IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 1, August 2023

Predicting Cloud Resource Provisioning using Machine Learning Techniques

Prasad Madhukarrao Nakhate[1], Vrushabh Vinod Sasane[2] , Vedant Rajendra Sanap[3], Siddhi Ramchandra Kore[4]

UG Students, Department of Computer Engineering[1,2,3,4]
Sinhgad College of Engineering, Pune, Maharashtra, India

Abstract: *It can be difficult to provision assets in a cloud environment. Both inadequate and excessive provisioning are bad for the system's overall performance. Researchers are increasingly turning to proactive provisioning techniques that foresee resource needs in advance and set the system up to accommodate such real-time demands. Although proactive provisioning methods are more complicated than reactive provisioning strategies, they offer a generally faster reaction time since decisions about resource provisioning are made in advance of the actual requirement for those resources. The adoption of an analytical framework that anticipates resource requirements is necessary for these proactive provisioning strategies to function well. The provisioning of resources for computational operations in cloud computing is a significant concern. Unsurprisingly, previous research has identified a number of ways to deliver Cloud resources effectively. However, a forecast of the impending computational jobs' future resource consumption is necessary to implement a comprehensive resource provisioning model is required. However, the subject of cloud prediction the use of resources is still in its early stages.*

Keywords: Proactive provisioning, Resource prediction, Prediction model, Machine learning, KNN, SVM, RF.

I. INTRODUCTION

When compared to reactive provisioning methods, proactive resource provisioning strategies offer faster reaction times, which lowers the system's overall cost. Such methods are more difficult, though, because a prediction model that can forecast future resource needs must be incorporated into the system.So having an effective load prediction model is the first step in proposing an efficient proactive provisioning solution. In the context of resource requirements in a cloud environment, the goal is to forecast the anticipated future workload, providing the system enough time to make the necessary arrangements in time, resulting in little to no delay in provisioning resources at the moment of actual necessity. Researchers have different ideas about what constitutes a "workload" for an application run in the cloud. Application request volume has been used as a measure of workload. The workload has been employed in resource utilization. Consider reaction time as well as CPU utilization when determining the workload estimation parameter.

Define future VM demand as the workload, whereas throughput and response time have been used as the workload for performance prediction. In this study, we employed five machine learning methods to estimate the value of the prediction parameter, which was CPU utilization. The machine learning algorithms create a model using previous data, train it, and then use the model to predict the future. These methods frequently represent the behaviour as a time series. The public cloud server Parallel Workloads Archive has provided the server log used for the experiment. WEKA 3.8 was used to conduct each experiment. Each log has been assessed using K-Nearest Neighbours, Support Vector Machine, and Random Forest machine learning approaches.

Large computations can now be completed more quickly than before thanks to advances in cloud computing. Large scientific process applications can now be run dispersed in the cloud as a result. A workflow is an application that has been divided into pieces and is made up of a number of computing activities linked together by dependencies on data and control flow. However, customers will be charged in accordance with their usage while using the computer power made now available through the cloud. A cloud customer frequently has a set spending limit or a limit they must not go over while yet attempting to complete computations or workflows as quickly as they can.

Scheduling tasks on certain accessible instances is how workflow applications are run in the cloud. There are two components to the issue: Resource provisioning involves choosing the instances to employ for computation and scheduling jobs for the selected instances. Computing power, price, and other factors vary between instances; some are better for memory, storage, or visual processing than others. The quickest execution time (makes pan) of a workflow within a given budget is an NP-hard issue, making it impractical to find a large-scale size of the problem through an exhaustive search. Heuristic or meta-heuristic algorithms are frequently used to solve problems.

1.1 OBJECTIVES

1. To research machine learning as a different approach to solving the resource provisioning issue for the issue of running scientific workflows on the cloud.
2. To determine and assess whether a trained machine learning model can provide solutions with an equivalent level of quality but much more quickly than an algorithm.
3. To choose (using machine learning) which resources to supply so that the algorithm can be fed the information.
4. To complete resource provisioning (using machine learning in combination with a scheduling algorithm) in order to more quickly identify effective scheduling options for scientific workflows while also preserving a good make span for the process subject to the specified budget.

1.2 PURPOSE

Investigating machine learning as a means of resource provisioning for running scientific workflows on the cloud is the key goal. Faster access to provisioning and scheduling tools would save time for scientists managing complex processes by allowing them to schedule their workflows more efficiently.Another goal would be to motivate academics to continue investigating how machine learning is used in the fields of resource provisioning and resource scheduling for cloud-based workflow execution.

II. LITERATURE REVIEW

1.TajwarMehmood, Dr. Seemab Latif et.al. Utilising resources effectively enables cloud providers to deliver great performance at minimal cost. A dynamic environment called cloud computing offers pay-as-you-go on-demand services over the internet. Because it is shared by many users. The resource allocator distributes resources from a limited pool to meet changing user demands. There shouldn't be excessive or insufficient resource provisioning. Overusing a resource can result in service degradation, while underusing it can waste money and resources. Resource allocators can make resource provisioning decisions more effectively if they can predict future resource demand. To help the resource allocator provide the best possible resource provisioning, a resource utilisation prediction technique is needed.

2. Padma D. Adane, O. G. Kakde et.al. Large computations, such as scientific workflow applications, can now be processed more quickly than ever before because to cloud computing's rapid improvements. In the cloud, workflow applications are executed by selecting instances and then planning the tasks to run on those instances (resource scheduling). It is usual to utilise heuristics or metaheuristics to address the NP-hard problem of determining the fastest execution time (makespan) of a scientific workflow within the certain budget.In order to solve the issue of running scientific workflows in the cloud, it is possible to use machine learning as an alternate method of resource provisioning. To find out more about this, it is assessed whether a trained machine learning model can forecast provisioning instances with solution quality that is comparable to a cutting-edge method (PACSA), but in a noticeably shorter amount of time. The PACSA technique provides solution instances that are used as labels for the machine learning models that are developed for the scientific processes Cybershake and Montage. An separate HEFT scheduler is used to plan the anticipated provisioning instances in order to obtain a makespan.

3. Samuel A. Ajila Akindele A. Bankoleet.el. By anticipating future resource demands a few minutes in advance due to Virtual Machine (VM) boot time, one can proactively provide resources and adhere to Service Level Agreements (SLA). In this study, we created and assessed cloud client prediction models for a benchmark online service using Support Vector Machine (SVM), Neural Networks (NN), and Linear Regression (LR) machine learning approaches. To

DOI: 10.48175/IJARSCT-12470

give the customer a more reliable scaling selection option, we have included two SLA metrics: Response Time and Throughput. To improve upon our earlier work, we applied our model to the Amazon EC2 public cloud infrastructure. We have increased the experimentation period by nearly 100%.

Finally, in order to reflect a more accurate simulation, we used a random workload pattern.

4. M. H. Hilman, M. A. Rodriguez et.al. The thesis topic relates to the use of machine learning in the areas of resource scheduling and provisioning for scientific workflow execution in the cloud. The resource provisioning problem is a component of the resource scheduling problem, and research frequently offers solutions to the two problems in tandem as a fix for the scheduling issue. In regard to the subject of this thesis, the earlier and similar works discovered throughout the literature review divided into 3 categories.

a. Currently used techniques for cloud-based scheduling of scientific procedures

b. Cloud-based execution of scientific workflows using machine learning

c. Resource provisioning for cloud service providers using machine learning approaches

5. Akindele A. Bankole Samuel A. Ajila et.al. Due to the VM boot-up time, Virtual Machine resources must be provisioned a few minutes in advance in order to meet Service Level Agreement requirements. Making predictions about future resource demands is one approach to do this. In this study, Support Vector Machine, Neural Networks, and Linear Regression are 3 machine learning techniques that we used to create and assess cloud client prediction models for the TPCW benchmark web application. In order to give the client a more reliable scaling decision option, we added the SLA metrics for Response Time and Throughput to the prediction model. Our findings demonstrate that the best prediction model is provided by the support vector machine.

III. MACHINE LEARNING APPROACH

We suggest using machine learning techniques to forecast how resources will be used in the Cloud. The reasoning behind this is that using straightforward linear regression models is insufficient because tasks carried out on cloud-based computational resources frequently do not scale linearly with the cardinality of their input. Additionally, a task can require a vector of input data rather than a single item, necessitating the use of multiple linear regression. For this reason, we suggest using ML to extract a model for generating future predictions from historic data, i.e., past task executions. Online and offline learning are differentiated by machine learning: In contrast to online learning, which presents problem examples one at a time, offline learning happens when all instances are shown at once. In our case, we choose offline learning because training the KNN,RF,SVM model requires processing a lot of data, which uses a lot of computing resources. Furthermore, we employ supervised learning, which assumes that the correct output is known for each instance of training data and may be applied to back-propagation.

IV. ALGORITHMS

K- Nearest Neighbours (KNN): The simplest machine learning algorithm, which can be applied to both classification and regression, is this one. The forecast is based on mean of the K-most comparable occurrences when used for regression. It directly uses the initial data set to create predictions. It employs a distance measure to find the k examples that are similar to a new input. KNN is classified as a lazy set of classifiers tab in Weka under the name IBk (Instance based k). By selecting various values for k and distance measures, the accuracy of the model may be adjusted. Because there are a lot of occurrences in the log, we chose k to be 7 for this experiment, and Manhattan distance was chosen for the distance function because the characteristics have different measures.

Support Vector Machine (SVM): Support-vector machines (also known as support-vector networks or SVMs) are supervised learning models with corresponding learning algorithms that examine data used for regression and classification analyses. An SVM training algorithm creates a model from a set of training examples that are each marked as belonging to one of two categories, making it a non-probabilistic binary linear classifier (although there are ways to use SVM in a probabilistic classification setting, like Platt scaling).An SVM model is a mapping of the examples as points in space with as much space between the examples of the various categories as possible. Then, based on the side of gap on which they fall, new samples are projected into that same area and predicted to belong to category.

DOI: 10.48175/IJARSCT-12470

Two types Of SVM:

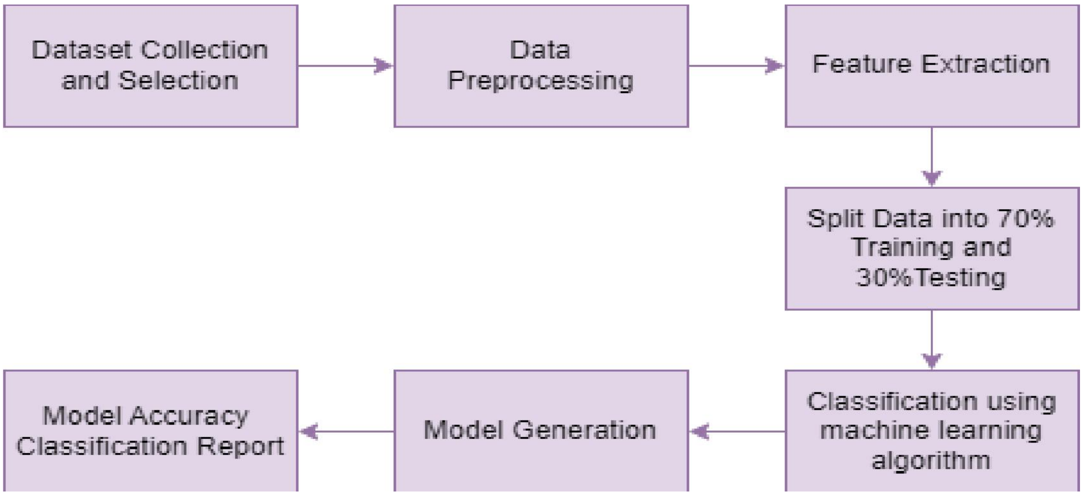
1. Linear SVM: The term "linearly separable data" refers to data that can be divided into two groups using only a single straight line. Linear support vector machine is used to classify such data, and the classifier utilized is known as the linear support vector machine classifier.
2. Non-Linear SVM: When a dataset cannot be classified using a straight line, it is said to have been non-linearly separated, and the classifier employed is known as a non-linear support vector machine classifier.

Random Forest (RF): A well-known machine learning algorithm Random Forest is a technique used in the supervised learning process. It can be used to solve machine learning problems involving both classification and regression. It is based on the concept of ensemble learning, which is a method of combining several classifiers to address tough difficulties and improve model performance.

According to what its name implies, "Random Forest is a classifier that includes a number of decision trees on different subsets of the given dataset and takes the average to enhance the predictive accuracy of that dataset." Instead than based on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

V. PROPOSED SYSTEM

- Dataset - Provide dataset (For a machine that doesn't see data the same way that people do, the acquired data must be standard and intelligible.)
- Pre-processing - Real-world data frequently contains sound, absent values, and may even be in an unfavourable format, disqualifying it from being used right away in machine learning models. Preparing the data for a machine learning model by cleaning it is important in order to increase the model's precision and efficacy.
- Feature Extraction: to produce new features from existing ones in order to reduce the no.of features in a dataset. This new, more condensed list of features should then act as a summary of almost all of the data contained in the original set of features.
- Classification - The Classification algorithm, which uses supervised learning to categorise new observations in light of training data, is used to recognise new observations. In classification, a programme makes use of the dataset or observations that are provided to learn how to categorise fresh observations into various classes or groups.


Dataset Collection and Selection
Data Preprocessing
Feature Extraction
Split Data into 70% Training and 30%Testing
Classification using machine learning algorithm
Model Generation
Model Accuracy Classification Report

5.1 Proposed Solution

Before Resource utilization is optimised through a sound resource allocation strategy. Predicting workload is essential for effective resource allocation. Therefore, this study's objective is to aid cloud service providers in increasing resource utilization. Workload trace data from Google is being examined in this study. Google made Google cluster usage trace data available to the public for research purposes in order to expose researchers to actual data and the complexity that cloud providers actually confront. The dataset is a collection of production workload traces from Google clusters.

VI. CONCLUSION AND FUTURE WORK

One of the top concerns for cloud providers is resource utilization. Predicting workload is one method of enhancing resource utilization. The ensemble model for workload prediction is provided and contrasted with the baseline study in this study. These models' accuracy and root mean square error (RMSE) were tested. To establish a benchmark against which to compare our results, state-of-the-art workload classification approaches for cloud resource utilization prediction were researched. Our data demonstrates that a Utilizing ensemble enhances performance. A stack generalization ensemble is what we call the "Ensemble based workload predictor" in our suggested approach. As base classifiers, KNN and RF have demonstrated good accuracy, and the ensemble created using them has also produced better results. An improvement of about 2% over the individual students. Baseline research revealed an RMSE of 0.37, and the proposed approach reduced error by 6.26% in CPU and 18.3% in memory usage.

A workload prediction technique for automated resource allocation may be suggested in the future. The secret to getting the most out of your resources is to choose a resource allocator with the best accurate prediction module. For cloud users, an ensemble prediction module can be created to aid in resource sizing decisions.

REFERENCES

[1]. Iglesias, J.O., et al. A Methodology for Online Consolidation of Tasks through More Accurate Resource Estimations. in Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on. 2014.

[2]. Caglar, F. and A. Gokhale. iOverbook: Intelligent Resource-Overbooking to Support Soft Real-Time Applications in the Cloud. in Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on. 2014.

[3]. Qi, Z, et al. Harmony: Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud. in Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on. 2013.

[4]. Hu, R., et al., Efficient Resources Provisioning Based on Load Forecasting in Cloud. The Scientific World Journal, 2014. 2014: p. 12.

[5]. Reiss, C., A. Tumanov, and G. Ganger, Towards understanding heterogeneous clouds at scale: Google trace analysis. Center for cloud computing, 2012.

[6]. Kundu, S. et al., “Modeling virtualized applications using machine learning techniques”, in Proc. Of 8th ACM SIGPLAN /Sigpos conference on Virtual Execution Environments, pp3 – 14, London, UK 2012

[7]. Kupferman, J. et al., “Scaling Into the Cloud”. University Of California, Santa Barbara, Tech. Rep. http://cs.ucsb.edu/~jkupferman/docs/ScalingIntoTheCloud s.pdf. 2009.

[8]. Quiroz, A et al., “Towards autonomic workload provisioning for enterprise Grids and clouds” in Grid Computing, 2009 10th IEEE/ACM International Conference pp 50-57, October, 2009

[9]. Sadeka, I. et al., “Empirical prediction models for adaptive resource provisioning in the cloud”, Future Generation Computer Systems, vol. 28, no. 1, pp 155 – 165, January, 2012

[10]. Sakr, G.E et al., "Artificial intelligence for forest fire prediction" IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pp.1311-1316, July 2010.

[11]. Sapankevych, N and Sankar, R., "Time Series Prediction Using Support Vector Machines: A Survey," Computational Intelligence Magazine, IEEE, vol.4, no.2, pp.24-38, May 2009.

[12]. H. Wang, H. Shen, and Z. Li, “Approaches for resilience against cascading failures in cloud datacenters,” in *Proc. of ICDCS*, 2018.

[13]. W. Wei, H. Fan, X.and Song, and J. Fan, X.and Yang, “Imperfect information dynamic stackelberg game based resource allocation using hidden markov for cloud computing,” *Trans. on SC*, 2018.

[14]. M. Xu and R. Buyya, “Brownout approach for adaptive management of resources and applications in cloud computing systems: A taxonomy and future directions,” *ACM Computing Surveys (CSUR)*, 2019.

DOI: 10.48175/IJARSCT-12470

[15]. Y. Yu, F. Jindal, V.andBastani, F. Li, and I. Yen, "Improving the smartness of cloud management via machine learning based workload prediction," in *Proc. of COMPSAC*, 2018.

[16]. M. Hassan, H. Chen, and Y. Liu, "Dears: A deep learning based elastic and automatic resource scheduling framework for cloud applications," in *Proc. of UBICOMP*, 2018.

DOI: 10.48175/IJARSCT-12470