matches criteria then load balancer returns -1 and the request is queued.

The throttled algorithm is as follows:

Step1. Throttled VM Load Balancer keep an index table of VMs and the state of the VM (BUSY/AVAILABLE). At the beginning all VM's are available.
Step2. Data Centre Controller receives a new request.
Step3. Data Centre Controller queries the Throttled VM Load Balancer for the next allocation.
Step4. Throttled VM Load Balancer parses the allocation table from top until the first available VM has been found or the table is parsed fully

If found:
i) The Throttled VM Load Balancer returns the VM id to Data Centre Controller.
ii) The Data Centre Controller send request to VM identified by that id.
iii) Data Centre Controller notifies the Throttled VM Load Balancer of new allocation.
iv) Throttled VM Load Balancer updates the allocation table consequently
If not found:
i) The Throttled VM Load Balancer returns -1.
ii) Data Centre Controller will queue the request.
Step5. When VM finishes processing the request, and DataCentre Controller obtains response cloudlet, it inform Throttled VM Load Balancer of VM de-allocation.
Step6. The Data Centre Controller check if there are any waiting request in queue. If there are, it carries on from step 3. Step7. Continue from step 2.

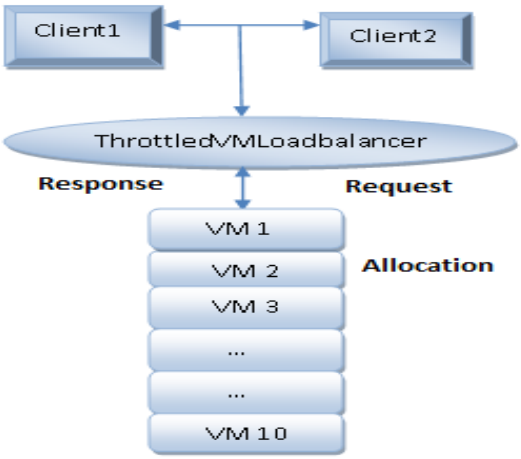

Fig 5 Throttled scheduling process

C: Cloud Analyst:

Cloud Analyst [7] [8] [9] is a GUI based tool that has been developed on a CloudSim architecture. CloudSim is an toolkit that is allowed to do modeling, simulation and other experimentation. The main problem inCloudSim is thatevery work has to be done programmatically. It allow user to do repeated simulation with small change in parameter very simply and rapidly. The cloud analyst allows setting location of the users that generating application and also location of data centers. In this various configuration parameter can be place like number of users, number of the requests generated per user per hour , amount of virtual machines, amount of processors, amount of storage, network bandwidth and other needed parameters. Based on parameters tool computes the simulation result and show them in a graphical form. The result consists of response time, processing time, and cost.

D:Map Reduce:

The interior concept of Map Reduce in a Hadoop is that input may divide into logical chunks, and every chunk may firstly processed in parallel, by a map job. The concequences of this individual processing chunk can physically partition into a different sets, which are then sort. Every sorted chunk has been passed to reduce job.

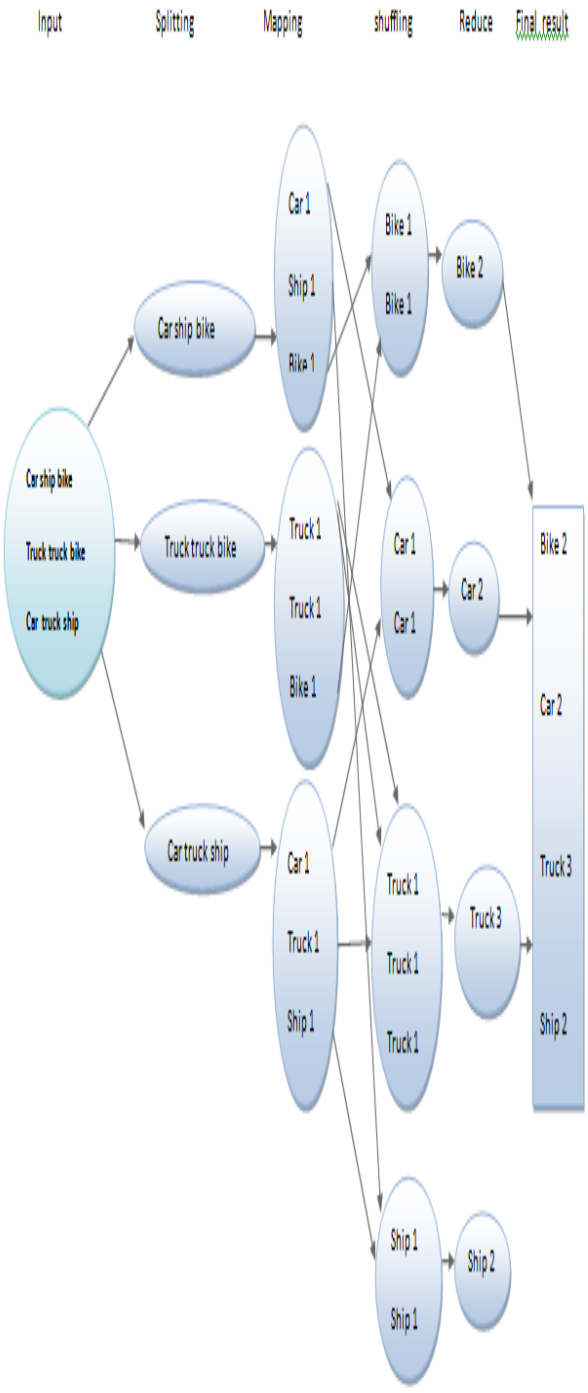

Figure6 :overall map reduce wordcount process

ISSN: 2278 – 1323

V. RESULT ANALYSIS

Table I: *parameters value*

parameter	value
Data centre OS	Windows 7
VM memory	512 MB
VM image size	10000
VM bandwidth	1000
Data centre arch	X86

Round Robin Load Balancer Algorithm and Throttled load balancer algorithm are implemented for a simulation. Java language has been used for implementing VM load balancing algorithm. Following table 2 and table 3 show a overall response time of VM load balancing algorithm.

Table II: Comparison of average response time of VM Load balancing

UB	RR	TR
UB 1	299.75	299.73
UB 2	301.19	301.262
UB 3	299.17	299.501
UB 4	300.43	299.941
UB 5	300.08	300.19

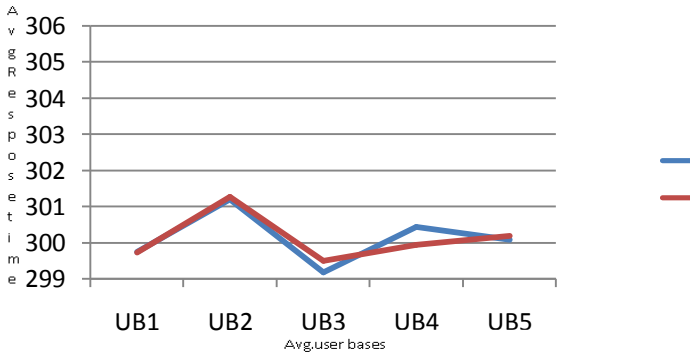

Fig7: Graph of variation of avg. response time based on number of users

VI. CONCLUSION

The whole goal of this paper is to balance load on cloud.Load balancing on the cloud will develop the performance ofcloud service substantiality. It will prevent overloading of the serverwhichdegrades the performance and response time will also beimproved.

This maybe used forwell organizeddata storage on clouds and the load balancing. This will help dynamically assigntasks (data) to least loaded server. Thus the whole performance of the cloud service will notaffect. It aims at having a backup plan in case system fails even partially. As well as work is done to be maintain system stability. There are requirements to accommodate the future modification in a system. Hence we have successfully gathered the information of project and confidentiality implement Load Balancing Model for better utilization and performances of the cloud service. We have simulated two different scheduling algorithms for executing user request in a cloud environment. Every algorithm is observed and their scheduling criterion likes average response time, data centre service time and total cost of different data centres are originate. We efficiently used the hadoop in order analyse the data which is enhanced in cloud.

ACKNOWLEDGMENT

We are grateful to express sincere thanks to our faculties who gave support and special thanks to our department for providing facilities that were offered to us for carrying out this paper.

REFERENCES

[1] M. Randles, D. Lamb, and A. Taleb-Bendiab, A comparative study into distributed load balancing algorithms for cloud computing, in Proc. IEEE 24th International Conference on Advanced Information Networking and Applications, Perth, Australia, 2010.

[2] B. Adler, Load balancing in the cloud: Tools, tips and Tools, tips and techniques, http://www.rightscale.com/infocenter/whitepapers/Load-Balancing-in-the-Cloud.pdf, 2012.

[3] Z. Chaczko, VenkateshMahadevan, ShahrzadAslanzadeh,ChristopherMcdermid. "Availability and Load Balancing in Cloud Computing" 2011 International Conference on Computer and Software Modeling.

[4] P. Mell and T. Grance, The NIST definition of cloud computing,http://csrc.nist.gov/ publications/nist pubs/800-145/SP800-145.pdf

[5] MR.Manan D. Shah, MR.Amit A. Kariyani, MR.Dipak L. Agrawal, Allocation Of Virtual Machines In Cloud Computing Using Load Balancing Algorithm, *International Journal of Computer Science and Information Technology & Security, Vol. 3, No.1,* February 2013

[6]YatendraSahu, R.K. Pateriya, Cloud Computing Overview with Load Balancing Techniques, *International Journal of Computer Applications, Volume 65– No.24,* March

[7] Bhathiya, Wickremasinghe."Cloud Analyst: A Cloud Sim-based Visual Modeller for Analysing Cloud Computing Environments and Applications", 2010,

[8] A. Singh, P. Goyal, S. Batra : Anoptimized round robin scheduling algorithm for CPU scheduling, International journal of computer and electrical engineerig (IJCEE), vol. 2, No. 7, pp 2383-2385, December, 2010.

ISSN: 2278 – 1323

[9] Tanvee Ahmed, Yogendra Singh "Analytic Study Of Load Balncing Techniques Using Tool Cloud Analyst" . International Journal Of Engineering Research And Applications. pp. 1027-1030, 2012.

[10] M. Randles, D. Lamb, and A. Taleb-Bendiab, A comparative study into distributed load balancing algorithms for cloud computing, in Proc. IEEE 24th International Conference on Advanced Information Networking and Applications, Perth, Australia, 2010, pp. 551-556.

[11]Modeling and Simulation of Cloud Computing Infrastructures and Services Grid Computing and Distributed Systems (GRIDS) LaboratoryDepartm[19] Rodrigo N. Calheiros Rajiv Ranjan, César A. F. De Rose and Rajkumar Buyya1(2011) : CloudSim: A Novel Framework for ent of Computer Science and Software Engineering The University of Melbourne, Australia

[12]ZhongXu, Rong Huang,(2009)"Performance Study of Load Balanacing Algorithms in Distributed Web Server Systems", CS213 Parallel and Distributed Processing

First Author: Nutan N was born in Karnataka, India. She received the B.E Degree in Computer Science and Engineering from Visvesvaraya Technological University Belgaum Dist., India in 2013 and M.Tech Degree in also same branch and University. Her research interests are in the area of Cloud Computing and Big Data Analytics.

Second Author: Girish .L. M-tech was born in Karnataka, India. He is working as Asst.Prof. In Computer Science Engineering Dept., CIT, Gubbi, Tumkur (district), Karnataka, INDIA. His area of interest are in the area of cloud computing, SDN and Big data.

ISSN: 2278 – 1323