INTISARI

ADAPTIF *HORIZONTAL AUTOSCALING* PADA PLATFORM KUBERNETES DENGAN *BIDIRECTIONAL LONG-SHORT TERM MEMORY* (Bi-LSTM) DAN *GRAPH NEURAL NETWORK* (GNN)

Oleh

Sahala Wahyu Wardana

22/501440/PPA/06391

Arsitektur *microservices* menjadi hal yang populer dalam pengembangan aplikasi *cloud native*. Dengan meningkatnya jumlah aplikasi yang dikembangkan melalui konsep *microservice*, efisiensi dalam pengelolaan *resource cloud* menjadi tantangan penting, terutama untuk menangani trafik yang fluktuatif. Beberapa penelitian sebelumnya telah menggunakan algoritme prediktif untuk mengalokasikan *resource*, namun umumnya belum mempertimbangkan fitur dependensi antar *services*. Hal ini menyebabkan prediksi penggunaan *resource* menjadi kurang akurat karena pengaruh *service* lain dalam dependensi belum diperhitungkan. Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan Graphen-HPA, sebuah strategi *horizontal pod autoscaling* yang memanfaatkan kombinasi algoritme *Bidirectional Long Short-Term Memory* (Bi-LSTM) dan *Graph Neural Network* (GNN). Model ini menambahkan fitur dependensi antar *service* dengan lapisan *Graph Convolutional Networks* (GCN) untuk meningkatkan akurasi prediksi penggunaan *resource*. Pengujian performa dari *autoscaling* dilakukan dengan membuat simulasi trafik secara realistis yang mencerminkan fluktuasi trafik agar *microservice* yang telah diluncurkan mendapatkan beban dan *autoscaling* diharapkan secara dinamis menyesuaikan nilai replika berdasarkan jumlah prediksi *resource*. Hasil evaluasi menunjukkan bahwa model Bi-LSTM-GCN belum dapat secara signifikan mengungguli Bi-LSTM, meskipun memiliki nilai rata-rata R^2 *score* sebesar 0,98, dibandingkan Bi-LSTM dengan nilai 0,97 dari pengujian KFold. Dari sisi skor efisiensi waktu *scaling*, model Bi-LSTM unggul sebesar 0,762 dan model Bi-LSTM-GCN lebih rendah yaitu 0,738.

Kata Kunci*: predictive autoscaling, horizontal autoscaling,* Bi-LSTM, *Graph Neural Network, microservice*s

ABSTRACT

ADAPTIVE HORIZONTAL AUTOSCALING ON KUBERNETES PLATFORM WITH *BIDIRECTIONAL LONG-SHORT TERM MEMORY* (Bi-LSTM) AND *GRAPH NEURAL NETWORK* (GNN)

By

Sahala Wahyu Wardana

22/501440/PPA/06391

The microservices architecture has become increasingly popular in the development of cloud-native applications. With the growing number of applications built using the microservice concept, achieving efficiency in cloud resource management poses a significant challenge, particularly in handling fluctuating traffic. Previous research has utilized predictive algorithms for resource allocation, but these approaches generally do not account for service dependency features. This oversight results in less accurate resource usage predictions, as the influence of other dependent services is not considered. To address this limitation, this study proposes Graphen-HPA, a horizontal pod autoscaling strategy that leverages a combination of Bidirectional Long Short-Term Memory (Bi-LSTM) algorithms and Graph Neural Networks (GNN). The model adds service dependency features using Graph Convolutional Network (GCN) layers to improve the accuracy of resource usage predictions. The performance of the autoscaling mechanism is evaluated through realistic traffic simulations that reflect traffic fluctuations, enabling deployed microservices to experience varying loads and the autoscaling mechanism is expected to dynamically adjust the number of replicas based on predicted resource usage. The evaluation results indicate that the Bi-LSTM-GCN model has not significantly outperformed the Bi-LSTM model, despite achieving an average R^2 score of 0,98 compared to Bi-LSTM's 0,97 in K-Fold testing. In terms of scaling time efficiency scores, the Bi-LSTM model achieved a higher score of 0.762, while the Bi-LSTM-GCN model scored lower at 0.738.

Keywords: predictive autoscaling, horizontal autoscaling, Bi-LSTM, Graph Neural Network, microservices