MSc Project Report

Predictive Based Multi factor Horizontal POD Auto Scalar for Kubernetes

A.Niroshan.G Dabare

2018

A report submitted as part of the requirements for the degree of

MSc in Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

Abstract

Kubernetes is a container orchestration tool which can be scaled to deploy application horizontally. Current horizontal pod auto scalar considers only average CPU utilization as a scaling factor. Also it scales the application once it reaches the threshold level, it called reactive based scaling. There could be a resource over utilization if the threshold set to a higher value and underutilization if the threshold set to low value. End of the day end users getting low quality of service from the application. CPU utilization is not the only factor to considering about application scaling. There should be other factors like average memory utilization, average incoming requests per minute which impacts the application availability.

Proposed solution is to rectify current drawbacks of the Kubernetes horizontal pod auto scalar. Its forecasting upcoming resource utilization using ARIMA model, which was used to predict time series dataset. Then the target application has pre-scaled to handle upcoming workload as a proactive based method. Also proposed solution considers multiple factors to scale application such as average CPU utilization, average memory utilization and average incoming request getting from load balancer. Ultimately proposed solution tries to keep resource utilization to an optimal level while keeping end users service availability high.

Proposed solution hosted under Google cloud platform using Google Kubernetes engine service and Python was used to developed prediction model as well as the desire pod counting and scaling components. Elastic search was used to store target application metrics statistics and it gathered by metric beat agent. Stack driver monitoring API also used to gather load balancer statistics. Implemented prototype was tested under several workload conditions under Kubernetes default auto scalar and predictive based multi factor auto scalar. Finally prototype was evaluated by domain experts and some future enhancements identified. Test results are proven that the introduced enhancements to the auto scalar effect positively to the end users quality of service.