ABSTRACT

Availability is a critical factor for application operation as it directly impacts user experience. Implementing autoscaling in a cloud computing environment can help maintain the availability and the scalability of an application while also improving operational cost efficiency. Garvan News, a public application, seeks to implement autoscaling but faces challenges due to limited human resources. This issue can be handled through the Infrastructure As Code (IAC) approach, which enhances consistency and efficiency in application development. The implementation of autoscaling with IAC involves analyzing requirements and creating both architectural and stack designs. The stack is divided into three parts, which are then sequentially implemented on AWS. Then, a load test was conducted using Apache JMeter on the web application to determine the maximum load that AWS can handle during a surge in requests. AWS autoscaling can handle requests well when tested with 25,000 and 40,000 users. However, it has a very high error percentage when loaded with 50,000 users.

Keywords: autoscaling, infrastructure as code, AWS, threshold, load testing