International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

A SURVEY ON ANALYSIS AND CLASSIFICATION OF WORKLOAD IN CLOUD

Chethan N[1], Mrs.Pushpalatha R[2] ,Dr.Ramesh Boraiah[3]
[1,2]Department of Computer Science and Engineering, VTU PG Centre, Mysuru,
[3]Department of Computer Science and Engineering, Malnad college of engineering, Hassan

Abstract: - Cloud computing is changing our lives in many ways. Any user adopting cloud computing services certainly expect the kind of improved performance that an cloud computing environment should provide. Workload is one of the major factor to accomplish high performance on Clouds. Workload classification would be a good solution to improve the performance. Analysis and Classification of the workload in cloud computing is challenging due to the virtualization layer overhead, complexity in workloads and insufficient availability of datacenter tracelogs. These elements adds lack of methodologies to characterize the applications hosted in the cloud. From the survey the workload can be classified based on the attributes like time bound, start time/end time etc, using clustering methods. The main objectives of this paper is to present an idea about why workload is important, why workload classification is important, complexities found in classifying the workloads in heterogeneous cloud environment, review of some methods for analysis and classification of workload. The workload can be either the synthetic or genuine workload. Google clusters traces, Yahoo M45 Hadoop cluster, PlanetLab cloud traces are some of the genuine cloud tracelogs available. Understanding characteristics of workload in cloud will help both cloud suppliers and researchers. Cloud suppliers can enhance their system Quality of Service (QOS) and researchers can assess new approaches using cloud simulators like CloudSim, CloudAnalyst, GreenCloud etc. For better resources management Workload must be classified as simple as possible. From workload classification, performance models can be constructed such as energy efficiency and resource management.

Keywords: Cloud Computing, Workload Classification ,Workload Analysis, Cloud trace.

I. INTRODUCTION

Cloud computing is a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. Cloud computing means it is a kind of Internet-based computing that provides shared processing resources and data to computers and other devices on demand.. The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, in consumer-oriented applications such as financial portfolios, to allow sharing of data-processing tasks, centralized data storage, online access to computer services or resources, immersive online computer games. To do this, cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing. Simple examples of cloud computing: Preparing documents over the Net is a newer example of cloud computing. Basically sign on to an online web-based service such as Google Documents and you can create a document, spreadsheet, presentation, or whatever you like using Web-based software. Rather than writing your words into a program like Microsoft Word or OpenOffice, running on your computer, you're utilizing comparative software running on a PC at one of Google's world-wide data centers. Clouds are being utilized as a platform for different types of applications with different Quality of Service (QoS) aspects, such as performance, availability and reliability.

Workload is defined as: "The average amount of work handled by a client, server or system in a given time period". The Cloud workloads have been recognized as follows: Websites, Online Transaction Processing, E-Commerce (E-Com), Mobile Computing Services and so forth. All of the online activity is delivered through data centers, and the more we send email, watch online videos, use social media like Facebook, and conduct business online, the more demands on data centers will grow. From the analysis of survey various workload handled by the cloud in 60second is shown in Table 1, close to 2.5 billion peoples are online around the world, and 70% of peoples use internet every day[16].Workload in cloud comprises of two components: tasks and users. Tasks are defined as the fundamental unit of computation assigned or performed in the Cloud and User is defined as the actor responsible for creating and configuring the volume of tasks to be processed. The workload comprises of a set of tasks, where every task belongs to a single job. A job may have multiple tasks. In order to better understand and describe tasks and improve the ability of Cloud, the analyzing of tasks is essential. The aim of workload analysis is to look at different aspects or characteristics of an enterprise application to determine the feasibility of moving the application in the Cloud. The following are some of the constraints with respect to workloads[2]:

a) A workload may have a particular begin time or elastic begin time.
b) A workload may have to finish by a certain time.
c) A workload may be time bound or time unbound.
d) A workload may have a certain lower limit of the resource that it needs.

VARIOUS WORKLOAD FOR EVERY 60 SECONDS	
EMAIL	**204 MILLION EMAIL ARE EXCHANGED**
ONLINE SHOPPING	**$272,000 OF ONLINE SHOPPING**
TWEETS ON TWITTER	**350,000 TWEETS ARE SENT ON TWITTER**
GOOGLE SEARCH	**5 MILLION SEARCH ARE MADE ON GOOGLE**
AUDIO DOWNLOAD	**15,000 TRACKS ARE DOWNLOADED VIA ITUNES**
LIKS ON FACEBOOK	**1.8 MILLION "LIKES" ARE GRANTED ON FACEBOOK**

Table 1: various workload handled by the cloud in 60second

Workload is one of the major factor to accomplish high performance on Clouds. Analysis and Classification of the workload in cloud computing is challenging because of the heterogeneous hardware is present in a data center , virtualization layer overhead brought about by I/O processing, complex workloads are made out of a wide variety of applications submitted at any time with various constraints and insufficient availability of genuine datacenter tracelogs for analysis due to business and confidentiality concerns. Understanding different attributes and constraints of workload in cloud will help both cloud suppliers and researchers. For suppliers, it enables a technique to upgrade resource management mechanisms to increase the productivity and Quality of Service of their systems. For researchers, it enables assessment of theoretical mechanisms supported by the characteristics of Cloud data centers through cloud simulators like cloudsim, cloudanalyst, Greencloud etc. because simulator contains abstractions for representing cloud infrastructures and power utilization[18].

1.1 Available Cloud Tracelogs

The workload can be either the synthetic or genuine workload. The synthetic workloads are valuable to carry out the controlled experiment. For performance evaluation of complex multitier applications,

the synthetic workload generation methods are required for example, in Banking, E-Commerce, Business deployed in the cloud computing environments. In every class, the application has its own particular attributes of workload. The important is that, the produced workload called synthetic workload should maintain the same attributes and behaviour of genuine workload [14]. At the present, there are a limited number of certifiable Cloud computing tracelogs are accessible. Tracelogs that are of limited observational period and to perform analyses are even more constrained. This is generally because of the business and confidentiality concerns of users and providers in commercial Clouds. The cloud vendors that has provided dataset from their Cloud computing clusters are GOOGLE, YAHOO, PlanetLab etc .As of late, Google has contributed by contributed two versions of tracelogs from their Hadoop MapReduce clusters. The first version ranges over a period of 7 hours with standardized processor and Memory use measurements at regular intervals 5 minutes. The trace describes the resource utilization for around in 9,174 jobs with 176,174 tasks . The trace has been public since December 2009. The second version of this trace spans 30 days and 12,583 servers in operation, giving data on 650,000 jobs with 25 million tasks. The trace has been public since November 2011[11].. The Google trace dataset is comprised of six separate data elements: Machine Event, Machine Attribute , Job Event, Task Event, Task Constraint and Task Usage. The total size of the data is approximately 250GB. Table 2 gives an overview of Statistical Analysis of Google tracelog dataset. Yahoo! which was made accessible for selected universities from their M45 Hadoop cluster. Here M45 data spans a period of 10 months from April 25, 2008 to April 24, 2009. The dataset comprises of 171,079 Hadoop jobs including large-scale graph mining, text and web mining natural language processing, machine translation problems, and data intensive file system applications, and large-scale computer graphics. Yahoo!, uses a Hadoop cluster of over 4,000 nodes, having 30,000 CPU cores, and 17 petabytes of disk space[12]. Table 3 gives an overview of Statistical Analysis of M45 dataset.

Trace span	29 Days	**Num of users**	430
Num of servers	12,532	**Avg users / day**	153.20
Num of tasks	17,752,951	**Avg task length**	61,575,043.48
Avg tasks / day	612,170.72	**Avg tasks / user**	3,981.06

Table 1: Statistical Analysis of Google tracelog

Log Period	Apr 25- Nov12, 2008, Jan 19- Apr 24, 2009	**Hadoop versions**	0.16: Apr 2008 0.17:Jun 2008 0.18:Jan 2009
Number of jobs **Successful jobs** **Failed jobs** **Cancelled jobs**	171079 165948 (97%) 4100 (2.4%) 1031 (0.6%)	**Average maps per job Average reduces per job Average nodes per job Maximum nodes per job**	154 ±558 σ 19 ± 145 σ 27 ±22 σ 299
Average job duration **Maximum job duration** **Node days used**	1214±13875 σ seconds 6.84 days 132624	**Number of active users**	31

Table 2: Statistical Analysis of M45 Dataset

The remainder of the paper is organized as follows. In Section 2, we provide an overview of work related. Section 3 Importance of Classification Workload in Cloud. Section 4 Challenges in Analysis

and Classification Workload. Section 5 Various Workload Analysis and Classification Techniques and finally, in Section 6, we conclude the paper.

II. RELATED WORK

A substantial amount of research has been devoted to the problem of analysis and classification of workload in cloud computing environment. In this section, the most relevant approaches are described, their limitations and gaps are also discussed.

Kuvulya et al. [12] present a statistical analysis of MapReduce traces. The analysis is based on 10 months of MapReduce from the yahoo M45 supercomputing cluster . Here, the authors present a set of coarse-grain statistical characteristics of the data related to resource utilization, source of failures, and job patterns. This work provides a detailed description of job completion times, but only provides very general information about the resource consumption and user behavioural patterns.
Aggarwal et al. [5] describe an approach to characterize Hadoop jobs. The analysis is performed on a data set spanning 24 hours from one of Yahoo!'s production clusters. This data set features metric generated by the Hadoop framework. The main objective of this work is to group jobs with similar characteristics using clustering to analyze the resulting centroids. This work only focuses on the usage of the storage system, neglecting other critical resources such as CPU , Memory , Disk space and network.

Mishra, et al. [1] describe an approach to construct Cloud computing workload classifications based on task resource consumption patterns. It is applied to the first version of Google tracelogs . The proposed approach identifies the workload characteristics, constructs the task classification, identifies the qualitative boundaries of each cluster, and then reduces the number of clusters by merging adjacent clusters. The approach presented is useful to create the classification of tasks. However it does not perform intra-cluster analysis to derive a detailed workload model. Finally, it is entirely focused on task modeling, neglecting the user patterns which are as important as the tasks in the overall workload model.

Solis et al. [11] provides an approach for characterizing Cloud workload based on user and task patterns using the second version of the Google tracelog. It presents coarse-grain statistical properties of the tracelog. This work has a number of limitations; the analysis performed is confined to only 2 days as opposed to the entire tracelog time span. Also, the cluster analysis and intra-cluster analysis do not contain sufficient detail to quantify the diversity of workload, instead presenting high-level observations. Finally, the validation of the simulated model against that of the empirical data is based only on a visual match of the patterns from one single execution, and does not consider more rigorous statistical techniques.

Sudha Pelluri et al[14] In this paper the real workload characteristics are used to generate the synthetic workload such that, the generated workload has similar characteristics and behavior as the real workload. The characteristic of real workload has been analyzed in IBM SPSS. The analyzed result was placed in the VMware workstation with Faban running on it.They have been able to generate synthetic workload which we are going to use in resource provisioning , load balancing energy management and other related research problems.

Rasheduzzaman et al.[15] This paper presents task shape and workload characterization of Google's compute cluster. The methodology for workload characterization consists of: (1) presenting the state of transitions of different job and how they are scheduled, failed, finished, and killed, (2)analyzing resource requests of memory, disk space and CPU-core using statistical tool, (3) showing the behaviour of different task type using cumulative distribution function, and (4) identifying the common job groups from resource usage table by using bisecting k-means algorithm. The results of

clustering analysis show that the largest clusters are very short time low memory core active jobs , while the smallest clusters are very long active jobs. This means that cluster management system do not need to keep inactive jobs in memory.

Xiaoyang et al[8]. In this paper, diverse sort of analysis such as coarse-grained analysis, cluster analysis and inner-cluster analysis were used to analyze task as well as task modeling. For analysis of workload they have utilized dataset from the second version of the Google MapReduce Cloud tracelog that features traces from over 12,000 servers over period of a month , which provides the normalized CPU, Memory and disk utilization per task in a timestamp every 5 minutes. Additionally, they have chosen CPU and Memory utilization attributes as dimensions of task model and compared it. And they have used k value of k-means clustering and some proper attributes can improve the accuracy of model. The experiment was done using MATLAB.

From the study of the related work it is clear that there are few available production tracelogs to analyze workload patterns in Cloud environments. By analysing these related work, we can identify the gaps that need to be addressed in order to achieve more realistic workload patterns. The analysis need to be other than croase-grain statics and cluster analysis. The workload is driven not only by task characteristics but also include the user behavioral patterns.

III. SIGNIFICANCE OF WORKLOAD CLASSIFICATION IN CLOUD

The rapid development of cloud computing technology have become the main service pattern on the Internet. The aim of cloud computing technology is to provide services that suit all of a user's demands. However, due to huge set of cloud service resources, it is difficult to identify the suitable services for the various types of users. Any user adopting cloud computing services certainly expect the kind of improved performance that an cloud computing environment should provide. Workload is one of the major factor to accomplish high performance on Clouds. Executing too many workloads on a single resource will cause workloads to interfere with each other and result in unpredictable performance which might discourages the users. On the other hand, users want their workloads to be done at minimal completion time. A classification of cloud workload would be a good solution to this problem.

ADVANTAGES OF WORKLOAD CLASSIFICATION:

- Different Workload patterns can be identified like different attribute, constraints etc .
- Understanding of workload patterns and leads to better resources management so that performance models can be constructed such as energy-efficiency.
- Workload can be analyzed at the group level, rather than at the individual server level.

In addition, classification of workload in cloud computing enables performance analysis and simulation, which brings benefits to cloud suppliers and researchers.

- Cloud suppliers can enhance the resource management mechanisms to effectively improve the diversity of users and tasks to increase the productivity and Quality of Service of their systems. For example, identifying the heterogeneity of task to reduce performance interference of physical servers or analyzing the correlation of failures to resource consumption.
- For researchers, simulation of Cloud workload enables evaluation of theoretical mechanisms supported by the characteristics of Cloud data centers. The use of simulation models enables the production of tracelogs based on realistic scenarios, filling the gap of previously identified in the cloud computing area. Nowadays, simulation-based approaches become popular in industry and academy to evaluate cloud computing systems, application behaviours and their security, the evaluation of these policies without deployment and execution of the applications in expensive large-scale environments

IV. CHALLENGES IN ANALYSIS AND CLASSIFICATION OF WORKLOAD

Workload is a critical factor to achieve high performance on Clouds. Workload analysis and classification is especially challenging when applied in a highly dynamic environment such as cloud computing environment, for various reasons:

1. Due to business and confidentiality reasons, only few genuine cloud tracelogs available for analysis. Thus, there is a lack of methodologies to characterize the different behavioural patterns of cloud applications. This is a particular challenge in academia, which relies on the very few publicly available Cloud tracelogs.
2. The cloud hosts a wide variety of applications submitted at any time, with different characteristics and user profiles, which have heterogeneous and competing QoS requirements. This leads to complex and massive size of workloads depending on users behaviour and resource consumption. Thus, it is challenging to predict workload patterns over time.
3. The virtualization layer promotes an overhead caused by I/O processing and interactions with the Virtual Machine Manager (VMM). This overhead depends on the hardware platform.
4. An in-depth statistical analysis and classification of workload diversity, within a large-scale production Cloud. This is due to the massive size and complexity of workload .

V. VARIOUS WORKLOAD ANALYSIS AND CLASSIFICATION TECHNIQUES

In this section, we summarize various techniques for analysis and classification of workload cloud computing identified during literature survey. The following are various workload analysis and classification techniques and these techniques are summarized. Table 3 gives the Comparison of Various Workload Analysis and Classification Techniques.

As mentioned earlier workload in cloud comprises of two components: tasks and users[1].
Users are responsible for driving the volume and behaviour of tasks based on the amount of resources requested for their execution. Therefore, three important characteristics that will refer to as dimensions are fundamental to describe the users' shape: the submission rate (α), Memory (ϕ) and the estimation ratios for CPU (β).
Tasks are defined by the type and amount of work dictated by users, resulting in different duration and resource utilization patterns. Consequently, essential dimensions to describe tasks are: Memory (π), length (χ) and average resource utilization for CPU (Υ) .

The Cloud workload can be described as a set of users with profiles *U* submitting tasks classified in profiles *T*, where each user profile u_i is defined by the probability functions of α β and ϕ, and each task profile t_i by χ , Υ and π determined from the tracelog analysis. The model components are formalized as below:

$$\mathbf{U = \{u_1, u_2, u_3, ...u_j\}}$$
$$\mathbf{T = \{t_1, t_2, t_3,.....,t_j\}}$$
$$\mathbf{u_i = \{f(\alpha) , f(\phi), f(\beta) \}}$$
$$\mathbf{t_i = \{ f(\pi), f(\chi), f(\Upsilon) \}}$$

Dempster–Shafer fusion[18] : It is a machine-learning based classifiers of different workload models. The classification among application workload running in virtualization gives interesting potential applications. For example, if the benchmarks are chosen appropriately, it may be determined what are the main characteristics of processes running in the virtual machine. Another possibility might be to know what are the processes that a given customer typically execute.

Feedback Based Workload Classification(FBWC)[10]: FBWC classification model to classify virtual machines efficiently and accurately. FBWC model consists of metrics collector, data preprocessor, TSRSVM classifier, decision maker and operating system tuner, these five elements cooperate with one another to work well.

k-means clustering [15]: The k-means clustering is a popular data-clustering algorithm to divide n observations into k clusters, in which values are partitioned in relation of the selected dimensions and grouped around cluster centroids. The k-means algorithm assigns every data point to cluster with the closest mean. Cluster centres are initialized randomly with the k-means function.

Cluster Analysis: Cluster analysis is used to classifying the different types of tassks. Clustering groups data, based on observations or features. The objective of cluster analysis is to identify groups of tasks with very similar resource consumption and assign them in clusters[8].

Intra-cluster analysis :The intra-cluster analysis consists of studying the data distributions for each one of the cluster dimensions. The process requires fitting the data from the logs to specific distributions using a Goodness of Fit (GoF) test to obtain the parameters of their Probabilistic Distribution Functions (PDF). The objective is to use these PDFs as components of the workload model[8].

Authors	**Analysis Methodology**	**Analyzed Components**	**Analyzed Parameters**	**Application type**
Kavulya [12]	Coarse-gain	Task	Task duration	MapReduce
Aggarwal[5]	Cluster centroids	Task	Task disk usage	MapReduce
Deborah Magalhães[13]	Distribution analysis	User & Task	Task resource usage	Web
Solis [11]	Cluster centroid & intra-cluster analysis	User & Task	User resource estimation & task resource usage	CloudSim simulator

Table 3: Comparison of Various Workload Analysis and Classification Techniques

5.1 LIMITATIONS

Although previously approaches offer some insights about workload characteristics, they do not provide a structured model which can be used for conducting simulations. The approaches, previously described completely focus on tasks, neglecting the impact of user behaviour. The workload is always driven by the users, therefore realistic workload models must include both user behavioural patterns and tasks characteristics. Only some of the workload characteristics like cpu, memory, network were considered. Have analyzed the workload for limited number of periods.
Some of the limitations can be overcome by proposed methodology

5.2 PROPOSED METHODOLOGY

The steps in the proposed methodology as follows:

a) Workload is analyzed. while analyzing the workload ,required resource metrics cpu, memory, network are also calculated.
b) In server, workload will be classified into high, medium and low workload respectively. This classification is based on the required resource metrics like cpu, memory , network.
c) Workload are clustered based on similar resource consumption, like the workload which required high resource are clustered in to one group, the workload which required medium resource are clustered into another group and workload which required low resource are clustered into another group respectively.
d) After clustering the workload which require similar resource, the virtual machine are allocated based on available resources.

The detailed explanation about proposed methodology will be addressed in the next result analysis paper.

VI. CONCLUSION

During the past few years, cloud computing has been one of the fastest growing parts in IT industry. Any user adopting cloud computing services certainly expect the kind of improved performance that an cloud computing environment should provide. Workload is one of the major factor to accomplish high performance on Clouds. In this paper we have survey on analysis and classification of Workload in cloud computing environment is described. In this article, the importance of workload in cloud computing and why classification of workload is important in cloud computing environmesnt is addressed. By workload classification how both cloud suppliers and researchers will get advantage from it is also mentioned. Various reasons that are challenging for analysis and classification of workload in cloud computing is also mentioned.

REFERENCES

1. "Towards Characterizing Cloud Backend Workloads:Insights from Google Compute Clusters", by Asit K. Mishra,Joseph L. Hellerstein,Walfredo Cirne,Chita R. Das, interning at Google during summer 2009.
2. "Metrics based Workload Analysis Technique for IaaS Cloud " , by Sukhpal Singh , Inderveer Chana ,International Conference on Next Generation Computing and Communication Technologies (ICNGCCT 2014), Dubai, UAE.
3. "An Analysis of the Server Characteristics and Resource Utilization in Google Cloud",by Peter Garraghan, Paul Townend, Jie Xu , 2013 IEEE International Conference on Cloud Engineering.
4. "Quality-of-service in cloud computing:modeling techniques and their applications",by Danilo Ardagna, Giuliano Casale, Michele Ciavotta, Juan F Pérez and Weikun Wang, Ardagna et al. Journal of Internet Services and Applications 2014.
5. "Characterization of Hadoop jobs using unsupervised learning," Aggarwal, S. Phadke, and M. Bhandarkar, in Proc. 2nd Int. Conf.Cloud Comput. Technol. Sci., 2010.
6. "Workload Classification for Efficient Auto-Scaling of Cloud Resources", by Ahmed Ali-Eldin, Johan Tordsson,Erik Elmroth, and Maria Kihl, May 21, 2013.
7. "A Simulation Based Analysis and Modeling of Workload Patterns using the CloudSim Toolkit", by Abhilasha Singh, International Journal of Scientific Engineering and Research (IJSER),ISSN (Online): 2347-3878, Impact Factor (2014).
8. "Modeling the Task of Google MapReduce Workload", by Xiaoyang Lin, Piyuan Lin*, Peijie Huang, Linxiao Chen, Ziwei Fan, Peisen Huang, 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing.
9. "Experimentally Prove The Workload Complexities of a Large Scale Utility Cloud", by S.Muthukaruppan, R.Ebenezer Princy, International Journal of Advance Research In Science And Engineering ,IJARSE, Vol. No.4, Issue 03, March 2015 .
10. "Workload Classification Model for Specializing Virtual Machine Operating System", by Xinkui Zhao, Jianwei Yin, Zuoning Chen, 2013 IEEE Sixth International Conference on Cloud Computing.
11. " An Approach for Characterizing Workloads in Google Cloud to Derive Realistic Resource Utilization Models", by Ismael Solis Moreno, Peter Garraghan, Paul Townend, Jie Xu, 2013 IEEE Seventh International Symposium on Service-Oriented System Engineering.
12. "An analysis of traces from a production MapReduce cluster," S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan, in Proc. IEEE/ACT Int. Conf. Cluster, Cloud Grid Comput., 2010,.
13. "Workload modeling for resource usage analysis and simulation in cloud computing" by Deborah Magalhães, Rodrigo N. Calheiros ,RajkumarBuyya,Danielo G. Gomes, Computers and Electrical Engineering 47 (2015)
14. "SYNTHETIC WORKLOAD GENERATION IN CLOUD", by Sudha Pelluri, Keerti Bangari, Volume: 04 Special Issue: 06 | NCEITCS-2015 | May-2015
15. "Task Shape Classification and Workload Characterization of Google Cluster Trace", by Md. Rasheduzzaman, Md. Amirul Islam, Tasvirul Islam, Tahmid Hossain and Rashedur M Rahman, 2014 IEEE.
16. "America's Data Centers Are Wasting Huge Amounts of Energy" ,by Pierre Delforge , Natural resources Defense Council.
17. Modeling and Simulation of Cloud Computing: A Review", by Wei Zhao, Yong Peng, Feng Xie, Zhonghua Dai, 2012 IEEE Asia Pacific Cloud Computing Congress.
18. "Coarse-grained Workload Categorization in Virtual Environments using the Dempster-Shafer Fusion", by 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD).

