ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Anticipatory Cloud Resource Allocation: A Machine Learning Frontier

Syed.Karimunnisa[1],

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Vaddesvaram, Guntur, AP, India-522302, karimun1.syed@gmail.com

Supriya Menon M[2]

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Vaddesvaram, Guntur, AP, India-522302,

supriyamenon05@gmail.com

Abstract.

Effective asset provisioning is a difficult task in the ever-changing cloud computing environment, as both insufficient and excessive allocation can negatively impact system performance. Due to this dilemma, scholars are investigating novel approaches, with a growing emphasis on proactive provisioning strategies that foresee and address resource requirements ahead of time. Proactive approaches contain intricate decision-making processes that allow the system to anticipate real-time requirements by allocating resources before they are actually needed, in contrast to reactive provisioning strategies, which react to demands as they arise. Although proactive provisioning techniques are more complex than reactive ones, they have a clear benefit in terms of quicker response times. A more flexible and responsive system is enhanced by the capacity to decide how best to provision resources before the need really materialises. But in order to successfully apply proactive provisioning solutions, an analytical framework that can precisely predict resource requirements must be adopted. In order to foresee and meet the changing demands of cloud environments, proactive provisioning must function smoothly, and this is made possible by the analytical framework. In such a circumstance, it becomes essential to incorporate a forecasting mechanism in order to put into practise a reliable and proactive resource provisioning strategy. The provisioning procedure gains sophistication from the capacity to forecast the future resource consumption of the upcoming computational jobs, guaranteeing that the system is ready to manage changing workloads effectively. The field of predicting resource utilisation in the cloud is

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

still in its early stages, despite the advances achieved in research on cloud resource usage. This underscores the continuous need for breakthroughs and innovations in this crucial area of cloud computing. The future of resource provisioning in cloud systems will surely be greatly influenced by the development and refining of cloud prediction models as technology advances.

Keywords: Virtual Machine, workload prediction, scheduling, SLA.

1. Introduction

As opposed to reactive provisioning techniques, proactive resource provisioning tactics provide faster response times, which lowers system costs overall. Although these proactive approaches have advantages, they also come with additional complexity because they require a prediction model that can predict future resource requirements. As a result, developing an efficient load prediction model is a prerequisite for suggesting an effective proactive provisioning solution [1]. Predicting the expected future workload is the main goal when it comes to resource requirements in cloud environments. This predictive feature minimises or completely eliminates delays in resource provisioning at the exact moment when it is needed by giving the system enough time to make the essential preparations. Scholars differ in their interpretation of what constitutes a "workload" for a cloud-based application [2]. Workload has been measured using metrics like resource utilisation and application request volume; response time and CPU utilisation are taken into account when estimating the characteristics of the workload.

In this study, workload was defined as future Virtual Machine (VM) demand, and reaction time and throughput were used as workload metrics to predict performance. Five machine learning techniques were used to estimate the CPU utilisation forecast parameter[3]. The selected machine learning techniques build a model using past data, train it, and then use it to make predictions in the future. These techniques frequently use a time series to depict system behaviour. The server log from the Parallel Workloads Archive public cloud server was used for the experimentation, and WEKA 3.8 was used for each experiment. The logs were evaluated using K-Nearest Neighbours, Support Vector Machine, and Random Forest

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

machine learning techniques, which yielded insightful results about the effectiveness and precision of the proactive provisioning models that were being considered [4].

New developments in cloud computing have made it possible to efficiently finish complicated jobs by greatly speeding up the execution of large-scale computations. This development has made it possible for significant scientific process applications to be distributed throughout the cloud infrastructure. A workflow, as used here, is an application that has been divided into sections and consists of several computing tasks that are closely related to each other due to dependencies on data and control flow [5].

Even with the cloud's improved computational capabilities, customers still have to pay according to how much of the available processing power they utilise. Cloud users frequently work within predetermined budgetary constraints, which means they must carefully strike a balance between completing computations[6] or operations quickly and adhering to the set expenditure limits. This paradigm highlights the difficulty faced by customers in the ever-changing cloud computing market as they attempt to maximise compute efficiency while adhering to financial constraints.

Scheduling on particular instances that are accessible is a task that comes with running workflow applications in the cloud. Resource provisioning, which involves choosing instances for computation, and job scheduling on the selected instances, are the two main components of this procedure. Cloud instances differ in terms of price, processing power, and other aspects; some are better suited for storage, memory, or image processing than others [7][8].

The task at hand is an NP-hard[9][10][11] computationally complicated problem: finding the workflow's best execution time within a specified budget. Due to the impracticality of using exhaustive search to solve such problems on a broad scale, heuristic or meta-heuristic algorithms have become common methodologies for problem resolution. These algorithms offer effective methods for handling the difficulties involved in optimising workflow execution time while staying within financial limitations.

Objectives

1. Examine machine learning as a substitute strategy to deal with resource provisioning issues when scientific operations are carried out on cloud platforms.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

2. Analyse how well a trained machine learning model does in providing answers that are similar in quality but come at a much faster speed than conventional algorithms.
3. Utilise machine learning to create well-informed recommendations regarding resource distribution, guaranteeing that the algorithm is provided with optimised and pertinent data.
4. For resource provisioning, combine machine learning with a scheduling algorithm to quickly identify the best scheduling possibilities for scientific workflows that nevertheless maintain a favourable make span under the given budgetary limitations.

2. Related Work

The main goal is to investigate how machine learning may be applied to resource provisioning for cloud-based scientific workflow execution. With faster access to provisioning and scheduling tools, scientists will be able to manage complex processes and schedule workflows more effectively. This is the goal of this exploration. The aim is also to stimulate more scholarly research into the application of machine learning to scheduling and resource provisioning for cloud-based processes.

H. Wang, H. Shen, and Z. Li [12] and others stress that in order for cloud providers to achieve optimal performance at the lowest possible cost, they must effectively utilise their resources. Resources are shared by several users in the dynamic world of cloud computing, which provides on-demand services via an internet-based pay-as-you-go paradigm. In order to distribute resources from a finite pool in a way that responds to changing user needs, the resource allocator is essential. It is critical to steer clear of both excessive and insufficient resource allocation because the former might cause a decline in service quality while the latter wastes funds and resources. Resource allocators can benefit from being able to forecast future resource demand, which will increase the effectiveness of their decisions about resource supply. In order to help resource allocators provide optimal resource provisioning and guarantee that resources are allocated as effectively as feasible, it is necessary to construct a resource utilisation prediction technique.

W. Wei, H. Fan, X.and Song, and J. Fan, X.and Yang [13] and colleagues emphasise how cloud computing's rapid improvements have made it possible to handle big computations more quickly, especially for scientific workflow applications. Resource scheduling, which is

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

the process of planning tasks on selected instances when they are selected, is what happens when workflow applications are executed in a cloud environment. Heuristics and metaheuristics are frequently used to solve the NP-hard problem of finding the fastest execution time (makespan) for scientific workflows within a certain budget. In order to address the challenges associated with managing scientific workflows in the cloud, a different approach to resource provisioning is investigated by incorporating machine learning. The goal of the project is to find out if a machine learning model that has been trained can predict provisioning instances with a level of solution quality that is on par with the state-of-the-art approach (PACSA), but in a lot less time. Solution instances are produced by the PACSA technique and are used as labels by machine learning models created for scientific processes like Montage and Cybershake. In order to acquire a makespan, an independent HEFT scheduler is utilised to schedule the expected provisioning instances, which adds to a thorough evaluation of the suggested machine learning-based resource provisioning methodology.

The significance of projecting future resource requirements several minutes in advance and utilising the Virtual Machine (VM) boot time to proactively assign resources and maintain Service Level Agreements (SLA) is emphasised by [14] M. Xu and R. Buyya. This paper explores the development and assessment of cloud client prediction models using Support Vector Machine (SVM), Neural Networks (NN), and Linear Regression (LR) machine learning techniques for a benchmark online service. In order to improve the customer's options for scaling selection, the study includes two SLA metrics: throughput and response time. Using a longer trial duration of approximately 100%, the model is applied to the public cloud infrastructure of Amazon EC2, building upon earlier studies. To further improve the accuracy of the study's conclusions, a random workload pattern is used in an attempt to create a more accurate simulation.

The use of machine learning in resource scheduling and provisioning for the cloud-based scientific processes is explored by [15] Y. Yu, F. Jindal, V. and Bastani, F. Li, and I. Yen. The resource provisioning problem, which is a part of the larger resource scheduling challenge, and its convoluted relationship are the main emphasis of the thesis. Research that has already been done frequently offers combined approaches to deal with both challenges simultaneously in an effort to reduce scheduling conflicts.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

A thorough literature analysis identifies and groups previous publications into three categories that are relevant to this thesis:

a. Methods currently used for scheduling scientific procedures on the cloud.

b. The use of machine learning to execute scientific procedures in the cloud.

c. Machine learning-based resource provisioning tactics for cloud service providers.

M. Hassan, H. Chen, and Y. Liu [16] emphasize that in order to satisfy Service Level Agreement (SLA) requirements, Virtual Machine (VM) resources must be provisioned a few minutes in advance, taking into account the VM boot-up time. One of the most important strategies for accomplishing this objective is to anticipate future resource demands. In this study, cloud client prediction models customized for the TPCW benchmark web application are developed and assessed using three machine learning techniques: Support Vector Machine, Neural Networks, and Linear Regression. The study aims to give clients a more dependable scaling decision option by incorporating SLA indicators for Response Time and Throughput into the prediction model. Interestingly, the results show that, out of the three machine learning methods examined, the support vector machine produces the best prediction model.

Machine Learning Algorithms

Our suggestion is to employ machine learning methods to forecast cloud resource usage. This recommendation is justified by the insufficiency of using basic linear regression models. This restriction results from the non-linear scaling of activities carried out on cloud-based computational resources, where job complexity is not systematically correlated with input magnitude. In addition, some tasks may require a vector of input data instead of a single item, in which case several linear regression models must be used. Therefore, in order to accurately anticipate future resource utilization, we recommend using machine learning to create models using historical data—specifically, previous task executions.

Online and offline learning are distinguished in the field of machine learning. Problem examples are presented in a sequential manner during online learning, while all instances are exposed at once during offline learning. Because training models like KNN, RF, and SVM requires processing a large amount of data, which uses a significant number of computer resources, we choose offline learning in our situation. We also use supervised learning, which

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

makes the assumption that every training data instance has the proper output known, making back-propagation techniques easier to use.

K- Nearest Neighbours (KNN): This is the simplest machine learning algorithm that can be used for both regression and classification. The forecast in regression applications is derived from the mean of the K most similar occurrences. To produce predictions, this algorithm makes use of the original dataset directly. To determine which K instances are most comparable to a new input, it uses a distance measure. This approach is called IBk (Instance-based k) in Weka and is classified as a lazy set of classifiers. Changing the values of K and distance measures allows one to modify the model's accuracy. We chose K to be 7 for this experiment based on the frequency of occurrences in the log. Furthermore, because the attributes were measured differently, we used the Manhattan distance as the distance function.

Support Vector Machine (SVM): Support-vector models, sometimes referred to as support-vector networks or SVMs, are supervised learning models with matching learning algorithms intended for classification and regression analysis. An SVM method uses a set of training examples, each labelled as falling into one of two categories, to create a model during the training phase. Although there are ways to use SVM in a probabilistic classification situation, like Platt scaling, this renders SVM a non-probabilistic binary linear classifier. In essence, the SVM model maps these examples as points in space, with the goal of optimising the distance between cases of various categories. As a result, fresh samples are projected into this area, and depending on which side of the gap they fall into, it is predicted which category they belong to.

Two types of SVM:

1.Linear SVM: Data that can be clearly separated into two groups by a single straight line are referred to be "linearly separable data". Such data are classified using the linear support vector machine, and the classifier that is used is known as the linear support vector machine classifier.

2. Non-Linear SVM: A dataset is said to be non-linearly segregated if a straight line cannot be used to classify it adequately. In these situations, the classifier that is used is called a non-linear support vector machine classifier.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Random Forest (RF): The supervised learning approach makes use of the well-known machine learning algorithm Random Forest. This method can be used to solve regression and classification-based machine learning challenges. based on the idea of ensemble learning, which blends several classifiers to improve model performance and tackle difficult issues.

The Random Forest classifier, as its name implies, uses many decision trees on various subsets of the provided dataset, averaging their predictions to increase the accuracy of predictions. Instead of depending just on one decision tree, the random forest combines predictions from all of the trees and selects the prediction that receives the most votes.

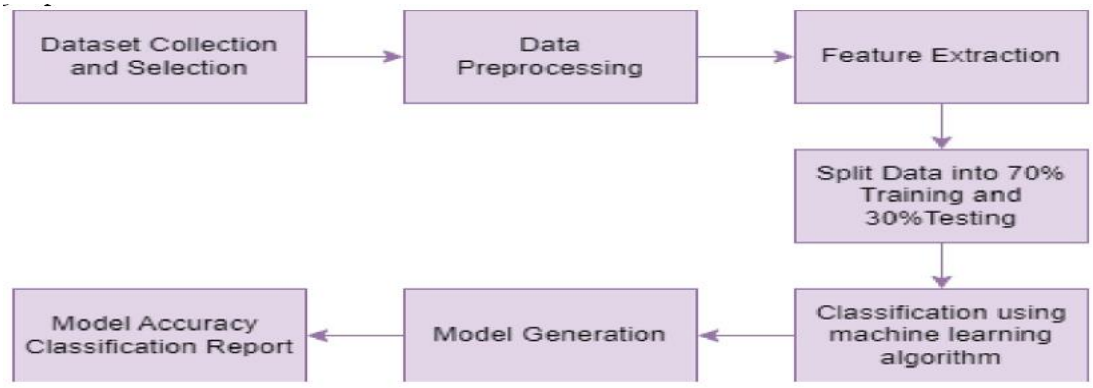
3. Proposed System

Dataset: The availability of a dataset is essential because robots interpret data differently than people do. The obtained data must follow common formats and be comprehensible in order to guarantee compatibility and understanding.

Pre-processing: Real-world data is frequently unsuitable for immediate use in machine learning models due to issues including noise, missing values, and unfavorable formats. Cleaning the data is a necessary preprocessing step that will improve the accuracy and performance of machine learning models.

Feature Extraction: In order to decrease the number of features in a dataset, this entails creating new features from ones that already exist. The information in the original set of features is thoroughly summarized in the resulting condensed list of features.

Classification: Based on training data, the Classification algorithm uses supervised learning to classify new observations. This algorithm is able to classify new observations into different groups or classes by using the dataset or observations that are supplied as training data.



ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Effective resource allocation must be in place before optimal resource utilisation can be achieved. The precise estimation of workload is essential for efficient resource allocation. Therefore, the main goal of this research is to help cloud service providers optimise their use of resources. The inquiry explores the examination of workload trace data that was acquired from Google. For research reasons, Google has kindly made their cluster usage trace data publicly accessible. This allows researchers to interact with real data and understand the challenges that cloud providers confront. The collection consists of Google cluster production workload traces.

4. Conclusions

Workload prediction is one way that cloud providers can solve their top worry, which is resource utilization. In this paper, an ensemble model for workload prediction is presented and compared methodically to a baseline investigation. These models' precision and root mean square error (RMSE) underwent extensive testing. Innovative workload classification techniques for cloud resource utilization prediction were looked into in order to create a standard for comparison. The results demonstrate that the use of an ensemble greatly improves performance. Our suggested method, which we call the "Ensemble-based Workload Predictor," makes use of a stack generalization ensemble. Specifically, base classifiers like as KNN and RF are quite accurate, and the ensemble that is generated from them performs better, showing an improvement of about 2% above single classifiers. An RMSE of 0.37 was found in the baseline investigation, however our suggested method showed a notable decrease in error—6.26% in CPU and 18.3% in memory consumption. This suggests a significant improvement in workload forecast accuracy. The study makes further development of a workload prediction method for automated resource allocation suggestions possible. Choosing a resource allocator with the best accurate prediction module is the key to maximizing resource use. The development of an ensemble prediction module is noteworthy for cloud users as a useful tool to help with well-informed resource sizing decisions.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

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.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

[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.

[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.