Expert System for Kubernetes Cluster Autoscaling and Resource Management

Lasal Sandeepa Hettiarachchi[1], Dilmi Palliyaguruge[1], Senura Vihan Jayadeva[1], Rusiru Abhisheak Vikum Bandara[1], Udara Srimath S.Samaratunge Arachchillage[1], and Dharshana Kasthurirathna[1]

[1]Department of Computer Science and Software Engineering, Sri Lanka Institute of formation Technology, Sri Lanka.
it19132310@my.sliit.lk, it19120980@my.sliit.lk, it19139036@my.sliit.lk, it19104218@my.sliit.lk, udara.s@sliit.lk, dharshana.k@sliit.lk

Abstract

The importance of orchestration tools such as Kubernetes has become paramount with the popularity of software architectural styles such as microservices. Furthermore, advancements in containerization technologies such as Docker has also played a vital role when it comes to advancements in the field of DevOps, enabling developers and system engineers to deploy are manage applications much more effectively. However, infrastructure configuration and management of resources are still challenging due to the disjointed nature of the infrastructure and resource management tools' failure to comprehend the deployed applications and create a holistic view of the services. This is partly due to the extensive knowledge required to operate these tools or due to the inability to perform specific tasks. As a result, multiple tools and platforms need to configure together to automate the deployment, monitoring and management processes to provide the optimal deployment strategy for the applications. In response to this issue, this research proposes an expert system that creates a centralized approach to cluster autoscaling and resource management, which also provides an automated low-latency container management system and resiliency evaluation for dynamic systems. Furthermore, the time series load prediction is done using a BiLSTM and periodically creates an optimized autoscaling policy for cluster performance, thus creating a seamless pipeline from deployment, monitoring scaling, and troubleshooting of distributed applications based on Kubernetes.

Keywords

Auto-Scaling, Chaos Engineering, Containerization, Docker, Kubernetes, Load Prediction, Machine Learning, Microservices