International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 5 (2020) pp. 532-536

Cloud Long Term Load Forecasting for Scaling of Compute Resources

Abhishek Awasthi[1] and Pratiba D[2]

[1,2]Department of Computer Science and Engineering,
R.V College of Engineering, Bangalore,
Karnataka – 560059, India

Abstract

This paper provides a Comparative study of various machine learning and statistical models for timeseries modelling in order to forecast future loads by an application-cluster on the cloud infrastructure based on virtual machines (VM). The work done in this paper emphasizes on data preparation in detail discussing about various parameters for long term forecasting. It also addresses the research gap between long-term forecasting and existing literature which is more focused on short-term forecasting restricted to VM level. It also explains the design of a real-time monitoring and forecasting system to automate the process of dynamic resource allocation based on forecasted demand.

Keywords: Time series analysis, machine learning, virtual machines, Cloud infrastructure, automation, dynamic resource allocation.

I. INTRODUCTION

One of the most important reason why cloud based infrastructures are getting so widely accepted is due to the high flexibility in terms of scaling resources as per demand. Setting up of new compute resources requires some overhead time which can lead to poor quality of service for cloud's clients .Hence for a smooth and cost effective scaling ,resources must be adjusted to demand ahead of time. To achieve this dynamic nature of cloud capacity, cloud load forecasting needs to be done.

Cloud providers need to manage both the quality of experience assured to the clients as well as reduce their own costs. In order to minimize the difference between reserved resources and utilization values, efficient time series forecasting needs to be developed for predicting future resource demands. This paper compares various statistical and machine learning models for long term forecasting.

A. Time Series Analysis

Time series can be defined as the collection of data points of one or more fields separated by regular time intervals. Example: amount of rain at a given location on a monthly basis.

The time series data can be classified into three types:

1. *Time series***:** collection of single data variable indexed using timestamp, pair of consecutive entries may or may not be separated by regular time interval.
2. *Cross-sectional time series data:* collection of two or more observed fields for a single timestamp.
3. *Pooled time series data***:** two or more time series collected at same timestamps.

The models which use target variable as the only field are called univariate models and the models which use one or more fields other than target variable are called multivariate models.

Time series consists of 3 components which are trend, seasonal and cyclic components. The *seasonal (S)* component of a time series is defined as the pattern observed due to weekly, monthly or annual cycles. A *cycle (C)* can be defined as the short-lived patterns in data which don't have a fixed time period. A *trend (T)* can defined as the long-term slope of the observed metric, it can be positive or negative. This relationship can be mathematically expressed as eqn. 1.

$$Y(t) = T(t) + S(t) + C(t) \text{ Eqn.} \qquad (1)$$

The load forecasting is done using various machine learning and statistical models which use time series analysis. Time series modelling is done using previously recorded time series data for single or multiple parameters.

B. Cloud infrastructure

The cloud infrastructure used in this study is based on virtual machines and is show in the figure (1) above. The highest level of operational level in this hierarchy is Data-Center (DC) level. DC level houses all the VMs related to individual cluster applications. But each cluster itself is localized to geographical locations grouped by DCs. For example if we had a cluster A then its cluster name in India will be IND_A while US_A in America.

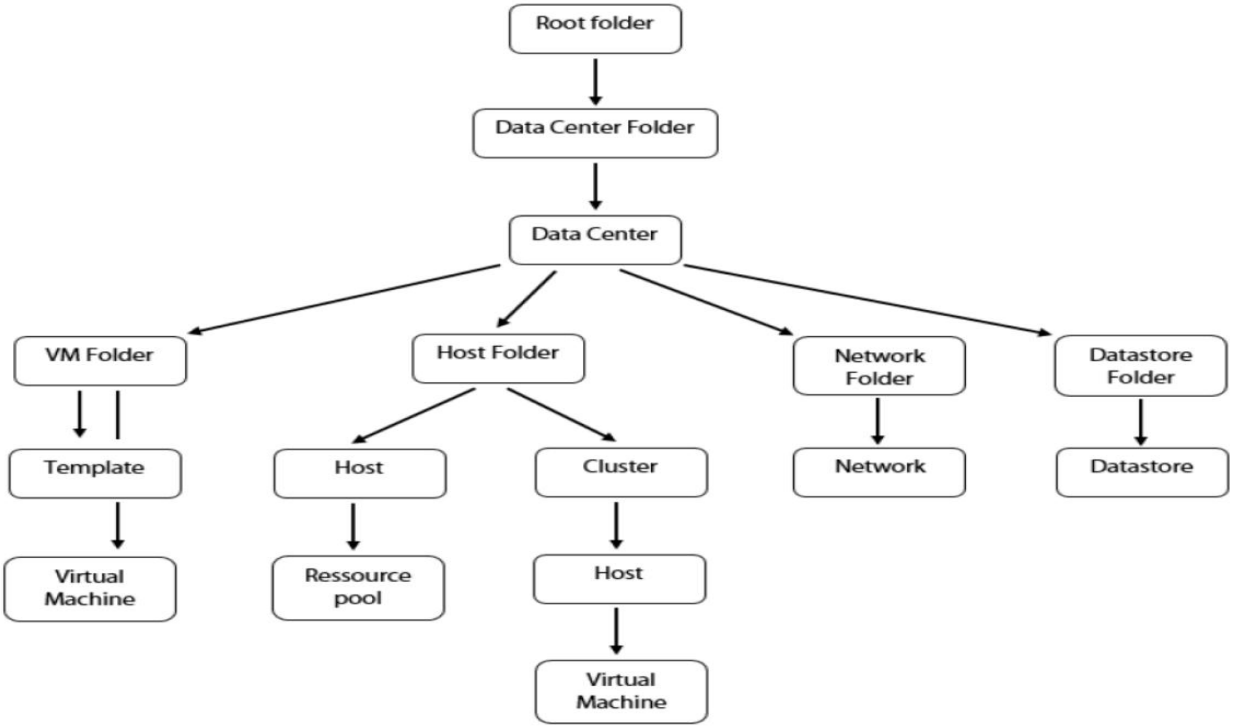

Fig -1: VM based cloud infrastructure

Each localized cluster consists of multiple ESXi hosts, and each host machine houses multiple VMs belonging to same cluster. All the disk requirements of a cluster are satisfied using a Datastore unit which is specialized for disk I/O operations with reduced read/write latencies. All the network related metrics are stored in Network level Node in a DC.

C. Research Gap

Extensive research has been done for short term forecasting (i.e. 10 to 15 minutes) for operations level optimizations by instantiation of new VMs as per requirements in advance. This study deals with quarterly forecasting which helps in efficient scaling by suggesting clients with expected cloud level usage predictions.

The algorithms applied for short term VM utilization predictions are not effective as they learn short lived cyclic patterns in data which are absent in long term data. The trends shown at cluster level are hugely contributed by business level strategies.

D. Dataset

Dataset used for this analysis is collected from various hierarchical levels of the cloud infrastructure discussed in the previous section. All the fields of the accumulated dataset are listed in the table -1.

This study is restricted to prediction of CPU utilization (MHz), Main memory utilization (GB) and disk usage (IOPS/GB) for a timestamp of up to 3 months into the future.

Table -I. Time series Dataset fields

Field Name	***Meaning***
CPU usage	CPU utilization in MHz
Mem usage	Memory utilization in GBs
Disk usage	Combined disk usage in GBs
CPU contention	Total Outstanding CPU requests
Mem contention	Total Outstanding memory requests
Mem balloon	Loaned mem from hosts
IOPS	Total disk IOPS from datastore
N/w latency	Network latency in seconds
OIOs	Outstanding IOPS for datastore
N/W usage rate	Avg transfer over N/W in KBs

II. RELATED WORK

Existing research has focused on short term time series load forecasting for VMs using machine learning techniques such as SVR and variations of RNN discussed in this section.

In [1], Proposed the usage of request time intervals along with virtual machine work load with request time interval data. The study also proposed the usage of a modified LSTM model known as N-LSTM [2] for machine workload forecasting for short-term predictions.

Nested LSTM [2] is an RNN architecture with multiple layers of memory unit cells. Nested LSTMs add depth to LSTMs via nesting as opposed to stacking. Nested LSTMs as opposed to stacking uses addition of depth to LSTMs via nesting to learn long term dependencies. Nested LSTMs outperform both stacked and single-layer LSTMs with exactly the same number of learned weights in the studies on various language modelling tasks, and the densely nested cells learn longer term dependencies compared with the higher-level units of a stacked LSTM layers.

Work done in [3] discusses the use of tensor canonical polyadic decomposition (CPD) [4] to reduce the amount of training time required for a deep learning model. The tensor CPD uses stacked autoencoder to reduce the input dimensionality of vectors which in turn reduce the number of weights to be learned in training.

In [5], LSTM is used to improve long-term pattern understanding in the data in order to increase the accuracy of long term load forecasts. Hourly resolution in load forecasts not only improves long term predictions but the resolution of predictions made on hourly basis . this study was done on England's real time electricity usage metrics.

In [6], a comparative study across three forecasting models are built using Linear Regressor (LR), Support Vector Regressor (SVR) and Recurrent Neural Network (RNN) has been done. This work proposes to include various SLA metrics rather than utilizing only single metric predictions with the help of utilization values such as memory , CPU and disk. This inclusion of service level parameters along with utilization provides a wider understanding of quality of experience.

Work done in [7] provides an analysis of effectiveness of time series modelling with ARIMA [8] model. The work addresses the hyper parameter tuning (p, q and d) specifically for long term predictions . The trends in stock price compare to our use case as stock prices are very random but long term trends are driven by high level factor such as market assessment.

Work done in [8] is discusses about open source algorithm from Facebook on Prophet algorithm to make effective time series predictions along with scaling strategies. Prophet algorithms is optimised for multiple regressors or for multivariate forecasting.

In [9] setting up an VM based cloud infrastructure for Infrastructure-as-a-Service (IaaS) which is based on various hierarchical levels described in section 1.2 . it allows for efficient scaling infrastructure.

III. METHODOLOGY

A. Data Preparation

As discussed in the introduction section that there is a 3 level hierarchy in the used cloud infrastructure. The data metrics are collected from individual VMs ,these collected metrics need to be rolled up to cluster level through host level. This rollup needs to be done using roll up queries on a Time Series database (TSDB).

Data collection at the VM is done in very short intervals (i.e. 20s), if a model is trained with data of such high granularity then it leads to very slow training with comparatively less performance improvement. Also effect of outliers is reduced by aggregation over larger intervals (i.e. 1 day). The original dataset consists of 819M rows across all cluster VMs which was down sampled to 189k records.

In the case of short term forecasting state of the art efficiencies can be achieved by just studying the utilization metrics of past without taking into the account for the *cause* that had the *effect* in utilization values. But in the case of long term predictions the information coming from sales or business side becomes key factor to explain data trends (for example, an abrupt surge in utilization can be explained if there is a new feature release for an application). Therefore, columns such as "*functionality_release"* is true if there has been functionality addition in past 4 weeks."*client_count"* parameter keeps the count of number of current registered user for the particular application/cluster, etc.

Columns such as "*week_day*" were added to the data set and are set to true if it's a week day and false otherwise, keeping account for usage patterns on weekends . Similarly track of holidays is kept using and additional column "*holiday*" was added, which is true for all national holidays based on geographical location of DC .

Missing values in the dataset was imputed using mean aggregation functions. All the numerical values were normalized between a range of -1 to 1 .For baseline modelling the dataset was split into train and test datasets in a 70:30 ratio respectively.

B. Real time prediction system

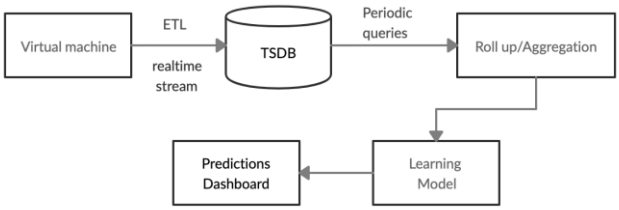

Fig -2: Real time prediction pipeline

Real time prediction pipeline starts with collection of real-time utilization metrics from each VM using an ETL (extract, transform and load) script, which gets the collected data stored in a TSDB. As discussed in section 3.1 the data collected at VM level needs to be rolled up to cluster level for application level analysis.

After having the rolled-up data at cluster level, data aggregation is performed to reduce the granularity of data which is not only helpful for faster learning but also the amount of storage space needed to store old data is reduced significantly. The roll up and aggregation needs to be performed periodically with a frequency same as the aggregation interval, these jobs are scheduled using CRON scheduler.

Aggregated data is used to tune the already existing baseline

models to new data trends. This update of in model weights is done in an online learning manner (i.e. incremental updates). Once the retuning of model is complete for current timestamps data, model is used to make predictions for the next quarter. Predictions for the next quarter are stored in a TSDB and same is visualized on monitoring platform.

IV. RESULTS

The data set was split into training and test datasets using a 70:30 ratio. Below tables show accuracies for quarterly predictions for various time series modelling algorithms.

Fig-3 to 5 show learned vs actual data , black dotted plot indicates actual utilization metrics whereas the light blue line indicates prediction curve modelled by our algorithm.

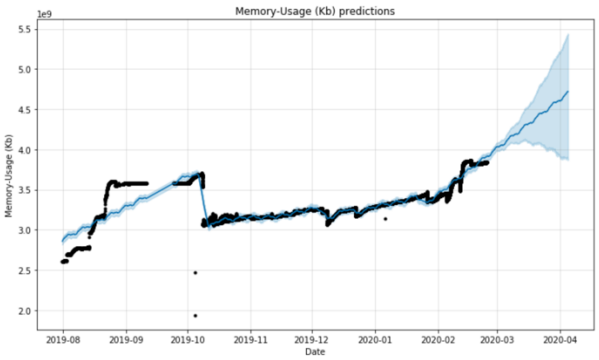

Fig -3: Memory usage model

Table -2: Main Memory usage Prediction

index	*Algorithm*	*Training (%)*	*Test (%)*
1	SVR	81.0	76.1
2	LSTM-RNN	92.1	89.4
3	ARIMA	92.7	91.8
4	Prophet	95.4	93.6

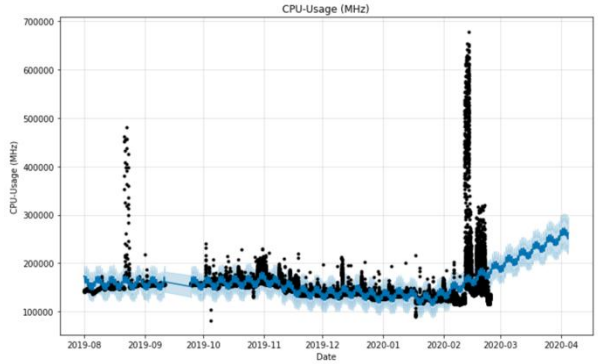

Fig -4: CPU usage model

Table -3: CPU usage Prediction

index	*Algorithm*	*Training (%)*	*Test (%)*
1	SVR	68.0	63.2
2	LSTM-RNN	72.4	72.1
3	ARIMA	86.1	79.4
4	Prophet	82.4	78.3

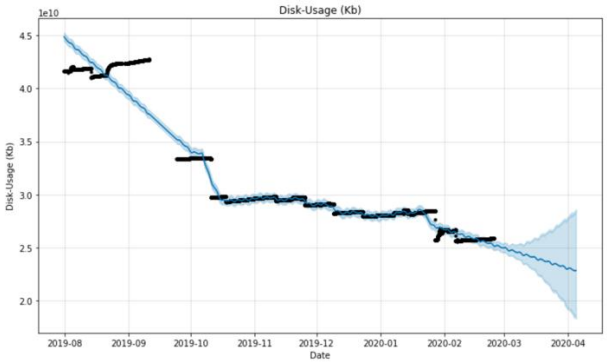

Fig -5: Disk usage model

Table -4: Disk usage Prediction

index	*Algorithm*	*Training (%)*	*Test (%)*
1	SVR	84.1	82.1
2	LSTM-RNN	87.8	87.0
3	ARIMA	98.3	93.4
4	Prophet	97.2	95.2

It can be observed from the actual utilization metrics there is no smooth trend associated with them. But the abrupt changes can be learned with the parameters described in section III-A which are not intrinsic to utilization metrics.

V. CONCLUSION

We can see from the results in the previous section that ARIMA and Prophet fare comparatively better than SVR and LSTM-RNN. LSTM-RNN are state of the are time series prediction models but they don't perform well when there is an absence of short-lived cyclic patterns While models like ARIMA and prophet are statistical models perform better when the data trends are random. RNN based architectures don't adapt to sudden changes in data trends quickly.

.

REFERENCES

[1] W. Guo, W. Ge, X. Lu and H. Li, "Short-Term Load Forecasting of Virtual Machines Based on Improved Neural Network," in IEEE Access, vol. 7, pp. 121037-121045, 2019.

[2] Q. Zhang, L. T. Yang, Z. Yan, Z. Chen and P. Li, "An Efficient Deep Learning Model to Predict Cloud

Workload for Industry Informatics," in IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3170-3178, July 2018.

[3] R. K. Agrawal, F. Muchahary and M. M. Tripathi, "Long term load forecasting with hourly predictions based on long-short-term-memory networks," 2018 IEEE Texas Power and Energy Conference (TPEC), College Station, TX, 2018, pp. 1-6.

[4] Samuel A. Ajila and Akindele A. Bankole ,'Using Machine Learning Algorithms for Cloud Client Prediction Models in a Web VM Resource Provisioning Environment' ,TMLAI Vol 1,issue 4 issn 2054-7390,2016

[5] B. Loganayagi and S. Sujatha, "Creating virtual platform for cloud computing," 2010 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2010, pp. 1-4.

[6] D. Alsadie, Z. Tari and E. J. Alzahrani, "Online VM Consolidation in Cloud Environments," 2019 IEEE 12th International Conference on Cloud Computing (CLOUD), Milan, Italy, 2019, pp. 137-145.

[7] A. Phan, A. Cichocki, I. Oseledets, G. G. Calvi, S. Ahmadi-Asl and D. P. Mandic, "Tensor Networks for Latent Variable Analysis: Higher Order Canonical Polyadic Decomposition," in IEEE Transactions on Neural Networks and Learning Systems

[8] Sean J. Taylor and Benjamin Letham "Prophet : forecasting at scale" https://peerj.com/preprints/3190.pdf

[9] Chenghao Liu, Steven C. H. Hoi, Peilin Zhao, Jianling Sun1 "Online ARIMA Algorithms for Time Series Prediction" Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16),2016

[10] Guosheng, H., Hu, L., Li, H., Li, K., and Liu, W., "Grid Resources Prediction with Support Vector Regression and Particle Swarm Optimization," 3rd International Joint Conference on Computational Science and Optimization (CSO), vol.1, pp.417-422, China. May, 2010.

[11] Bankole A., and Ajila S.A., "Cloud Client Prediction Models for Cloud Resource Provisioning in a Multitier Web Application Environment", in 7th IEEE International Symposium on Service-Oriented System Engineering (IEEESOSE 2013), San Francisco Bay, USA March 25 – 28, 2013.

[12] Bertholon, B., Varrette, S., Bouvry, P., "Certicloud: A Novel TPM-based Approach to Ensure Cloud IaaS Security". IEEE International Conference on Cloud Computing (CLOUD). pp. 121–130. 2011.

[13] Boniface, M. et al., "Platform-as-a-Service Architecture for Real-Time Quality of Service Management in Clouds". 5th International Conference on Internet and Web Applications and Services (ICIW). pp. 155–160, Barcelona, Spain. May, 2010.

[14] Borgetto, D., Maurer, M., Da-Costa, G., Pierson, J., and Brandic, I., "Energy-Efficient and SLA-Aware Management of IaaS clouds". 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy). pp. 1–10. Madrid, Spain. May, 2012.

[15] Cain , H. W., Rajwar, R., Marden, M., and Lipasti, M., "An Architectural Evaluation of Java TPC-W" in Proceedings of the Seventh International Symposium on High- Performance Computer Architecture, Nuevo Leone, Mexico. January, 2001