RINGKASAN

IMPLEMENTASI AUTOSCALING MENGGUNAKAN HORIZONTAL POD AUTOSCALER (HPA) DI KUBERNETES PADA APLIKASI BERBASIS CLOUD SERVER. Ferdiansyah Maula Syarif NIM E32222371, Tahun 2025, Teknologi Informasi, Politeknik Negeri Jember, Bekti Maryuni Susanto, S.Pd. T, M.Kom., (Dosen Pembimbing).

Dalam era digital saat ini, aplikasi web sering kali menghadapi fluktuasi trafik yang signifikan, yang dapat mempengaruhi kinerja dan ketersediaan layanan. Untuk mengatasi tantangan ini, autoscaling menjadi solusi yang sangat penting, karena memungkinkan aplikasi untuk secara otomatis menyesuaikan jumlah sumber daya yang digunakan berdasarkan beban kerja yang sedang berlangsung. *Autoscaling* adalah proses yang secara dinamis menambah atau mengurangi kapasitas sumber daya, seperti server atau kontainer, untuk memastikan bahwa aplikasi dapat menangani lonjakan trafik tanpa mengalami *downtime* atau penurunan kinerja. Dalam konteks Kubernetes, salah satu mekanisme *autoscaling* yang paling umum digunakan adalah *Horizontal Pod Autoscaler* (HPA). HPA secara otomatis mengelola jumlah pod yang berjalan dalam sebuah *deployment* berdasarkan metrik yang ditentukan, seperti penggunaan CPU atau memori, serta metrik kustom lainnya. Dengan HPA, ketika beban kerja meningkat, HPA akan menambah jumlah *pod* untuk memastikan aplikasi tetap responsif, dan sebaliknya, ketika beban kerja menurun, HPA akan mengurangi jumlah *pod* untuk mengoptimalkan penggunaan sumber daya. Hal ini tidak hanya meningkatkan efisiensi operasional, tetapi juga mengurangi biaya, karena pengguna hanya membayar untuk sumber daya yang benar-benar digunakan. Dengan demikian, integrasi antara web dan *autoscaling*, khususnya melalui *Horizontal Pod Autoscaler*, menjadi sangat krusial dalam menciptakan aplikasi yang skalabel, responsif, dan efisien di lingkungan *cloud* yang dinamis.