Auto Scaling of Cloud Resources using Time Series and Machine Learning Prediction

A thesis submitted for the Degree of

Master of Science (Research)

By

Sivasankari Bhagavathiperumal

In

School of Computer Science

UNIVERSITY OF TECHNOLOGY SYDNEY

AUSTRALIA

MARCH 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

Date: March 2020

Author: Sivasankari Bhagavathiperumal

Title: Auto Scaling of Cloud Resources Using Time Series and Machine Learning

Degree: Master of Science (Research) in Computing Sciences

I, Sivasankari Bhagavathiperumal declare that this thesis, is submitted in fulfilment of the requirements for the award of Master of Science (Research) in computing Sciences, in the School of Computer Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 02/10/2020

Acknowledgments

My first acknowledgement is to God for being my source of strength. I would like to thank my supervisor Dr. Madhu Goyal for the opportunity to work with her on the thesis, for all the guidance she has given me along the way. I thank the University of Technology Sydney for providing me a wonderful education environment and resources that made this work possible. I would also thank the experts who were involved in validating my project. Finally, I thank my parents, husband and kids for supporting me in my years of study. Thank you all.

Table of Contents

Table of Contents

List of Tables

List of Figures

Abstract

Cloud platforms are becoming very popular as a means to host business applications, and most businesses are starting to use the cloud platform in order to reduce costs. A cloud platform is a shared resource that enables businesses to share the services Software as a service (SAAS), Infrastructure as a service (IAAS) or Anything as a service (XAAS), which are required to develop and deploy any business application. These cloud services are provided as virtual machines (VM) that can handle the end user's requirements. The cloud providers must ensure efficient resource handling mechanisms for different time intervals to avoid wastage of resources. Auto-scaling mechanisms would take care of using these resources appropriately along with providing an excellent quality of service. Therefore, a process to predict the workload is required so that the cloud servers can handle the end user's request and provide the required resources as Virtual Machines (VM) disruptively, so that the businesses using the cloud service only pay for the service they use, thus increasing the popularity of cloud computing. The workload consists of the application programs running on the machine and the users connected to and communicating with the computer's applications. The computing resources are released based on the workload identified using the resource provisioning techniques, which are models used for enabling convenient on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, applications and services).

In order to provide these resources dynamically without any interruption, an auto-scaling system is required to both manage the load and balance the service. The application should be capable of handling the load as it comes in and to ensure the cloud resources are not supplied in abundance when there are fewer requests coming to the server. This thesis introduces a framework to identify the current workload and predict the future load to the server. The results show the actual workload to the server and the level of requests expected in the future. But to fully appreciate the flexibility of identifying the future load in advance, it was important to identify a method that could in turn provide guidance with adjusting the resources supplied reducing the cost. Also, this would ensure the service level agreement (SLA) was met. The framework identified in this thesis carefully monitors the inputs to the server and analyses the load. The researches also applied deep learning techniques to predict the future load to the server, which produced a result with less root mean square error (RMSE) compared the other techniques. This thesis investigates five different methods to identify the future workload: Naïve, ARIMA, Linear regression, Exponential smoothing and Forward propagation. The logs of two datasets were analysed to predict the future resource for different time intervals along with computing the error between the predicted and actual value.