Statistical Techniques for Characterizing Cloud Workloads: A Survey

Kenga Mosoti Derdus*, Vincent Omwenga O.
Faculty of Information Technology, Strathmore University, Nairobi, Kenya
*Email: derduskenga [AT] gmail.com

Patrick Ogao J.
School of Computing and Information Technology, Technical University of Kenya, Nairobi, Kenya

Abstract - **Cloud computing infrastructure is becoming indispensable in modern IT. Understanding the behavior nd resource demands of cloud application workloads is key in data center capacity planning, cloud infrastructure testing, performance tuning and cloud computing research. Additionally, cloud providers want to ensure Quality of Service (QoS), reduce Service Level Agreement (SLA) violations and minimize energy consumption in data centers. To achieve this, cloud workload analysis is critical. However, scanty information is known about the characteristics of these workloads because cloud providers are not willing to share such information for confidentiality and business reasons. Besides, there is lack of documented techniques for workload characterization. In this paper, we perform the first meticulous review on statistical techniques that can be used to characterize cloud workloads. In this review, we identify a statistical technique and its role in understanding cloud workload characteristics. Throughout the review, we point out relevant examples where and how such techniques have been applied. Additionally, we have shown the sources cloud workloads and their nature.**

Keywords: Cloud computing, characterizing cloud workload, statistical techniques, cloud workload analysis.

I. INTRODUCTION

Today, cloud computing has grown and has become indispensable on modern IT in supporting cloud service consumers, businesses, education entities and learning institutions [1]. This growth is because of the benefits cloud computing offer as compared to traditional computing. These benefits include cost saving, mobile access, flexibility and scalability and resource maximization [1]. Over a billion cloud users access a variety of cloud services such as search, financial services, gaming, social media and video streaming on a daily basis [2]. Cloud computing paradigm includes three service layers, which includes Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [3], and four cloud deployment models, which includes public cloud, private cloud, community cloud and hybrid cloud [4].

Cloud computing is mainly backed up by virtualization technology, which is based on physical resources abstraction in a way that several virtual resources are multiplexed on a physical one [5]. With virtualization, a physical machine (PM) or physical server is divided multiple small servers known as Virtual Machines (VMs), which can run different applications independently. The hypervisor or Virtual Machine Manager (VMM) is software layer, which induces the partitioning capability and may run directly on the hardware or on a host operating system [5]. Currently, many companies offering cloud computing services in all the service models mentioned earlier. They ranged from large companies such as Google, Amazon, IBM, HP, Facebook, and Salesforce to small companies such as Linode, Vultr, Cloud Sigma and Digital Ocean [1].

To continue the adoption of datacenters as well as improving existing datacenters and designing new ones, understanding cloud workload characteristics is critical. This requires a thorough analysis of cloud backend traces. This is important for datacenter engineers, researchers and cloud providers. Workload analysis can achieve many objectives such as understanding datacenter resource demand for planning, predicting system failures and performance evaluation [2]. Some of the largest cloud service providers such as Google, Facebook and Yahoo have published their backend traces. Although the publicly available backend traces represents a small subset of cloud service market, it has contributed invaluably to understanding cloud workloads [2] [6] [7] [1] [8].

To understand workload characteristics, analysis techniques have to be chosen wisely for achieving relevant goals. In this regard, statistical analysis techniques such as mean, moving averages, Coefficient of Variance, Correlation, Autocorrelation and clustering, which are all time series analysis techniques have been very critical. Time series analysis can reliably predict future resource demands in datacenters because cloud workloads collect application and human activity on the web over a period. Clustering generally employs K-means algorithm and can be used to group workloads with similar characteristics for purposes of scheduling. The goal of this work is to assess the applicability of statistical techniques for workload characterization. Particularly, we perform an end-to-end review of statistical techniques available for cloud workloads characterization and the rationale for each characterization. To the best of our knowledge, this is the first survey to comprehensively achieve this goal. The rest of this paper is organized as follows: Section II describes how cloud workloads are obtained. Section III describes the nature of cloud workloads and its attributes.

Section IV presents the rationale for cloud workload characterization. Section V presents comprehensive statistical techniques used in workload techniques, the goals they achieve, their weaknesses while drawing examples on how they have been applied in workload characterization. Finally, the paper is concluded in section IV.

II. NATURE OF CLOUD WORKLOADS

Ismael et al., [9] defines workload as "*a specific amount of work computed or processed within the datacenter with defined resource consumption patterns*". A typical system workload may include tasks to be performed, resource demands and task durations. Cloud workloads is a time series because it observes and collects collect application and human activity on the web over a period.

The nature of workloads can be described in different ways. According to [10], workload can be described according to its pattern - static workload, growing workload, on-and-off workload, periodic workload and unpredictable workload. A static workload has a constant number of user requests per minute. A growing workload's user requests per minute grow rapidly. A periodic workload shows seasonal changes. An on-and-off workload shows user requests that are processed periodical such as in batch processing or in experimentation scenario. Unpredictable workloads are of type periodic but are hard to predict. Due to the growing complexity of cloud computing paradigm, cloud workload elicit dynamism. Workload can also be described according to the information that is presented in system logs such resource usage tasks, task duration, task arrival rate and task volume [11]. For instance, GCT contains very useful information relating to submitted jobs and their tasks and machine properties in which they execute [12]. Using this workload, one can identify submitted jobs and its associated tasks, machine in which the tasks executed, task resource request, resources actually used, resources allocated in the cloud, user submitting the request, task constraints, priority, task start time and task end time. Workloads obtained for different purposes will not be similar.

Additionally, workload logs can be presented as raw or in obfuscated. Raw workloads traces are presented the way they were recorded from the system such that of GWA-T-12 [2]. On the other hand, obfuscating means transforming the raw workload traces to a new form, which makes it hard to decipher the infrastructure from which they were obtained [12]. However, the original patterns and relationship remain unchanged. Obfuscation is used for business and confidentiality reasons [12] [13].

Because cloud workload traces is collected over time interval, any analysis whose objective is to make prediction (such as future compute resource usage) may treat it as a time series [14] [7]

III. SOURCE OF CLOUD WORKLOADS AND WORKLOAD LOGS

There are three major sources of cloud workloads- real workloads, synthetic workload and workloads obtained by using workload generators [15] [13] [16] [17]. Real workloads are collected directly from a running system. Real workloads are the best for investigating cloud computing characteristics through characterization. Unfortunately, real workloads are scarce in the internet because clouds providers are not willing to publicly make them available because of business and confidentiality reasons [15]. Some of the publicly available real workloads include Clarknet, WorldCup trace 98, GCT, GWA-T-12, Facebook Hadoop workloads, OpenCloud Hadoop workload, Yahoo cluster traces and Eucalyptus IaaS cluster traces [2] [12] [15] [18].

Because of the scarcity of real workloads, synthetic workloads have become popular [15] [19]. Synthetic workloads are generated such that their characteristics resemble those of a real workload. [15] successfully generated synthetic workloads based on the characteristics of GCT using IBM's SPSS. Finally, workload generators are benchmark applications, which, by running then, generate workloads that meet certain characteristics [17]. Workload generators are configurable and can produce a wide range of workloads to meet different scenario. Already, many workload generators exist such as Rice University Bidding System (RUBiS) [4], Phoronix Test Suit [17] and TPC-W [20].

IV. IMPORTANCE OF WORKLOAD CHARACTERIZATION

Workload characterization is very important because it is applied in a number of areas such workload scheduling [17] [20] [9], workload prediction and resource planning [10] [7] [2] [21], synthetic workload generation [15] [16] [19], evaluation of workload failures [22] [23] and system performance [11] and security analysis [11].

Before mapping VMs and jobs to the PM, it is crucial to understand the characteristics of the workloads to be scheduled. For instance [4] and [24] notes that, workloads, which consume similar resources (homogenous workloads) should not be placed in the PM. For this reason, before mapping workloads to PMs, it is advisable to evaluate their resource consumption properties. Further, [25] has used statistical techniques by identifying the running times of jobs submitted into a cloud cluster. Scheduling entirely long jobs in a PM may have a detrimental effect. Mixing long and medium to short jobs is advisable. For this reason, jobs characteristics have to be known before hand through analysis.

Since cloud service providers are not willing to publish real cloud clouds, characterizing workloads can enable researchers to generate synthetic workloads, which behave like real workloads. This can be done for teaching and research. Using GCT on IBM SPSS, [15] has successfully characterized real workloads then generated synthetic workloads based on the results of characterization. Similarly, [26] and [16] have

used GCT and Georgia Tech Cloud Workload Specification Language (GT-CWSL) to generate synthetic workloads.

[22] has analyzed workloads from a production environment and as a result, failures in the cloud is understood. For example, failures of jobs and tasks has been analyzed and correlated with their scheduling constrains. This way, there are opportunities of predicting failures proactively thus reducing resource wastage.

V. STATISTICAL METHODS OF WORKLOAD CHARACTERIZATION

Statistical techniques for workload characterization can be categorized depending on the complexity or purpose. For purposes of this work, we categories these techniques into basic such as means and percentile, correlations and clustering.

A. Basic Statistics

The basic statistics that can be used to characterize workloads include minimum (min), maximum (max) or peak, percentages, percentile, frequency, standard deviation (SDev), unitless Coefficient of variation (CoV), cumulative distribution function (CDF).

During workload characterization max and min is used to report the maximum and minimum resources used by a given workload and this is a clear show on how resource requested by application workloads vary [2]. The resources include CPU, memory, hard disk and network bandwidth. [2] has successfully characterized VMs in GWA-T-12 workloads by reporting max and min resource usage. Mean is a very simple statistical method but very important. It can be used to report average CPU, RAM, hard disk and network bandwidth used by application workloads. It can also be used to quickly compare the difference between resources reserved by cloud service provider, resources requested by applications workloads and resources actually used by applications [16] [15] [26]. Moreover, mean of resource usage by VMs has been used to customize VM sizes [4]. The ratio of peak resource consumption and mean resource consumption has also been used to measure dynamicity of business critical workloads. In some cases, static consolidation techniques may use peak resource usage as a way of allocation resources to application workloads [27]. Percentages can also be used to compare the resources used by VMs and the resources that is actually used [28]. For example if servers use 10% of the memory used, it is said that memory is underutilized. Percentage resource usage has been used to create resources usage reference models. For instance, VMware Knowledge Base (VMware KB) uses a threshold setting for memory and CPU [29]. According to this reference, 80% CPU utilization is considered a ceiling and a warning if CPU utilization is 90% for 5 minutes. On the other hand, 85% memory utilization should be considered a ceiling and above 95% for 10 minutes is an alarm state

To understand the behavior of tasks or jobs submitted to a cloud environment, frequency is used. In relation to workload tasks, frequency of a given task is the number of times a particular task is submitted to a data center. This is very important because it can be used as a clustering feature as well as predicting resources to be used by that task [4]. SDev of each resource over time is used to report how the resources reserved, resources requested and resources used is spread [2] [30]. Because compute resources use different units, comparing the spread in different datasets is misleading. In this scenario, the unitless CoV, which is the ratio of SDev and mean, is used [31]. CoV has been used in [2] to compare variability of CPU, RAM and HDD used by application workloads in GWA-T-12 workload traces. Other research work, which have used SDev and CoV to report stability of workloads include [32] [33] and [23].

VM resource usage collected over time is treated as a time series. Use of statistical multiplexing is thus a reason why understanding resource usage peaks is very important. This is important when many VMs are co-located in the same physical server. Statistical multiplexing exploited where unutilized resources of one VM can be borrowed by a co-located VMs [34]. To achieve this, a time series is decomposed into remove the seasonal and trend component so as to predict the fluctuating component. This forecasting is used to ensure that futures resources needs do not affect current consolidation decisions. Clustering, based on VM peak points, can be used to ensure that VMs, which peak simultaneously are not co-residence in a physical server.

Percentile, which is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall, is yet another basic technique used to characterize cloud workloads. For a given application or VM workload, percentile is used to show the percentage of resources below a give value [35]. For instance, if there were data showing the resource consumption of an application running in a VM over a period, then one would determine the percentage of instances, when memory consumption was below a give point. This technique is more often used by researchers to allocate resources to VMs [2] [30] and has recently been used to design Googles GCE and Amazon's burstable instances [36] [37]. Use of percentiles has also been used to estimate transactions or tasks execution time, which is very crucial in negotiating SLA [38] [13]. Quartiles, which is in the same family as percentiles has been used to customize VM resources in [39]

A simple moving average (SMA) is a technique for characterizing resource usage captures in a time series and can achieve prediction, visualization or data preparation [40]. When used for visualization, it is normally used to remove noise frequencies and can bring out patterns otherwise hidden without smoothening [41].

As forecasting is common in time series data, the method chosen for forecasting depends on the behavior of time series data. For instance, in [41] the author uses Jargue-Bera test to

check if a smoothed time series follows normal distribution to decide on the method to use to forecast between neural networks (NN) and Autoregressive Integrated Moving Average (ARIMA). If smoothed time series is has normal distribution, ARIMA is used for forecasting, otherwise, NN is used. Many studies such as [42] and [43] have shown that NN can model a non-linear function thus able to handle complex time series than ARIMA models. Distribution of data in a time series can also visualized by using frequency distribution in a histogram and symmetry or skewness [44]

B. Correlations

Correlation characterization is used to study the relationships between different resource consumption over time [2]. This is revenant because workload logs corrected over time is considered a time series data. The commonly used techniques are Pearson Correlation Coefficient (PCC), which measures a linear relationship between two variable and spearman rank correlation coefficient (SRCC), which measures the relationships between two ranked series.

Understanding the relationship between resource usages can be used to achieve better VM consolidation techniques. [27] has built a technique of VM consolidation by first studying the correlation between workloads in terms of resource consumption. This work has also used correlations to determine workload, which peak (in terms of resource consumption) simultaneously and those that at peak at different times for purposes of consolidation. For instance, applications, which peak simultaneously, may not be co-located. The work presented in [2] investigated the correlation between resources requested and the one actually used in GCT. The author reports very weak correlation between the two. Moreover, with correlation analysis on workloads, one is able to choose between univariate and multivariate time series prediction. For instance, [7] and [45] has used correlations to identify relationships between co-clustered VMs for group level workload prediction.

Correlation analysis is also very important when determining the input variables for an NN prediction model [46]. The choice of input variable becomes a challenge because of a number of reasons such as availability of a large number variables for consideration, variable that may have no or little influence to the predicted variable and redundancy that may be caused by existence of highly correlated variables. The work in [47] has successfully used correlation analysis to select input variables to be input in predicting cloud resource provisioning in for web applications. According to the author, least correlated variables to the prediction target class are eliminated.

Autocorrelation of individual time series has been used by [2] and [7] to identify time patterns in the usage of resources. This is achieved by using the auto-correlation function (ACF) tool. Using ACF, [2] has shown setting different lag values, one is able to determine window sizes for resource prediction. To this end, the author has been able to identify short-term (few hours) predictions and daily patterns for different resources. Autocorrelation has also been used in [21] to measure workload periodicity.

Because the resources consumed by applications vary over time, it is advisable to report probability density function (PDF) and cumulative distribution function of values (such as PCCs and SRCCs) reported at different times [2].

C. Clustering

Clustering is an unsupervised learning technique where a larger group of items can be subdivided into several smaller groups [47]. As a result, clustering can be used in identifying objects with similar characteristics and those that are dissimilar. 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 centroid [47] [6]. The selected dimensions used to cluster objects is called a clustering feature set [4].

In [47], k-means has been used to group jobs into different groups based on task duration on GCT logs. Additionally, the same work has grouped workloads based resource usage patterns. In [48], k-means has been used to cluster VM on real traces collected for a period of 180 days from a 10-node data center. For each VM, the authors have considered 11 metrics as clustering feature set, which includes resource usages such as CPU, memory, disk and network bandwidth. The aim of this work is to identify the correlation between the resources usage of VMs. In [4], the author uses x-means, a modified version of k-means, to cluster GCT logs using resource usage, task priority, task length and submission rate as clustering feature set. The aim of this work is to group applications tasks, which are them mapped to the appropriate VMs to reduce energy consumption in a data center.

In [25], the authors have uses to k-means to cluster GCT jobs using 11 characteristics as a feature set. The authors discovered that 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. In [49], the authors have used a clustering techniques to group data center requests into groups. The obtained groups are then used to generate optimum mix ratio for workloads and an estimation of server capacity to minimize SLA violations. In [7], the authors applied a clustering technique on a time series to identify groups of VMs, which exhibit correlated workload patterns. Results from this technique is to be used to predict variations of workload patterns for resource planning.

VI. CONCLUSION

Understanding the behavior of cloud workloads is very important for many data center operations such as workload scheduling, workload prediction and resource planning, synthetic workload generation, evaluation of workload failures,

testing system performance and security analysis. This can be achieved through workload characterization. In this work, we have performed a meticulous review of statistical techniques, which can be used to characterize cloud workloads. As future work, we plan to use these techniques to characterize real cloud workload logs.

REFERENCES

[1] P. Jemishkumar, I.-L. Y. Vasu, B. Farokh, Jindal, X. Jie and G. Peter, "Workload Estimation for Improving Resource Management Decisions in the Cloud.," in *2015 IEEE TwelfthInternational Symbosium on Autonomous Decentralized Systems* , 2015.

[2] S. Shen, V. v. Beek and A. Iosup, "Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters," Parallel and Distributed Systems Section, 2015.

[3] R. Neha and J. Rishabh, "Cloud Computing: Architecture and Concept of Virtualization," *International Journal of Science, Technology & Management*, vol. 4, no. 1, 2015.

[4] F. P. Sareh, "Energy-Efficient Management of Resources in Enterprise and Container-based Clouds," The University of Melbourne , 2016.

[5] R. M. Sharma, "The Impact of Virtulization in Cloud Computing," *International Journal of Recent Development in Engineering and Technology*, vol. 3, no. 1, 2014.

[6] N. Chethan, R. Pushpalatha and R. Boraiah, "A Survey on Analysis and classifications of Workloads in the Cloud," *International Journal of Recent Trends in Engineering and Research*, vol. 2, no. 4, 2016.

[7] A. Khan, X. Yan, S. Tao and i. Anerousis, "Workload characterization and prediction in the cloud: A multiple time series approach," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*, 2012.

[8] Q. Xia, Y. Lan and L. Zhao, "Energy-saving analysis of Cloud workload based on K-means clustering," in *Computing, Communications and IT Applications Conference (ComComAp)* , 2014.

[9] S. M. Ismael, Y. Renyu, X. Jie and W. Tianyu, "Improved Energy-Efficiency in Cloud Datacenters with Interference-Aware Virtual Machine Placement," in *Autonomous Decentralized Systems (ISADS), 2013 IEEE Eleventh International Symposium*, 2013.

[10] A. Y. Nikravesh, S. A. Ajila and C.-H. Lung, "An autonomic prediction suite for cloud resource provisioning," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 6, no. 3, 2017.

[11] C. C. Maria, L. D. V. Marco, M. Luisa, P. Dana, I. M. Momin and D. T. Tabash, "Workloads in the Clouds," 2016.

[12] C. Reiss and J. Wilkes, "Google cluster-usage traces: format + schema," Google , 2011.

[13] M. Deborah, N. C. Rodrigo, B. Rajkumar and G. G. Danielo, "Workload modeling for resource usage analysis and simulation in cloud computing," *Computers and Electrical Engineering*, vol. 47, no. 2015, pp. 69-81, 2015.

[14] M. Amiria and L. Mohammad-Khanlia, "Survey on Prediction Models of Applications for Resources Provisioning in Cloud," *Journal of Network and Computer Applications* , vol. 82, no. C, 2017.

[15] K. Bangari and C. Rao, "Real Workload Characterization and Synthetic Workload Generation," *International Journal of Research in Engineering and Technology* , vol. 5, no. 5, 2016.

[16] A. Bahga and V. K. Madisetti, "Synthetic Workload Generation for Cloud Computing Applications," *Journal of Software Engineering and Applications*, vol. 4, no. 7, 2011.

[17] J. Smith and I. Sommerville, "Workload Classification & Software Energy Measurement for Efficient Scheduling on Private Cloud Platforms," in *Conference'10 University of St Andrews*, 2011.

[18] Delft University of Technology, "GWA-T-12 Bitbrains," 2015. [Online]. Available: http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains. [Accessed 5 March 2017].

[19] G. D. Costa, L. Grange and I. D. Courchelle, "Modeling and Generating large-scale Google-like Workload," in *The Seventh International Green and Sustainable Computing Conference* , Hangzhou, China , 2016.

[20] C.-Z. Mar, S. Lavinia, A.-C. Orgerie and P. Guillaume, "An experiment-driven energy consumption model for virtual machine management systems," 2016.

[21] A. A.-E. Hassan, "Workload Characterization, Controller Design and Performance Evaluation for Cloud Capacity Autoscaling," Ume°a University, Ume, Sweden, 2015.

[22] X. Chen, "Failure Analysis and Prediction in Compute Clouds," University of Science and Technology of China , 2014.

[23] I. Cano, S. Aiyar and A. Krishnamurthy, "Characterizing Private Clouds: A Large-Scale Empirical Analysis of Enterprise Clusters," in *SoCC '16 Proceedings of the Seventh ACM Symposium on Cloud Computing* , 2016.

[24] G. Dhiman, K. Mihic and T. Rosing, "A System for Online Power Prediction in Virtualized Environments Using Gaussian Mixture Models," in *in Proceedings of the 47th Annual ACM/IEEE Design Automation Conference (DAC)*, 2010.

[25] M. Rasheduzzaman, M. A. Islam, T. Islam, T. Hossain and R. M. Rahman, "Task shape classification and workload characterization of google cluster trace," in *Advance Computing Conference (IACC), 2014 IEEE International*, Gurgaon, India, 2014.

[26] S. Pelluri and K. Bangari, "SYNTHETIC WORKLOAD GENERATION IN CLOUD," *International Journal of Research in Engineering and Technology*, vol. 4, no. 6, 2015.

[27] A. Verma, G. Dasgupta, T. Nayak, P. De and R. Kothari, "Server Workload Analysis for Power Minimization using Consolidation," in *USENIX Annual technical conference* , CA, USA, 2009.

[28] Z. Ren, J. Dong, Y. Ren, R. Zhou and X. You, "Workload characterization on a Cloud Plartform: An early Experience," *International journal of Grid and Dsitributed Computing*, vol. 9, no. 6, pp. 259-268, 2016.

[29] VMware, "Performance Best Practices for VMware vSphere 6.0," VMware, Inc, Palo Alto, CA, 2015.

[30] R. Ganesan, S. Sarkar and A. Narayan, "Analysis of SaaS Business Platform Workloads for Sizing and Collocation," in *2012 IEEE 5th International Conference on Cloud Computing (CLOUD*, 2012.

[31] H. Abdi, "Coefficient of Variation," Sage, CA , 2010.

[32] A. Verma, G. Dasgupta, T. K. Nayak, P. De and R. Kothari, "Server Workload Analysis for Power Minimization using Consolidation," in *USENIX'09 Proceedings of the 2009 conference on USENIX Annual technical conference* , San Diego, California, 2009.

[33] A. Mishra, J. Hellerstein, W. Cirne and C. Das, "Towards Characterizing Cloud Backend Workloads: insights from Google compute clusters," in *ACM SIGMETRICS Performance Evaluation Review*, 2010.

[34] X. Meng, C. Isci, J. Kephart, J. Kephart, E. Bouillet and E. Bouillet, "Efficient resource provisioning in compute clouds via VM multiplexing," in *Proceedings of the 7th international conference on Autonomic computing*, Washington DC, USA, 2010.

[35] D. Feitelson, Workload Modeling for Computer Systems Performance Evaluation, Cambridge University Press , 2014.

[36] N. Nasiriani, C. Wang, G. Kesidis and B. Urgaonkar, "Using Burstable Instances in the Public Cloud: When and How?," 2016.

[37] A. Gilgur, t. Gunn, D. Browning, X. Di, W. Chen and R. Krishnaswamy, "Percentile-Based Approach to Forecasting Workload Growth," in *IT Capacity and Performance 41st International Conference*, San Antonio, TX , 2015.

[38] B. Ciciani, D. Didona, P. D. Sanzo, R. Palmieri, S. Peluso, F. Quaglia and P. Romano, "Automated Workload Characterization in Cloud-based Transactional Data Grids," in *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW)* , Shanghai, China , 2012.

[39] F. P. Sareh, R. N. Calheiros, J. Chan, A. V. Dastjerdi and R. Buyya, "Virtual Machine Customization and Task Mapping Architecture for Efficient Allocation of Cloud Data Center Resources," *The Computer Journal*, 2015.

[40] J. Brownlee, "Moving Average Smoothing for Data Preparation, Feature Engineering, and Time Series Forecasting with Python," Machine Learning Mastery , 2016. [Online]. Available: https://machinelearningmastery.com/moving-average-smoothing-for-time-series-forecasting-python/. [Accessed 01 November 2018].

[41] Q. Z. Ullah, S. Hassan and G. M. Khan, "Adaptive Resource Utilization Prediction System for Infrastructure as a Service Cloud," *Journal of Computational Intelligence and Neuroscience: Hidawi*, vol. 2017, 2017.

[42] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Systems with Applications: Elsevier*, vol. 37, no. 1, pp. 479-489, 2010.

[43] H. Xu, X. Zuo, C. Liu and X. Zhao, "Predicting Virtual Machine's Power via a RBF Neural Network," in *International Conference in Swarm Intelligence*, Bali, Indonesia, 2016.

[44] K. R. Das and R. Imon, "A Brief Review of Tests for Normality," *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, pp. 6-12, 2016.

[45] A. Adegboyega, "Time-series models for cloud workload prediction: A comparison," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, Lisbon, Portugal, 2017.

[46] V. Landassuri, R. Marcial-Romero, H. A. Montes-Venegas and M. A. R. Corchado, "Review of Input Variable Selection Methods for Artificial Neural Networks," in *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, K. Suzuki, Ed., InTech, 2011, pp. 20-44.

[47] A. A. Bankole, "Cloud Client Prediction Models for Cloud Resource Provisioning in a Multitier Web Application Environment," Carleton University, Ontario, Canada, 2013.

[48] M. Alam, K. A. Shakil and S. Sethi, "Analysis and Clustering of Workload in Google Cluster Trace based on Resource Usage," in *19th IEEE International Conference on Computational Science and Engineering* , Paris, France, 2015.

[49] C. Canali and R. Lancellotti, "Automated Clustering of Virtual Machines based on Correlation of Resource Usage," *Journal of Communications Software and Systems*, vol. 8, 2012.

[50] R. Singh, U. Sharma, E. Cecchet and P. Shenoy, "Autonomic Mix-Aware Provisioning for Non-Stationary Data Center Workloads," in *Proceedings of the 7th international conference on Autonomic g*, 2010.