International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-6)

Research Article | **June 2017**

Performance Analysis of Web Applications Working on Cloud Environment Using Workload Prediction Model Based on ANN

Supreet Kaur Sahi, Dr. V.S.Dhaka
JNU, Jaipur, Rajasthan,
India

Abstract—

Cloud computing is good fit for deployment of different applications but workload and instances requirement will vary depending upon type of applications. Workload estimation of cloud computing is tedious task. In cloud computing number of instances of cloud need to be reserved based on certain parameters. If these instance are under estimated then performance of system will reduce and if over estimated then cost will increase. In order to optimize this cost there must be some algorithm working that can help in reserving number of instances based on certain parameters. Web applications have unpredictable workload. Certain steps of capacity planning need to follow for predicting workload of web applications. This paper analysis performance of ANN based workload estimation model for web applications on cloud environment. A brief survey of literature is also presented to find out different parameters necessary for capacity planning of website.

Keywords—Capacity Planning, Cloud Computing, E-business, Neural Network, Workload

I. INTRODUCTION

According to recent survey by IDC (2017) global spending on public cloud, services and infrastructure will increase to $122.5 billion by the end 2017, which is almost 25% increase as compared to year 2016 [1]. According to Stamford (2016) IT expenditure is shifting from traditional IT offerings to cloud services. The collective amount of cloud shift in 2016 has reached $111 billion. It is expected to increase to $216 billion in 2020 [2]. Due to agility and operational efficiencies of cloud computing, organizations are curious to move different applications on cloud.

Identifying different workloads on cloud is cumbersome task. As for example web applications have possibility for unpredictable spikes however cloud computing can help as cloud providers can provide unlimited resources on demand. But this may lead to high cost. In order to optimize this cost there must be some algorithm working that can help in reserving number of cloud instances required for given applications. Workload of cloud can have different characteristics. It is different from traditional workload. Workload of cloud needs to be comprehended from different perspectives. There are different cloud service providers. Data available on their website can be used for analysis purpose [3].

Workload of cloud can also base on deployment models. Private cloud plays pivotal role when security and privacy is major workload requirement. In this type of workload maximum control is provided to organizations for data hosting. Public model supports elasticity and infinite capacity that are needed by IT organizations. By resource sharing peak workload of customers can be handled due to geographic distribution. Workloads with high security and processing requirements can use hybrid cloud as an option. This deployment model has features of both public and private cloud. Workload of a hybrid cloud integrates different hosting environments that can be accessed using different tenants. Community cloud comprises all pooled data and functionality that all participating companies need for their business. Based on current requirements of privacy and security, a community cloud can be stipulated in a private data centre regulated by the collaborating companies [4].

Cloud computing has changed the way IT resources allocated and use. In order to understand workload it is equally crucial to know how cloud offers services and the ways these services are delivered to end-users on different levels of application stack. It is also important to understand that under what conditions different applications gets benefits from these properties.

While calculating workload of different applications, multiple dimensions need to be considered. Some of these requirements are Legacy, Standard Front and Back office Applications, Batch or online applications, Frequency and cost involved.

Two types of workloads exist:

- Batch Mode Workload
- Real time Workload.

Workloads are self-contained entities. In this paper web applications are considered. Web applications play crucial role in different enterprise and cloud services. With growing social network admiration and high speed of Internet an online system faces unpredictable peak and mounting loads. Auto scaling is characteristic of cloud computing that supports workload management [5]. However, solutions available are subject to some of the following constraints [6].

- Relying on user for providing threshold values and scaling metrics.
- Scaling algorithm applied.

Different types of workloads are as follow:

- Static workload that almost remains constant over time period.
- Periodic workload with recurring peak.
- Once in a lifetime workload that achieve peak once during lifetime.
- Unpredictable workload varies randomly and frequently.
- Continuously changing workload will grow and shrink with time.

In cloud environment user only wish to pay for resources used by them so provider needs to apply rapid elasticity to meet user needs. Two cloud computing properties 1) Rapid Elasticity 2) pay-per usage plays crucial role in handling cloud workload. Elastic scaling flexibly allocates resource and hence once the increase is found new resources can be provisioned and when number of resources not required they can be decommissioned [6].

Some methodologies need to be followed to understand and forecast workload of cloud based application. Web application data is complex, as it will vary with respect to load. Neural networks have ability to originate meaning from complicated or imprecise data. It can be used to get patterns and identifying trends that are too difficult to be noticed. Hence, Neural network can be applied in order to forecast workload. MATLAB neural network toolbox can be used for implementing results as well as for analysing and understanding of model [7][8][9][10]

In this paper author has presented the workload prediction method for web-based applications running on the cloud environment. Also performance analysis is done using regression graphs to understand the results of model.

II. BACKGROUND

The growth of the cloud computing signifies a important change in the way information technology services are invented, deployed, developed, updated, scaled, maintained and paid. Computing today reflects a major change — on one hand, computers are becoming exponentially more and more powerful and the per-unit cost is falling rapidly [11][12]. On the other hand, as computing is becoming more pervasive, there is rising difficulty in management of entire infrastructure [13]. The cloud computing delivers all the functionalities of existing IT services with reduction in upfront costs of computing which prevent many organizations from deploying many IT services [14].

Cloud computing has emerged as a convincing paradigm for management and delivering services over the Internet [15]. The rise of cloud computing has changed the design of information technology, and turning the long-held potential of utility computing into an actuality [16]. Capital outlay is not required with the advancement of cloud computing for developers with inventive ideas for Internet services [17]. Workload prediction problem exist as a result of users continuously accessing different application on cloud that need to be handled automatically [18][19]. With advancement of technology customer wants to pay only for amount of resources used, there must be provision for rapid elasticity where number of resources allocated must increase or decrease on the basis of customer demand [20].

Quiroz et al [21] described four steps of resource management of cloud computing: Virtual Machine Provisioning, Resource Provisioning, Run-time Management and Workload Modelling. In this work focus will be on workload management. According to Chaisiri et al different types of provisioning plan provided by cloud provider are categorized as reservation and on-demand plans. Resources provisioning performed by reservation plan are inexpensive as compare to that provision by on-demand plan. The problem is that it is difficult to provision resources on demand due to uncertainty of consumer future demands. In order to address this problem, an optimal cloud resource-provisioning algorithm is proposed by framing a stochastic programming model. The purpose of model is to minimize cloud consumer total provisioning cost. Bender decomposition is also used in this algorithm. In this at every step, firstly complicated variables are identified to form master problem and then sub-problems comprises of decision variables are solved. The step repeats itself and it will stop when there is convergence in lower and upper bounds [22].

Ching-Chi Lin et al [23] discusses resource scaling for web-based applications on cloud. In this paper effective auto scaling strategy is presented. Data set used consists of 24 physical servers, each with the: quad-core X5460 CPU * 2 with hyper-threading, 16 GB memory, and 250 GB disk. The author is using incoming request per second for scaling of cloud resources.

Mian et al [24] has considered case of data intensive workload of cloud computing by mapping same request to same virtual machine. In this, the assumption is that any data intensive workload consists of requests that are issued by particular client of cloud belongs to set of particular data object. There is a mapping between request set and virtual machine. The provisioning problem that author has discussed has objective to optimize the solution to map same type of requests to same virtual machines. In this, paper a manager machine is proposed which have three parts:

a) Configurator - which is brain of system with the purpose of finding best configuration for executing workload
b) Scheduler - that maps workload request with corresponding execution platforms
c) Provisioner- that is assigned the task of preparing execution platforms.

Public cloud i.e. Amazon EC2 is used for experiment in this paper. The objective function of this paper has used minimal dollar-cost given by Amazon EC2. From this paper it can be concluded that mapping same type of request to same virtual machine in cloud environment plays a key role. Amazon EC2 can be considered for getting details of different parameters.

Zhu [25] suggests resource allocation plan with budget constraints for different adaptive applications in cloud environment. This paper presents Multi-input-multi-output feedback control model-based dynamic resource provisioning algorithm. The aim of this paper is the development of feedback control based model for application quality maximization within budget and time constraint. In this paper a framework design, implementation and validation is presented which supports dynamic application adaption in cloud environment. Chunlin [26] aims at profit maximization by serving SaaS user request. Amini [27] discusses virtual machine importance for resource provisioning. This paper presents large-scale system with both internal and external request types that need to be served.

The Van [28] and Ahmed [29] described that Workload burst and spikes as the main reasons behind service disruptions of website and decreasing Quality-of-Service (QoS). Some spikes are result of non-predictable event and load volumes while others are due to planned event with non-predictable load volumes. Ahmed [29] has taken sample size of seventy workloads in which ten are real and sixty are artificial traces. Samples of data include data from Google, wiki etc. From studying this paper it can be concluded that service rate plays critical role in workload prediction. The strength of the cloud as described by Border [30] does not only depends on single server implementations or rely on a single technology, it is due arrangement of building block in which multiple services can be deployed on multi-server architectures located in datacentres around the world to create global applications for the Internet. Roy [31] presents challenges in auto scaling of cloud applications whereas Sun [32] discusses applications where traces of Internet broadcast services workload on cloud is presented. The concept of resource pooling and rapid elasticity is used in this case. Raquel [33] have proposed analytical model that can help in planning IT infrastructure comprise of resources acquired from provider. The aim is cost optimized model of acquiring resources by guiding planning phase that consists of deciding how many cloud instances must be reserved from reservation market to be used in the future.
Following table presents list of different findings:

Table1 Findings from Literature

SNo.	Author	Findings
1	Chaisiri et al [22]	This Algorithm is able to adjust tradeoff between reservation of resources and allocation of on demand resources using different virtual machine classes.
2	Ching-Chi Lin et al [23]	Request per second is important parameter for workload prediction
3	Mian et al [24]	Amazon EC2 can be considered for getting details of different parameters.
4	Zhu [25]	This can be used as the base to find different hardware requirements. Also dynamic resource provisioning algorithm achieves is more beneficial as compared to static provisioning scheme.
5	Chunlin [26]	It considers SaaS layer for resource provisioning.
6	Amini [27]	The solution based on preemption of Virtual Machine based resources according to priority is suggested for resource allocation.
7	Van [28] and Ahmed [29]	Service rate plays critical role in workload prediction
8	Border [30]	Strength of cloud is discussed.
9	Roy[31] and Sun [32]	Concept of resource pooling and rapid elasticity is presented.
10	Raquel[33]	Business-driven approach and a model to plan these long-term contracts which aims at maximizing the profit is proposed. The data available in this paper can help in workload forecasting. This paper also consider threshold and peak time as parameters.

From above table different parameters obtained can be divided into different equivalence classes. Parameters belonging to same equivalence class can be grouped together. Each parameter can be taken as instance of a class Qi from a set Q = {Q1, Q2…,Qn} for Application A.

Examples of Qi are service rate, arrival rate, threshold etc.

The value used by A is a set of data objects Oi= {O1, O2…,Om}.

Values of application can be value of particular parameter on different conditions. Examples of conditions can be high load, average load, low load on cloud infrastructure,

From available literature it is also found that under provisioned resources leads to performance reduction and overprovisioned resources leads to higher cost. Hence it can be concluded that some optimization tools are required that can predict number of instances of cloud required in order to forecast the workload. Pattern for workload is described by user behaviour that result changes in utilization of IT resources. Different metrics on which workload can be measured are as follow:

- Number of user request
- Processing load
- Network traffic and congestion
- Data storage
- Access to database servers.
- Messaging system

Almedia et al [34] presented the workload characteristics of e-commerce servers with respect to different customer behaviour patterns. From studying this paper it can be concluded that customer behaviour pattern plays key role

in deciding workload of websites. It is found that when analysing the scalability of e-business sites, one has to consider the business, functional, customer behaviour, and IT resource aspects of the problem.

III. PROPOSED STEPS FOR PERFORMANCE ANALYSIS

In this paper workload of web sites are considered. Series of steps need to be followed to predict workload [35].

I. The initial step is to understand kind of hardware and software resources required by website, connectivity, protocols applied. This information is collected using meetings, audit, planning, interview and questionnaires. The author has studied different e-business websites like Amazon in order to gather information.

II. Workload characterization is done in order to identify workload of each component of system. Identifying intensity and service rate of each component is necessary to understand different parameters involved in workload. Available data and logs are used to characterize workload.

III. In the third step value of different parameters of workload are identified. It also involves monitoring the performance of web services from different reference points. Logs play key role in obtaining values. Study of available literature also helps in deriving different parameters involved.

IV. Good forecasting plan plays key role in website development as undesirable scenarios lowers profits of company. Hence number of cloud instances need should be accurately forecasted. Graphs generated with the help of Matlab neural toolbox helps in accurate forecasting of results.

V. In the fifth step, models are developed according to forecasted plan. Two types of models will be used. Analytical model stipulates the interactions between different components. Simulation model will mimic the system behaviour.

VI. In the final step performance will be validated based on results. Errors must be within specified limits.

VII. In the last step models will be analyse to predict the results. This step should indicate actions that must be taken so that Web services can meet the business goals set.

Based on above steps workload prediction model is presented in next session.

IV. WORKLOAD PREDICTION MODEL

Matlab neural network toolbox is used where input consist of 11 variables and output is of one variable. By applying different combinations different data samples are generated.

Let a be mean arrival rate of E-business application client.

Let s is the mean service time per request.

It is difficult to predict exact demand but approximate values can be predicted. Let ah be arrival rate during high load and al during low load.

At each instant, cloud instances need to be reserved. They are acquired from two markets on demand market and reservation market. Also price of reserved market instances are lower as compared to on demand market instances [36]. Hence total cost of cloud infrastructure is given by following formula:

$$C(nr)= cr(nr)+co(nr) \quad (1)$$

Where cr(nr) is cost of reserved market instances and co(nr) is cost of on demand market instances.

If any request takes more time than allocated time in that case there is no utility gain *ug*. Thus total utility gain of web applications are

$$ug(nr)=ugn(nr)+ugl(nr)+ugh(nr) \quad (2)$$

Where *ugn(nr)* is utility gain under normal load, ugl(nr) is utility gain under low load and *ugh(nr*) is utility gain under high load.

Another parameter is threshold TA. It is established during service level agreements. It tells the percentage of request successfully finished during the period the service is monitored. If the number of instances available to satisfy the request is less than threshold than penalty incurred. Hence the goal of this planning model is to maximize the value of nr. The inputs given to model based on literature survey are as follow:

- Available Instances (n)
- Service rate (s)
- Arrival rate (Normal a)
- Arrival rate (high demands ah)
- Arrival rate (low demands al)
- Amount of time (low load tl)
- Amount of time (high load th)
- Amount of time (normal load tn)
- Utility gain (ug)
- Threshold (TA)
- Unsuccessful request Penalty (p).

Applying different combinations generates data set:

- Changing only service time keeping other same. Only change the active Mean service time per request at a reference cloud instance, the others remain the same
- Changing load times keeping others constant.

- Changing utility gains for successful request, others are constant.
- Changing Penalty for unsuccessful requests and others remain the same.
- Change all the parameters.

Output of this model predicts number of cloud instances need to be reserved for given applications under certain set of parameters. Mean squared error is used as performance function to calculate network performance. Data is collected from available literature as discussed in literature review session.

In the paper trainlm method is applied for obtaining result. It gives best performance at epoch 28.

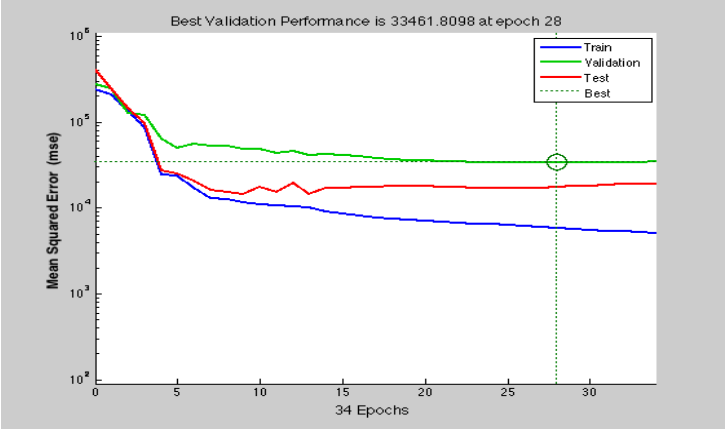

Figure 1. Performance graph

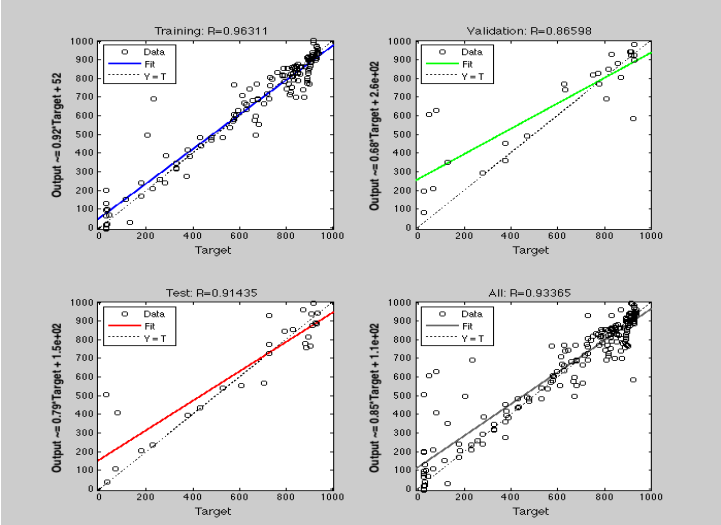

Figure 2. Regression graph

The Regression value is an indicator of the association between the outputs and targets. The value R = 1 indicates an exact linear association between expected and actual values. If R is close to zero, then there is no linear relationship between actual outputs and expected outputs [37]. This also indicated that there is error in selection of value. In figure 2 regression value of training is 0.96311 and that of testing and validation is 0.91435 and 0.86598 respectively. Overall value of regression coefficient comes out to be 0.93365 [38].

V. CONCLUSIONS

For above discussed case, the training data indicates a good fit. The performance graph is not indicating case of over fitting. As regression coefficient value is approximately 1, number of instances required for provided data is predicted accurately by algorithm. The model can be used as a base model to predict workload of unpredictable spikes of cloud applications. The steps proposed in this model can help in predicting workload of any application running on cloud.The problem with it is that algorithm is assuming homogenous environment. Also number of parameters selected will depend on type of applications. In this workload for websites are considered which is highly unpredictable workload.

REFERENCES

[1] James Bourne (2017). "IDC says global spending on public cloud services to hit $122.5bn in 2017", retrieved from http://www.cloudcomputing-news.net/news/2017/feb/21/idc-says-global-spending-public-cloud-services-hit-1225bn-2017/.

[2] Stamford, Conn. (2016). "Gartner Says by 2020 Cloud Shift Will Affect More Than $1 Trillion in IT Spending", retrieved from http://www.gartner.com/newsroom/id/3384720.

[3] Sahi S.K., Dhaka V.S. (2016). "A survey paper on workload prediction requirements of cloud computing":" Computing for Sustainable Global Development (INDIACom), 3rd International, Publisher IEEE.
[4] C. Fehling et al (2014). "Cloud Computing Fundamentals", Cloud Computing Patterns, 21, DOI 10.1007/978-3-7091-1568-8_2, Springer-Verlag Wien.
[5] The Open Group Platinum, (Dec 2015). "Maximizing the Value of Cloud for Small-Medium Enterprises", Applicable Workloads for Cloud.
[6] Judith Hurwitz, Robin Bloor, Marcia Kaufman, and Fern Halper, (Dec 2015). "How to Handle Workloads in Cloud Computing", Cloud Computing For Dummies.
[7] Carlos Gershenson (2001). "Artificial Neural Networks for Beginners", School of Cognitive and Computer Sciences, 9 pages.
[8] Christos Stergiou, Dimitrios Siganos (2015). "NEURAL NETWORKS", retrieved from http: //www.doc.ic.ac.uk /~nd/ surprise_96 / journal /vol4 /cs11 /report.html.
[9] Neural Network Toolbox, MathWorks (July 2015). Retrieved from http://in. mathworks.com/products/neural-network/.
[10] Venkateshwar Rao, Sarika Rao (2012). "Application of Artificial neural networks In Capacity planning of Cloud based IT Infrastructure", IEEE.
[11] S. Hackett (2008). Managed Services: An Industry Built on Trust, IDC.
[12] J.T. Hamill, R.F. Deckro, J.M.K. Jr. (2005). "Evaluating information assurance strategies", Decision Support Systems, pp. 463–484.
[13] P. Roehrig (2009). "New Market Pressures Will Drive Next-Generation IT Services Outsourcing", Forrester Research, Inc., 2009.
[14] J. Staten (2009). Hollow Out The MOOSE: Reducing Cost With Strategic Rightsourcing, Forrester Research, Inc., 2009.
[15] Q. Zhang, L. Cheng, R. Boutaba (2010). "Cloud computing: state-of-the-art and research challenges", Journal of Internet Services and Applications, pp. 7–18, http://dx.doi.org/10.1007/s13174-010-0007-6.
[16] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, Matei Zaharia (Apr, 2010). "A View of Cloud Computing", Communications of the ACM, Vol. 53 No. 4, pp. 50-58, doi: 10.1145/1721654.1721672.
[17] P. Noordhuis, M. Heijkoop, A. Lazovik (2010). "Mining twitter in the cloud: A case study, Cloud Computing (CLOUD)", Proceedings of IEEE 3rd International Conference on, IEEE, Miami, FL, pp. 107–114.
[18] L. Tucker (2009). Introduction to cloud computing for Enterprise Users, Sun Microsystems, Inc.
[19] Smith, D.M. (2012). "Hype cycle for cloud computing" from Technical Report, Gartner 2012, http://www. gartner.com/id1/42102116.
[20] Fehling, C., Ewald, T., Leymann, F., Pauly, M., Rutschlin, J., Schumm, D. (2012). "cloud computing knowledge and experience in patterns", proceedings of the 5th IEEE International Conference on Cloud Computing (CLOUD), Honolulu.
[21] Quiroz, A et al. (2009), "Towards autonomic workload provisioning for enterprise Grids and clouds" in Grid Computing, 10th IEEE/ACM International Conference. pp. 50-57, Banff, Alberta, Canada. October.
[22] Sivadon Chaisiri, Bu-Sung Lee and Dusit Niyato (2011). "Optimization of Resource Provisioning Cost in Cloud Computing", January 31 2011, DRAFT Digital Object Identifier 10.1109/TSC.2011.7 1939-1374/11, IEEE, Pages: 32.
[23] Ching-Chi Lin, Jan-Jan Wu, Pangfeng Liu, Jeng-An Lin, Li-Chung Song (2013). "Automatic Resource Scaling for Web Applications in the Cloud", J. Park et al. (Eds.): GPC 2013, LNCS 7861, pp. 81–90, Springer-Verlag Berlin Heidelberg.
[24] Rizwan Mian, Patrick Martin 2012). "Executing data-intensive workloads in a Cloud", 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing; 78-0-7695-4691-9/12, IEEE, DOI 10.1109/CCGrid.2012.18.
[25] Qian Zhu, Gagan Agrawal (2012). "Resource Provisioning with Budget Constraints for Adaptive Applications in Cloud Environments", IEEE Transactions on Services Computing, vol.5, no. 4, pp. 497-511, doi: 10.1109/TSC.2011.61.
[26] Chunlin Li, La Yuan Li. (2012). "Optimal resource provisioning for cloud computing", The Journal of Supercomputing, Vol. 62, Issue 2, pp. 989-1022.
[27] Amini Salehi M, Javadi B and Buyya R. (2013), "Resource Provisioning based on Pre-empting Virtual Machines in Distributed Systems", The Journal of Concurrency and Computation: Practice and Experience, Vol. 26, No. 2, pp. 412-433.
[28] H.N. Van, F.D. Tran, J.M. Menaud (2010). "Performance and power management for cloud infrastructures", IEEE 3rd International Conference on Cloud Computing, IEEE, pp. 329-336.
[29] Ahmed Ali-Eldin, Oleg Seleznjev, Sara Sjo stedt-de Luna, Johan Tordsson, Erik Elmroth (2014). "Measuring Cloud Workload Burstiness", IEEE/ACM 7th International Conference on Utility and Cloud Computing.
[30] Border, C. (2013). "Cloud Computing in the Curriculum: Fundamental and Enabling Technologies", Proceedings of The 44th ACM Technical Symposium on Computer Science Education, March 06 - 09, Denver, CO, USA.

[31] N. Roy, A. Dubey and A. Gokhale (2011), "Efficient Auto scaling in the Cloud using Predictive Models for Workload Forecasting," IEEE International Conference on Cloud Computing (CLOUD), Washington, DC, pp. 500 – 507.
[32] Y. Sun, Y.Chen and M. Chen (2013) "A Workload Analysis of Live Event Broadcast Service in Cloud", Procedia Computer Science, vol. 19, pp. 1028–1033.
[33] Raquel Lopes, Francisco Brasileiro, Paulo Ditarso Maciel Jr. (2010). "Business-Driven Capacity Planning of a Cloud-based IT Infrastructure for the Execution of Web Applications", IEEE International Symposium on Parallel & Distributed Processing, Workshops and PhD Forum (IPDPSW).
[34] Almeida, V. A. and Menascé, D. A. (2002). Capacity Planning: An Essential Tool for Managing Web Services. IT Professional 4, 4 (Jul. 2002), 33-38. DOI= http://dx.doi.org/10.1109/MITP.2002.1046642.
[35] Supreet Kaur Sahi, Dr. V.S.Dhaka (January 2015). "A Review on Workload Prediction of Cloud Services, International Journal of Computer Applications" 109(9): 1-4,
[36] ''Amazon Elastic Compute Cloud (EC2)'. Retrieved from http: //aws. amazon.com/ec2/, Jan 2016.
[37] Neural Network Toolbox, MathWorks (Nov 2015). Retrieved from http://in. mathworks.com /products/neural-network/.
[38] J. O. Rawlings, S. G. Pantula, D. A. Dickey (1998). "Applied Regression Analysis", New York, Springer-Verlag.

