UNIVERSITAS
GADJAH MADA
Implementasi Time Series Forecasting Berbasis Machine Learning untuk Autoscaling Proaktif Pada Kubernetes
Safarini Gandira, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

IMPLEMENTASI TIME SERIES FORECASTING BERBASIS MACHINE LEARNING UNTUK AUTOSCALING PROAKTIF PADA KUBERNETES

Safarini Gandira
22/503022/SV/21462

Mekanisme penskalaan pada Kubernetes cenderung memiliki waktu jeda dalam melakukan replikasi, hal ini berpotensi menyebabkan beban berlebih pada sistem dan pemborosan biaya akibat alokasi sumber daya yang berlebih. Untuk mengatasi permasalahan tersebut, proyek ini membangun sebuah sistem *autoscaling* proaktif berbasis *machine learning* dengan pendekatan prediksi deret waktu guna mengoptimalkan penggunaan CPU dan memori. Dalam penerapannya, model akan memprediksi penggunaan CPU yang kemudian dijadikan acuan untuk menyesuaikan jumlah replika pod secara preventif sebelum terjadinya lonjakan beban. Arsitektur model yang digunakan pada penelitian ini yaitu *Gated Recurrent Unit* (GRU), *Long Short-Term Memory* (LSTM), *Bidirectional* LSTM (Bi-LSTM), dan Bi-LSTM-transformer yang kemudian dibandingkan performanya dengan *Horizontal Pod Autoscaler* (HPA) bawaan Kubernetes yang bersifat reaktif. Berdasarkan hasil pengujian beban yang dilakukan, *autoscaler* berbasis Bi-LSTM paling efisien dalam mengalokasikan penggunaan CPU. *Autoscaler* berbasis Bi-LSTM mencapai tingkat akurasi nilai R2 *score* sebesar 0,9746, MAE 0,0058 dan RMSE 0,0080. Sementara itu, sistem *autoscaling* yang mengalokasikan memori terefektif adalah *autoscaler* berbasis GRU, dengan nilai memori terendah 173,9MB dan nilai R2 *score* 0,9678.

Kata kunci : *Autoscaling*, Bi-LSTM, GRU, Kubernetes, LSTM, prediksi deret waktu

UNIVERSITAS
GADJAH MADA
Implementasi Time Series Forecasting Berbasis Machine Learning untuk Autoscaling Proaktif Pada Kubernetes
Safarini Gandira, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

IMPLEMENTING MACHINE LEARNING-BASED TIME SERIES FORECASTING FOR PROACTIVE AUTOSCALING IN KUBERNETES

Safarini Gandira
22/503022/SV/21462

The scaling mechanism in Kubernetes tends to experience a lag time during replication, potentially causing system overload and cost inefficiencies due to resource over-provisioning. To address this issue, this project developed a proactive machine learning-based autoscaling system utilizing a time-series forecasting approach to optimize CPU and memory utilization. In its implementation, the model predicts CPU usage, which serves as a reference to preventively adjust the number of pod replicas before a load spike occurs. The model architectures evaluated in this study include Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Bi-LSTM-Transformer, whose performances were then compared against the reactive, default Kubernetes Horizontal Pod Autoscaler (HPA). Based on the load testing results, the Bi-LSTM-based autoscaler proved to be the most efficient in allocating CPU usage, achieving an accuracy level with an R2 score of 0.9746, MAE of 0.0058, and RMSE of 0.0080. Meanwhile, the most effective autoscaling system for memory allocation is the GRU-based autoscaler, recording the lowest memory consumption at 173.9 MB with an R2 score of 0.9678.

Keyword: Autoscaling, Bi-LSTM, GRU, Kubernetes, LSTM, time series forecasting