A Review of improving load balancing in cloud environments using machine learning

[1]Shital Haribhau Yadgire, [2]Prof. Manoj Kathane, [3]Prof. Yogesh B Jadhao

***Abstract*: Cloud Computing is a rising area in the field of information technology (IT). Cloud computing helps to contribute data and provide many resources to users. Users compensate only for those resources as much they used. Cloud computing stores the data and disseminated resources in the open environment. The amount of data storage increases rapidly in open environment. Load balancing is one of the main challenges in cloud computing environment. It is a technique which is required to distribute the active workload across multiple nodes to ensure that no single node is overloaded. Load balancing techniques helps in best possible consumption of resources ultimately enhancing the performance of the system. The goal of load balancing is to minimize the resource utilization which will further reduce energy consumption and carbon emission rate that is the dire need of cloud computing. There are a variety of scheduling algorithms that maintain load balancing through knowledgeable job scheduling and resource allocation techniques. The aim of this paper is to discuss briefly some of the cloud concepts, the existing load balancing techniques and present a proportional study of the same.**

***Keywords*: Cloud computing, Deployment Models, Service Models, Challenges, Load Balancing, Algorithms.**

I Introduction

Cloud computing is today’s most current technical topic nowadays with extensive range of effects across IT, Infrastructure, Information Architecture, Business, Software Engineering, and Data Storage or Data Analysis and many more. Cloud Computing is new and inventive technology that is revolutionizing day by day. The key concept of cloud computing states that you need not to buy the hardware or software, you need much more. Rather you rent some of computational power, storage space, databases, and other resource you need to by a provider according to a pay-as-you-go model that makes your investment smaller. It's an additional model for empowering help to any user on-request organize access to common pool of configurable registering assets, for example, system, server, stockpiling, administrations and applications. Cloud provides a cost effective pay only for the resources which they consume. Cost is one of the main reasons for the success of the cloud.

Characteristics of the cloud

On demand service: Cloud computing provides resources and services as per the demand of the users. This does not involve interacting with the cloud service providers.

Broad Network Access: The cloud resources can be accessed anywhere on the network together with laptops, tablets and smart phones. Resource pooling: Both the storage and computing resources are pooled to accomplish multi-tenancy.

Rapid elasticity: The cloud services can be rapidly scaled up or down based on demand.

1. Horizontal scaling: It refers to introducing and removing server resources as per the demand.
2. Vertical scaling: It refers to altering the computing capacity of the already assigned server resources.
3. Measured service: Cloud computing implements the pay as you go model. The precise resources that are used are charged based on a previously specified metric.

II Architecture of the Cloud system

The architecture of the cloud refers to the components. These components are back end platform and a front end platform. The front end may contain thin clients or thick clients or mobile devices. The back end refers to the servers and storage space.

1. Clients:

A cloud client includes computer hardware and/or software that relies on cloud computing for application delivery, or that is particularly designed for delivery of cloud services and that, in either case, is essentially useless without it.

□ Thin clients: These naturally are used for display and not do any kind of computation. They do not have any internal memory as the servers do all the calculation and other work.

□ Thick clients: These typically use dissimilar browsers to connect to the cloud and internet.

□ Mobile client: The end devices are the mobile devices like smart phones and tablets. Mobile cloud computing is a division of cloud computing where at least some of the devices are mobile.

□ Datacenter: Datacenter is group of servers hosting different applications. An end user will have to join to the data centre to make use of the cloud applications a datacenter, geographically may be located at any distance from the cloud.

□ Distributed servers: A distributed server is a server that frequently checks the services of their hosts. Distributed servers are the part of a cloud which is hosted on the internet to provide services for various applications. The user gets a intellect that the application is being run on the user’s machine when one accesses the cloud.

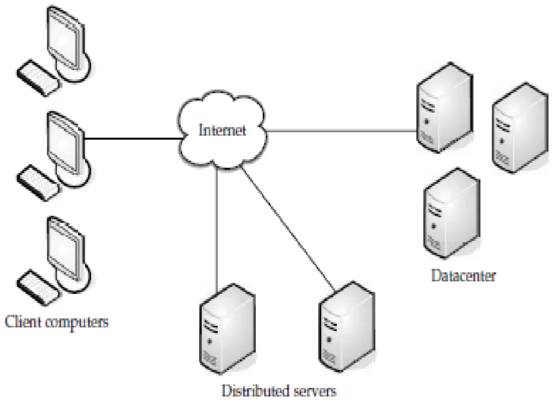

Fig.1 Components of the cloud

III Cloud models

1 Service models

3.1 Service models

□ Infrastructure as a Service (IaaS): Iaas provides the users the ability to provision computing and storage resources. These instances are provided to the users as virtual storage and virtual machine instances for use. Users can start, stop, configure and manage the virtual machine instances and virtual storage using services. Amazon Web Services (AWS), Microsoft Azure, Google Compute Engine (GCE), Joyent. Are all the examples for IaaS.

□ Platform as a service (PaaS): PaaS provides the users the capability to develop and deploy application in the cloud using development tools, application programming interfaces (APIs), software libraries and services provided by the cloud provider. The cloud service manages the underlying cloud infrastructure including servers, networks, operating systems and storage. Apprenda is an example for enterprise platform as a service model.

□ Storage as a Service (SaaS): SaaS provided the users a place for storage purpose. These servers may be spread all over the world.

3.2 Deployment models

1 Public cloud
2 Private cloud
3 Hybrid cloud
4 Community cloud

IV Load balancing

The process of redistributing and reassigning the larger processing load to the smaller processing nodes, in order to improve the performance of the system is called load balancing. IT basically must have a mechanism to process user requests and make the application run quicker. This entirely is optional and mostly depends on the clients business needs. Load balancing takes care of two significant things primarily to make sure of the availability of Cloud resources and secondarily to enhance the performance. This will ensure following things,

□ Resources are easily presented on demand.
□ Resources are competently utilized under condition of high/low load.
□ Reduced energy expenditure in case of low load, when the usage of the CPU cycles and memory falls below a certain threshold.
□ Decrease in the resource usage cost Load balancing helps in the allocation of computing resources to achieve proper resource utilization. High resource utilization with proper load balancing helps in minimizing resource consumption. It helps in implementing scalability and to avoid Bottlenecks. Load balancing techniques help networks and resources by providing a highest throughput with minimum response time. Load balancing is separating the traffic between all servers, so data can be sent and received without any delay with load balancing.

V Need of Load Balancing in Cloud Computing

Load balancing in clouds is a mechanism which distributes the surplus dynamic local workload evenly across all the nodes. It is used to achieve a high user approval and resource operation ratio, making sure that no single node is overwhelmed, hence improving the overall performance of the system. Suitable load balancing can help in utilizing the available resources optimally, thereby minimizing the resource consumption. It also helps in proper implementation of fail-over, enabling scalability, to avoid bottlenecks and over-provisioning, reducing response time etc.

Classification of load balancing algorithms:

1) Static approach: This is an approach that is mostly defined in the design or accomplishment of the system. In this approach, the load balancing algorithms split the traffic equally between all servers.

2) Dynamic approach: This is the approach that considers the current state of the system throughout load balancing decisions. This approach is more appropriate for widely distributed systems such as cloud computing.

Dynamic load balancing approaches have two types as follows

- distributed approach
- centralized approach.

a) Centralized approach: - In this approach, a single node is accountable for managing and distribution within the whole system.

b) Distributed approach: - In this approach, each node separately builds its own load vector. Vector collecting the load information of other nodes. The decisions are made close with the use of the load vectors. There are many load balancing algorithms which helps for better throughput and improve the response time in cloud environment. Each of them has their own pros and cons.

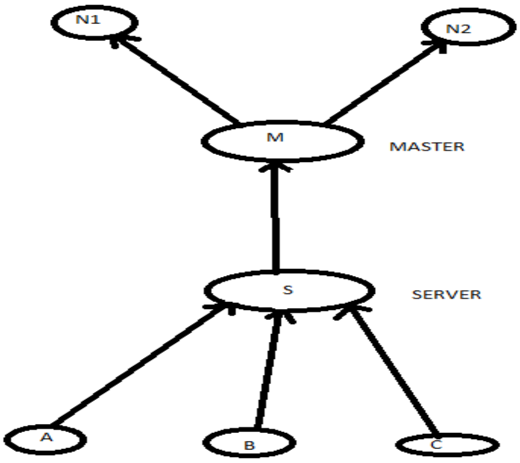

Fig 2 Flow of cloud computing

A. Dynamic Algorithms:

1. Honey Bee Foraging Algorithm: This whole algorithm is based on the process of honeybees' discovery of the food and alarming others to go and eat the food. First forager bees go and find their food and then after coming back to their respective beehive, they dance. After seeing the power of their dance, the scout bees follow the forager bees and get the food. The more active the dance is, the more food available is. So this whole process is mapped to the overloaded or under loaded virtual servers. The server processes the requests of the clients which is parallel to the food of the bees. As the server gets grave or is overloaded, the bees search for another location i.e. client is moved to any of the other virtual server. In this way, this whole technique works in honey bee foraging algorithm.

2. Ant Colony Optimization Technique: In this technique, a pheromone table was being planned which was updated by ants as per the resource consumption and node selection formulae. Ants move in forward direction in search of the overloaded or under loaded node for selection. As the overloaded node is traversed, then ants move back to fill the newly encountered under loaded node, so a single table is updated every time.

3. OLB and LBMM: Wang et al proposed a combination of opportunistic load balancing (OLB) and load balancing min-min (LBMM) algorithms to get the better performance of each tasks and to increase the task proficiency. All the tasks are assigned to each node in a specific manner and here the results are better than all other algorithms and also it can be used in LBMM. In LBMM it starts with carrying out of Min-Min algorithm at the first step. At the second step it chooses the smallest size task from the heaviest loaded resource and calculates the completion time for that task on all other possessions. Then the minimum completion time of that task is compared with the make span produced by Min-Min and if it is less than make span then the task is reassigned to the resource that produce it, and the ready time of both resources are updated .

5. Load Balancing MinMin Algorithm: An algorithm called LBMM which have a three level load balancing framework. In first level LBMM architecture is the request manager which is responsible for in receipt of the task and assigning it to service manager, when the service manager receives the request; it divides it into subtask and assigns the subtask to a service node based on availability of node, remaining amount of memory and the transmission rate which is responsible for execution the task.

B .Static Algorithms:

1. Map Reduce-based Entity Resolution: Map Reduce is a computing model and related implementation for processing and to generate large datasets. Map task and reduce task two main tasks in this model which are written by the user, Map takes an input pair and produces a set of middle value pair and Reduce task accepts an middle key and a set of values for that key and merges these values to form a smaller set of value. Map task read entities in parallel and process them, this will cause the Reduce task to be overloaded.

2. Min-Min Load Balancing Algorithm: Min-min load balancing algorithm begins by discovery of the minimum completion time for all tasks. Then among these minimum times, the minimum value is chosen which is the minimum time among all tasks on any resource. According to the minimum time, the task is then scheduled on the consequent machine. The execution time for all other tasks is updated on that machine and that task is detached from the list. This procedure is followed until all the tasks are assigned the resource. In scenarios where the number of smaller tasks is more than the number of larger tasks, this algorithm achieves better performance. However, this approach can lead to starvation.

3. Max-Min Scheduling Algorithm: U.bhoi proposed a Min-Min algorithm in cloud load balancing. Max-Min algorithm is approximately same as Min-Min algorithm, but in Max-Min it selects the task with maximum value and gives into respective machine. In this algorithm first verdict out minimum completion times, then the maximum value is chosen which is the maximum time among all the tasks on any resources. After that maximum time finding, the task is allocated on the particular chosen machine. Then the execution time for all tasks is upgraded on that machine, this is done by adding the execution time of all assigned task to the execution times of other tasks on that machine. Then all assigned task is detached from the list that executed by the system. Since the supplies are known beforehand, this algorithm is expected to perform well.

4. Round Robin Algorithm: In round robin the processes are distributed over all processors. Allocation order is maintained locally independent of the allocations from remote processors. In Round Robin, it send the requests to the node with the minimum number of connections, so at any point of time some node may be greatly loaded and other remain idle, this problem is reduced by CLBDM.

5. Round Robin Load Balancer (RR): Radojevic proposed an algorithm called as CLBDM (Central Load Balancing Decision Model). CLBDM is an improvement over in Round Robin Algorithm which is totally based on session switching at the application layer. A Round Robin algorithm which is an arrangement of choosing all elements in a collection equally in some well advised order, usually from the top to the bottom of a list and then again it'sstartingat the top of the list and so on. In this algorithm all the processes are separated between all processors. In this each process is allocated to the processor in a round robin order. The work load distributions between processors are equal. Different processes have different processing time. At sometimes the some nodes are heavily loaded and others stand immobile in web servers where http requests are of similar nature and distributed equally then RR algorithm is used.

6. Weighted Round-Robin Load Balancing Algorithm

Weighted round-robin was developed to improve the serious issues with round robin algorithm. In weighted round robin algorithm, each server is assigned a weight and according to the values of the weights, jobs are circulated. Processors which are having greater capacities are assigned with a larger value. Hence the highest weighted servers will receive more tasks. In circumstances where all weights become equal, servers will receive balanced traffic.

7. Genetic Algorithm: Load Balancing Improved Min-Min (LBIMM) or Map-Reduce algorithm are selected based on size of the task. Furthermore identifying the task with minimum execution time and resources and then working for the same. If task size is less than 50MB than load balancing is managed using load balancing improved min-min (LBIMM) and if task size is greater than load balancing is performed using Map-Reduce. For lower task, minimum completion time is calculated because some resources are already selected based on which it switches to honey bee or ant colony strategy. For huge task, minimum completion time is calculated using Map-Reduce. Here load balancing improved min-min (LBIMM) and Map-Reduce are offered as base strategy for lower task and huge task respectively and assuming flags.

VI Challenges for Load Balancing

There are some qualitative metrics that can be enhanced for improved load balancing in cloud computing as follows:

Throughput: It is the total number of tasks that have finished execution in a given period of time. It is necessary to have high through put for better execution of the system.

Associated Overhead: It describes the amount of operating cost during the execution of the load balancing algorithm. It is a masterpiece of progress of tasks, inter process communication and inter processor. For load balancing technique to work properly, lowest operating cost should be there.

Fault tolerant: We can define it as the ability to execute load balancing by the suitable algorithm without arbitrary link or node breakdown. Every load balancing algorithm should have a good fault tolerance approach.

Migration time: It is the amount of time for a process which is to be transferred from one system node to another node for execution. For better routine of the system this time should be always less.

Response time: In Distributed system, it is the time taken by a specific load balancing method to respond. This time should be minimum for better performance of system.

Resource Utilization: It is the parameter which gives the information within which the present resources are utilized. For competent load balancing in system, optimal resource should be utilized.

Scalability: It is the capability of load balancing algorithm for a system with any finite number of processor and machines. This parameter can be enhanced for better system performance.

Performance: It is the overall effectiveness of the system. If all the parameters are enhanced then the overall system performance can be enhanced. Load balancing is one of the main goals related to cloud computing. It can be a cache, memory, CPU capacity, network or delay load. Good Load balancing makes cloud computing more efficient, powerful and improves user satisfaction.

VII Literature Work

Gulshan Soni[3] et al. have presented "A Novel Approach for Load Balancing In Cloud Data Center". In this article researcher projected the logic of central Load balancer that manage the load on the cloud and assign the load corresponding to their priority and the response time is less as contrast to other load balancing algorithm.

Ruhi Gupta[1] have shown the "Review on Existing load balancing techniques in Cloud Computing". In this paper shows comparison of dissimilar techniques of load balancing in cloud computing. The researcher had done comparison and highlighted most recent techniques used in load balancing and based on that compares different parameters.

Vikas Kumar[4] et al. presented "A load balancing based cloud Computing Techniques and Challenges". In this paper researcher presented that load balancing is the main issue in cloud computing. It is compulsory requirement to distribute the surplus dynamic local workload evenly to entire load in the whole cloud system to achieve a high user satisfaction, efficient computer system and resource utilization ratio. The work of various researchers is analyzed and compared.

Foram F.Kherani[5] et al. presented “Load balancing in cloud computing” described the load balancing in cloud computing and how to increase and maintain the performance or throughput of cloud computing and also converse the comparison of different existing static load balancer and conventional dynamic load balancer method. In this paper mention the algorithms round robin, equally spread current implementation of load and throttled load balancing.

Rajvinder Kaur and Pawan Luthra [6] presented “Load balancing in cloud computing” presented in 2014 which states there are many algorithms in cloud computing which consist of many factors like scalability, better resource utilization, high performance and better response time for load balancing.

Kalyani Ghuge,Prof. Minaxi Doorwar[7] presented “A survey of various load balancing techniques and enhanced load balancing approach in cloud computing” in 2014 which states the concept of linear programming for load balancing which considered the factors such as cost for balancing and resource utilization.

Bhushan Ghutke, Urmila Shrawankar[8] presented “Pros and cons of load balancing algorithms for cloud computing” in 2014 which states that ESWLC concentrates on efficient load balancing and provides accurate results in cloud computing. CLDBM and ant colony algorithm reduces need of the human administration and provide faster information. LBMM use less response time for calculation and more efficient resource utilization in load balancing.

Bhatt Hirenkumar .H , Prof Hitesh A. Bheda presented [9]“An overview of load balancing techniques in cloud computing environments” IJECS in 2015 which did a comparison on of load balancing algorithms on the different criteria such as scalability, network overhead, resource utilization, algorithm complexity and response time.

“Process of load balancing in cloud Computing using genetic algorithm”(2015) This paper can do better load balancing with the help of Genetic Algorithm in the cloud environments.

“A Novel Approach for Enhancing Selection of Load Balancing Algorithms Dynamically Cloud Computing”(2017) By studying many researches and survey papers, different types of algorithms for load balancing came to know which are dynamic and static in nature. Furthermore it was decided to merge them to get better efficiency than to perform individually. Load balancing algorithms are good in distributing load equally to all virtual machines and perform task by increasing response time of the system. For enhancing such performances, we came up with novel approach to select algorithms dynamically based on some condition and situation in which they are enhanced, for which we need to calculate some parameters. Result shown proves that proposed algorithm.

Algorithm

```
1. For all tasks Ti
2. For all resources
3. Cij=Eij+Rj
4. Do until all tasks are mapped
5. For each task find the earliest completion time and the resource that obtains it
6. Find the task Ti with the minimum earliest completion time
7. Assign task Ti to the resource Ri that gives the earliest completion time
8. Check size of assigned task.
9. If task size > 50MB
Call Map-Reduce();
Go to Step 14
Else
Call LBMM();
10. Calculate parameters
11. Compute workload in wli , response time of each node in rti
12. Wl = wli , Rt = rti
13. If Wl > max
a. than max = Wl;
b. If Rt > max1
i. than max1 = Rt;
ii. Reschedule task with HoneyBee()
Endif
Else
iii. Reschedule task with AntColny()
End If
14. Delete task Tk from list
15. Update ready time of resource Rl
16. Update Cil for all i
17. End do
18. For all resources R Compute
makespan = max(CT(R)) End for for all resources
19. For all tasks
20. Find task Ti that has minimum ET in Rj
```

```
21. Find MCT of task Ti
22. If MCT < makespan
Reschedule the task Ti to the resource that produces it
End If
23. Update the ready time of both resources
24. End If
25. End For
26. End For
```

VII Future scope

Existing system suffers from drawbacks like, non optimal cloud usage, reduction in cloud efficiency, and increased delay. In the future, the overall efficiency of the cloud load balancing using machine learning based on genetic algorithm can be improved. The input parameters are server capacities, server memory, processing power of server cpu clock cycles as input to the genetic algorithm, which will produce effective load balancing solutions for the system. This will provide better results as compare to genetic algorithm.

Algorithm based on machine learning:

```
1. Initialize a random solution for load balancing
Nr = the selected nodes.
Lr = the load to each node.
and
∑Lr = total load.
2. Fitness = function of (Capacity of node, Memory of node and speed of the node)
3. Repeat steps 1 and 2 for k Rounds
4. Find mean value of fitness after k rounds, and pass the solutions where,
fitness > mean value
and discard all other solutions
5. Repeat steps 1 to 4 for 'i' iterations
6. Select the best fitness value solution.
```

In such a way the algorithm for load balancing can be somewhat more effective to the existing algorithm.

Conclusion

By studying many research and survey papers, different types of algorithms for load balancing we came to know which are dynamic and static in nature. Furthermore it is decided to merge them to get better efficiency than to perform individually. Load balancing algorithms are good in distributing load equally to all virtual machines and perform task by increasing response time and efficiency. For enhancing such performance, we came up with novel approach to select algorithms dynamically based on some condition and situation in which they are better, for which we need to calculate some parameters. In future, one can add one more algorithm for prediction of work load of each VM in cloud by improving the overall efficiency of the cloud load balancing using machine learning based on genetic algorithm.

REFERENCES

[1] Mayanka Katyal, Atul Mishra "A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment" IJDCC, 2013

[2] A Performance Analysis of Laod Balancing Algorithms in Cloud Environment, 2015 International Conference on Computer Communication and Informatics(ICCCI-2015),Jan,08-10,2015,Coimatore,INDIA

[3] Methodical analysis of various balancer conditions on public cloud division,@2015 IEEE, Anisaara Nadaph, Vikas Maral

[4] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi, Jameela Al-Jaroodi"A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms" IEEE, 2012

[5] Moumita Chatterjee, S.K.Setua "A New Clustered Load Balancing Approach for Distributed Systems" IEEE, 2015.

[6] Monika Rathore, Sarvesh Rai, Navdeep Saluja "Load Balancing of Virtual Machine Using Honey Bee Galvanizing Algorithm in Cloud" IJCSIT,2015.

[7] Huankai Chen, Professor Frank Wang, Dr Na Helian, Gbola Akanmu "User-Priority Guided Min-Min Scheduling Algorithm For Load Balancing in Cloud Computing" IEEE, 2013

[8] Lars Kolb, Andreas Thor, Erhard Rahm, “Load Balancing for Map Reduce based Entity Resolution” in IEEE 28th International Conference on Data Engineering 2012.

[9] Cloud Computing Features, Issues and Challenges: A Big Picture, 2015 International Conference on Computational Intelligence & Networks.Deepak Puthal DOI:10.1109/CINE.2015.31

[10] Shagufta Khan, Niresh Sharma “Effective Scheduling Algorithm for Load Balancing (SALB) using Ant Colony Optimization in Cloud Computing” in IJARCSSE, 2014.

[11] Ruhi Gupta,"Review on Existing Load Balancing Techniques of Cloud Computing" in International Journal of Advanced Research in Computer Science and Software Engineering,Volume 4, Issue 2, February 2014

[12] Randles, M., 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 Workshops (WAINA), Perth, Australia, April 2010.
[13] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al-Jaroodi, “A Survey of Load Balancing in Cloud Computing:Challenges and Algorithms” in IEEE Second Symposium on Network Cloud Computing and Applications 2012.
[14] A. M. Alakeel, "A Guide to dynamic Load balancing in Distributed Computer Systems", International Journal of Computer Science and Network Security (IJCSNS), Vol. 10, No. 6, June 2010, pages 153- 160.
[15] B. P. Rimal, E. Choi, and I. Lumb, "A Taxonomy, Survey, and Issues of Cloud Computing Ecosystems, Cloud Computing: Principles, Systems and Applications", Computer Communications and Networks, Chapter 2 , pages 21-46, DOl 10.1007/978-1-84996-241- 42, Springer - V erlagLondonLimited, 2010.
[16] R. Mata-Toledo, and P. Gupta, "Green data center: how green can we perfonn", Journal of Technology Research, Academic and Business Research Institute, Vol. 2, No. 1, May 2010, pages 1-8.
[17] S. Kabiraj, V. Topka, and R. C. Walke, "Going Green: A Holistic Approach to Transform Business", International Journal of Managing Information Technology (IJMIT), Vol. 2, No. 3, August 2010, pages22-3l.
[18] 1. Baliga, R. W. A. Ayre, K. Hinton, and R. S. Tucker, "Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport", Proceedings of the IEEE, Vol. 99, No. 1, January 2011, pages 149-167.
[19] Sandeep Sharma, Sarabjeet Singh, Meenaksshi Sharma, "Performance Analysis of Load Balancing Algorithms", World Academy of Science, Engineering and Technology, 2008.
20] Lars Kolb, Andreas Thor, Erhard Rahm, “Load Balancing for Map Reduce based Entity Resolution” in IEEE 28th International Conference on Data Engineering 2012.
21] Cloud Computing Features, Issues and Challenges: A Big Picture ,2015 International Conference on Computational Intelligence & Networks. Deepak Puthal DOI:10.1109/CINE.2015.31
[22] Shagufta Khan, Niresh Sharma “Effective Scheduling Algorithm for Load Balancing (SALB) using Ant Colony Optimization in Cloud Computing” in IJARCSSE, 2014.
23] Randles, M., 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 Workshops (WAINA), Perth, Australia, April 2010.