EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

A Comparative study about Workload prediction from one time forecast with cyclic forecasts using ARIMA model for cloud environment

Yuvha Secaran R[1,*], Sathiyamoorthy E[2]

School Of Information Technology and Engineering, VIT University, Vellore, Tamilnadu, India

Abstract

Auto-scaling systems help provisioning resources on demand which helps tap into the elastic nature of the cloud. The applications hosted on the cloud tend to face workload surges which causes the response to be slow or denied. To tackle provisioning resources on demand there are reactive and proactive strategies in place. The topic of interest is the proactive strategies which uses a quantified metric as an input to provision resources before the demand arises. The quantified metric is the prediction obtained as a result of analysing the historical data of a application. This paper focuses using historical data of requests served by a web application to obtain a forecast value. The forecast value is the quantified metric which influences the scaling decisions. Conclusions are drawn about the accuracy of the metric based on prediction intervals along with the varied ways of forecast.

Keywords: Auto scaling, Time series, Workload prediction, environment, Energy, Cloud.

Received on 09 March 2020, accepted on 04 April 2020, published on 14 April 2020

doi: 10.4108/eai.13-7-2018.163977

*Corresponding author. Email: aspiringcoder9313@gmail.com

1. Introduction

Cloud computing transitioned the way organizations think about hosting services online. This transition has brought in several challenges in terms of robustness, reliability and security. Over a period of time since its inception, cloud computing has evolved to be a mature model. Facilitating administrations online means enlisting a pool of virtual assets from a cloud supplier dependent on an evaluating plan. Foundation as-an administration cloud suppliers offer examples normally under two valuing plans, in particular on-request where a cloud client will embrace a pay-more only as costs arise methodology and the held occasions where a client pays forthright for the necessary measure of occurrences.

Amazon came up with the idea of spot instances back in 2009 to sell their idle capacity resulting in a competitive advantage over other providers with a trade off between price and reliability. The auction based mechanism lets a cloud user to bid unused spot instance for an hourly price set by Amazon EC2. The pricing may fluctuate based upon the supply and demand for instances. Cloud user will participate in the bid by setting a bid price as part of the request. When the market price is lesser or equal to the bid price then the instance is made available to theuser. If the spot price goes beyond the bid price then the instances will be terminated. No matter what type of resources are used it will be scaled based on the requirement if an auto scaling system is in place. Scaling strategies work predominantly based on reactive or proactive behavior.Receptive methodologies center around provisioning assets after the flood in outstanding task at hand occurred while the

proactive systems handle the flood by provisioning assets before hand by knowing the interest through forecast procedures. Proactive methodologies rely entirely upon a measured measurement as contribution to scale assets. The quantified metric is the outcome of the prediction system after analysing the historical data. Current work focuses on analysing the quantified metric in terms of different prediction intervals and forecasts thereby drawing conclusion about the accuracy of the prediction.

2. Literature Survey

The literature specific to this study focuses on a two point perspective viz, auto scaling and workload prediction. Authors of [Chenhao Qu et al.2016] use spot instances with a fault tolerant model. The model primarily focuses on hosting web applications using a combination of both on-demand and spot instances with carefully defined scaling policies and fault tolerant semantics. The cost efficient reactive auto scaling strategy optimize the response time of the requests and the cost of the instances. [Roy N et al.2011] use a look-ahead optimization technique to ascertain a time interval for changing the resources based on the prediction. The auto scaling algorithm is proactive in nature which considers the cost factors that control the behavior of the system for scale up and scale down. Cost factors include SLA cost, cost of reconfiguration and cost to lease the machines. [Anshul Gandhi et al.2014] designed a modeling engine to characterize the workload and assess the scaling options. The scaling policy is proactive and the assessment is done with the help of the predictor component which does a medium term prediction outputs the workload periodically through a monitoring window. The result is that the scale up and the scale out combination gives the optimal scaling policy. [Anshuman Biswas et al.2015] allow user to request resources through the broker based architecture and bill the user for usage in terms of seconds as opposed to the cost per hour. The broker is responsible for creating the virtual private cloud with the resources leased from the public cloud. The predictive approach uses machine learning with a deadline constraint. The predicted 'K' request along with the five parameters as the characteristics of the request help acquire resources or extend the stop time for the resources. Efforts were focused to increase the broker profit and reduce the user cost. [Tania Lordio-Botran et al.2014] have reviewed the auto scaling systems and concluded that it is advisable to focus efforts towards developing proactive scaling systems. As an added note, authors have opined that the scaling systems should take advantage of time series analysis techniques especially the prediction capabilities.

Table 1. Table showing Actual,One time forecast and Cyclic forecast values

ACTUAL	*ONE TIME FORECAST*	*CYCLIC FORECAST*
10605152	10660567	10660567
9938548	9958829	9842539
9525577	9612000	9646089
9424458	9440583	9283901
9297976	9355861	9414700
8939574	9313987	9216971
8390395	9293292	8552662
7703478	9283063	7918568
7121185	9278008	7172993

According to [H. Shumway and David S. Stoffer2011] time series is the systematic approach to answer questions posed by the time correlations. The questions are mostly of mathematical and statistical in nature. Observations 'Xt' recorded at specific time 't' can be termed as a time series. RThe arrangement can be continous when the perceptions are made between interim [0,1] or it very well may be discrete if the perception made at fixed time interims. A period arrangement can be spoken to graphically with the irregular variables(Xt) as the vertical hub and the time scale as the even one. Plotting helps associating the qualities at the adjoining timeframes to remake some theoretical consistent time arrangement that may have created such discrete example. The time series could differ for each scenario and the fundamental visual characteristic distinguishing is the differing degree of smoothness. Auto scaling strategies use prediction to predict the workload at a future time slot. Authors of [Rodrigo N. Calheiros et al2015] focus on resource utilization and low QoS impact by associating a workload prediction mechanism to the auto scaling system. The prediction is done using the ARIMA model and achieves a accuracy of about 91 percent. [Yazhou Hu et al.2016] use time series to analyze the monitoring data and then use kalman filtering to predict the workload time series. The scale up or down is predicted by using a trigger strategy. The strategy uses pattern matching technique. The performance of the trigger strategy is better than the threshold approach. [Joseph Doyle et al 2016] focus on identifying a metric compute unit seconds (CUS) which specifies the execution time of the workload. Kalman filtering is used to predict CUS. [Wei Fang et al.2012] use ARIMA model to predict workload for proactive provisioning of resources for cloud applications. The CPU intensive applications were handled well by the system. [Samuel A. Ajila et al.2013] evaluates three machine learning techniques Support vector machines, Neural networks and Linear regression with SLA metrics like throughput and response time. A random workload pattern is employed for realistic simulation. The Support Vector machine scores better than the other two techniques. [Anshuman Biswas et al.2014] proposes a proactive auto scaling technique for resources where in the number of resources are adapted based upon the system load. Support vector machine scores better than the linear regression. ARIMA has been used for prediction because previous research have presented that web workloads tend to have strong correlation between them as specified by [G .Urdaneta et al.2009].

3. Existing System

The existing system in context uses historical information of the wikipedia website in order to obtain the forecast value which is used as the quantified metric[Kumar, N., & Kharel, R. (2018]. The effect of this metric on resource utilization and its impact on QoS is analysed. The system uses the wikipedia traces obtained from the wikimedia foundation. The request traces as well as the project wise count is open to access. Using this data as the historical information for the wikipedia application , ARIMA model is fit to it and prediction is obtained for the next one hour. The prediction obtained is along with the 80 and 95 percent confidence levels.

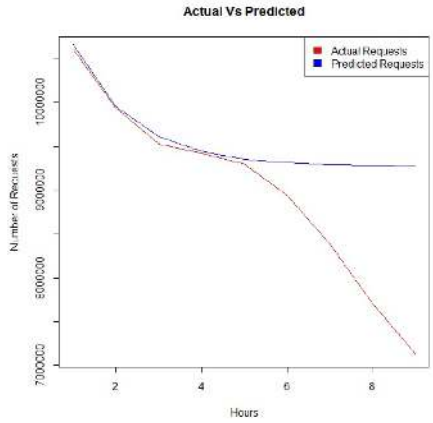


Figure 1. Actual VS One Time prediction

The existing system focus on analyzing the impact of using the lower and higher confidence levels as a trade-off between SL[A Memon, M. S., & Parveen, S. (2017)], Resource Utilization and response times. The higher edges of confidence intervals has a prediction accuracy of 78 percent at the high 95 percent but the intervals minimize the underestimations. The lower edges of the interval increases the underestimations but has a 85 percent prediction accuracy at low 95 percent. The lower edges should be used if the main priority is system utilization with minimal operational cost. Finally, for practical purposes 80 percent confidence interval is best suited [Shah, H. A., & Karn, N. K. (2017)]. The entire process of procuring the historical information, fitting the model and forecasting the next hour prediction are done entirely by statistical engine.

4. Proposed System

Table 2. Table showing MAPE scores for Cyclic and One Time Forecast

	Cyclic Forecast MAPE	*One Time Forecast MAPE*
3hr Window	0.91789889479695	0.54
6hr Window	1.43391599558	1.1026
9hr Window	1.5618991322454	7.5743

The proposed system uses the functionality of the afore mentioned statistical engine to obtain the forecasts in two manner.The authentic information for preparing the ARIMA model can't promptly for use. Subsequently as a pre-imperative the crude information from the wikimedia establishment is gotten and it is dependent upon investigation to acquire the quantity of solicitations for the english wikipedia assets. The preparation information for the model ranges across 3 weeks of January 2011 from 01 to 21. The test information is the real qualities for the quantity of solicitations saw on January 22 2011 from 0 to 9 hours. The request count with the training data period is transformed to a time series. Once the historical data is available for use with the statistical engine then the system uses it for forecast. The first forecast is obtained as a one time forecast for the next 10 hrs based on the historical data. The second forecast uses the cyclic behavior of the existing system where in the actual data available at the end of the hour is used for the next round of prediction.

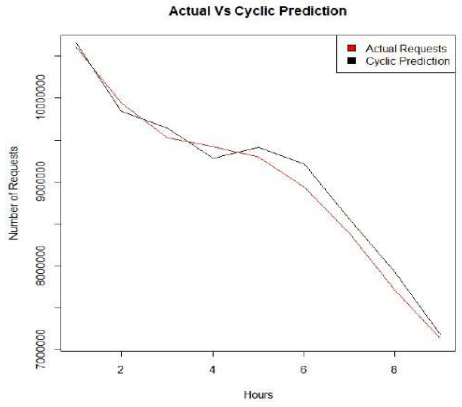


Figure 2. Actual VS Cyclic prediction

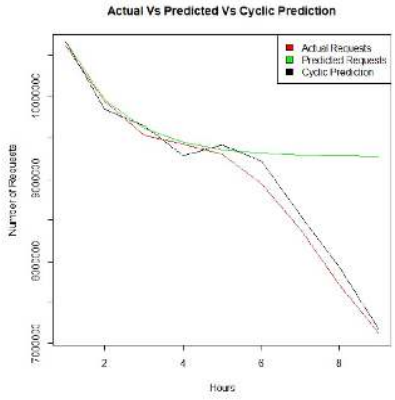
For the purpose of comparative study the entire act of reading the raw data, converting it to a time series, fitting ARIMA model to the data, obtaining the forecasts and returning the forecast estimates in a formatted manner are automated. This automation makes it easy to deduce the forecasts for our study purpose and also to draw conclusions out of it. The main aim is to split the forecasts

into hourly windows as multiples of three and to compare the behavior of one time forecasts and the cyclic forecasts. The forecast data is split into 3 , 6 and 9 hours respectively. The intention of the study is to evaluate the one time forecasts in relation to the cyclic forecasts. Mean Absolute Percentage Error (MAPE) is used as the metric to to better understand which one adapts when there is a change in actual value as well as which one has better accuracy across the hourly trend.

Mean Absolute Percentage Error (MAPE) is used to measure the prediction accuracy of forecast. Its widely used in statistics. The accuracy is expressed as a difference between the actual value and the forecast value divided by the actual value again. This value is again summed up for each forecast point and divided by the number of fitted points on which the forecast was based on. $A_t = Actualvalues\ F_t = Forecastvalues$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (1)$$

5. Results and Discussion



6

Figure 3. Actual VS OneTime VS Cyclic prediction

The one time forecast and the cyclic forecast are obtained for the next 9hrs which are then compared to the actual data as shown in (Table 1) MAPE score is determined to think about the precision of the two conjectures. For better comprehension of the conjecture conduct the MAPE score is determined for perceptions split across products of 3hrs windows like 3hour, 6 hour and 9 hour. The outcomes are outlined in (Table 2) for reference. The MAPE score of one time forecast for 3hour and 6 hour observations are slightly better than the cyclic forecasts. This is because by looking into the actual and forecasted value ,the one time forecast gives closer predictions for the first 5 hrs and starts to deviate from actual observations from the 6th hour. This is also evident in the (Figure 1).The MAPE score of Cyclic forecast for the 9 hour window is very much better than one time forecast which signifies the adapting nature of the forecast model to the hourly data that is getting reduced at each interval this is shown in (Figure 2).

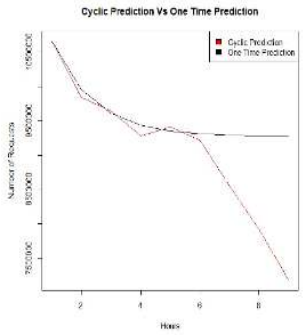


Figure 4. OneTime VS Cyclic prediction

When observing the actual data the number of requests starts to decrease each hour which is not factored in with the one time forecasts but the cyclic forecast which takes in the actual hourly input of requests is aware of the decrease in each hour and forecasts start to adapt accordingly as show in (Figure 3).(Figure 4) looks much similiar to the (Figure 1) showing that the one time forecasted values deviate from the cyclic predictions much early. Thus when dealing with systems requiring forecasts at regular intervals it is better to perform cyclic forecasts by updating the actual values for each hour to the forecast model and obtaining predictions based upon the actual data.

6. Conclusion and Future work

Workload prediction helps to make proactive decisions in an auto scaling system. Prediction involves analysing the historical information of a particular application and draw conclusions out of it using a prediction model which forecasts the expected value in the near future. The usage of the prediction can influence dynamic nature of workload there by streamlining proper resource usage and enable the system to scale without wastage of resources. Dynamic workloads affect factors like resource utilization and QoS. In order to better understand the prediction, our proposed system splits the prediction across hourly windows and analyses the accuracy of the same. The accuracy is compared for both one time and cyclic forecasts in order to better understand which one aligns best with the actual data. Analysis shows that the cyclic forecasts in the long run better align to the actual data at the end of each regular interval. The findings from the compartive study will be used by proactive auto scaling system. Autoscaling systems helps adjusting the resources when there is a workload surge.

The proactive strategy uses predictive mechanisms to estimate the future workload and streamlines the surge by provisioning resource requirements. Spot instances have been a low cost alternative and the dynamic changes in the spot market based on the demand supply have forced autoscaling systems to be smarter. This means the scaling system must consider varied factors like calculating optimal bid price, choosing amongst heterogeneous resources and placing request in the appropriate time. The future work is to tackle such factors associating the proactive scaling strategy with spot instances.

References

[1] Shumway and David S. Stoffer2011 Robert H. Shumway and David S. Stoffer "Time Series Analysis and Its Applications with R Examples"; *Springer Texts in Statistics*, Third Edition(2011)

[2] Roy N et al.2011 Roy.N, Dubey.A, Gokhale.A. (2011)'Efficient autoscaling in the cloud using predictive models for workload forecasting', *Proc. IEEE 4th International Conference on Cloud Computing(CLOUD)*, pp.500-507.

[3] Chenhao Qu et al.2016 Chenhao Qu, Rodrigo.N.Calheiros, Rajkumar Buyya (2016) 'A Reliable and cost-efficient auto-scaling system for web applications using heterogeneous spot instances', *Journal of networking and Computer Applications*, Vol. 65, pp.167–180.

[4] Anshul Gandhi et al.2014 Anshul Gandhi, Parijat Dube, Alexei Karve, Andrzej Kochut, Li Zhang (2014) 'Modeling the impact of workload on Cloud Resource Scaling' *Proc IEEE 26th International Symposium on Computer Architecture and High Performance Computing*, pp.310-317.

[5] Yazhou Hu et al.2016 Yazhou Hu, Bo Deng, Fuyang Peng, Dongxia Wang (2016) 'Workload prediction for cloud computing elasticity mechanism' *Proc IEEE International Conference on Cloud Computing and Big Data Analysis*, pp.244–249.

[6] Samuel A. Ajila et al.2013 Samuel A. Ajila, Akindele A. Bankole (2013) 'Cloud client prediction models using machine learning techniques' *Proc IEEE 37th Annual Computer Software and Applications Conference*, pp.134–142.

[7] Rodrigo N. Calheiros et al2015 Rodrigo N. Calheiros, Enayat Masoumi, Rajiv Ranjan, and Rajkumar Buyya (2015) 'Workload Prediction Using ARIMA Model and its Impact on Cloud Applications QoS' *IEEE Transactions on Cloud Computing*, Vol. 3, No. 4, pp.449–458.

[8] Joseph Doyle et al2016 Joseph Doyle, Vasileios Giotsas, Mohammad Ashraful Anam and Yiannis Andreopoulos 2016 'Cloud instance management and resource prediction for Computation-as-a-Service platforms' *Proc IEEE International Conference on Cloud Engineering*, pp.89–98.

[9] Wei Fang et al.2012 Wei Fang, ZhiHui Lu, Jie Wu, ZhenYin Cao (2012) 'RPPS: A Novel Resource Prediction and Provsioning scheme in cloud data center' *Proc IEEE Ninth International Conference on Services Computing*, pp.609–616.

[10] Tania Lordio-Botran et al.2014 Tania Lorido-Botran, Jose Miguel-Alonso, Jose A. Lozano (2014) 'A Review of auto-scaling techniques for elastic applications in Cloud Environments'*Journal of Grid Computing*, Vol.12, pp.559-592.

[11] Anshuman Biswas et al.2014 Anshuman Biswas, Shikharesh Majumdar, Biswajit Nandy Ali El-Haraki (2014) 'Automatic resource provisioning: A Machine learning based proactive approach' *Proc 6th IEEE International Conference on Cloud Computing Technology and Science*, pp.168–173.

[12] Anshuman Biswas et al.2015 Anshuman Biswas, Shikharesh Majumdar, Biswajit Nandy Ali El-Haraki (2015) 'Predictive auto scaling techniques for clouds subjected to requests with SLA's' *Proc IEEE World Congress on Services*, pp.311-318.

[13] G.Urdaneta et al.2009 G. Urdaneta, G. Pierre, and M. van Steen (2009) 'Wikipedia workload analysis for decentralized hosting,' *Comput. Netw.*, Vol. 53, no. 11, pp. 1830-1845.