INTISARI

Analisis Komparatif Arsitektur Monolitik dan Arsitektur *Microservices* pada Pengembangan Aplikasi Chatbot Berbasi AI Generatif

Oleh

Ahmad Ali Masykur

Perkembangan teknologi kecerdasan buatan, khususnya AI generatif, mendorong kebutuhan akan sistem aplikasi yang efisien, *scalable*, dan tangguh. Dalam konteks tersebut, pemilihan arsitektur perangkat lunak menjadi aspek krusial, terutama antara arsitektur monolitik dan microservices. Penelitian ini bertujuan untuk menganalisis secara komparatif kedua arsitektur dalam konteks pengembangan aplikasi chatbot berbasis AI generatif, dengan fokus pada performa sistem dan aplikasi di bawah berbagai beban kerja.

Penelitian ini menggunakan pendekatan eksperimental melalui implementasi dua versi sistem chatbot—monolitik dan microservices—yang diuji dalam tiga skenario beban (rendah, sedang, dan tinggi). Proses pengujian dilakukan di lingkungan cloud menggunakan Google Cloud Platform. Metrik performa dikumpulkan melalui *load testing* menggunakan k6 dan pemantauan sistem menggunakan Glances, yang mencakup penggunaan CPU, memori, aktivitas disk I/O, jaringan, serta metrik performa HTTP. Analisis data dilakukan menggunakan pendekatan regresi non-parametrik serta penilaian kualitas layanan (Quality of Service/ QoS) untuk membandingkan kinerja kedua arsitektur.

Hasil penelitian menunjukkan bahwa arsitektur monolitik cenderung lebih cepat dalam respon pada beban ringan hingga sedang, sementara microservices menunjukkan stabilitas yang lebih baik pada beban tinggi. Perbandingan berdasarkan dimensi QoS seperti efisiensi sumber daya, keandalan, dan responsivitas menunjukkan bahwa tidak ada arsitektur yang unggul secara mutlak; masing-masing memiliki kekuatan pada skenario tertentu. Temuan ini dapat menjadi acuan dalam pengambilan keputusan pada pengembangan sistem berbasis AI generatif dengan pertimbangan konteks dan kebutuhan spesifik.

Kata Kunci: Arsitektur Monolitik, Arsitektur *Microservices*, Chatbot, AI Generatif, Pengujian Beban, k6, Glances, Google Cloud Platform, *Quality of Service* (QoS), Regresi Non-parametrik

ABSTRACT

Comparative Analysis of Monolithic Architecture and Microservices Architecture on Generative AI-Based Chatbot Application Development

By

Ahmad Ali Masykur

The advancement of artificial intelligence, particularly generative AI, has increased the demand for application systems that are efficient, scalable, and resilient. In this context, the choice of software architecture plays a crucial role, especially between monolithic and microservices architectures. This study aims to conduct a comparative analysis of both architectures in the development of a generative AI-based chatbot application, focusing on system and application performance under varying workloads.

The research follows an experimental approach by implementing two versions of the chatbot system—monolithic and microservices—and evaluating their performance under three load scenarios: low, medium, and high. Testing was conducted on the Google Cloud Platform, using k6 for load testing and Glances for system-level monitoring. Performance metrics collected include CPU utilization, memory usage, disk I/O activity, network throughput, and HTTP-level metrics. Data were analyzed using non-parametric regression techniques and Quality of Service (QoS) evaluation to objectively compare the two architectural models.

The results indicate that the monolithic architecture delivers faster response times under low to medium loads, whereas the microservices architecture exhibits better stability under high load conditions. When evaluated across QoS dimensions such as resource efficiency, reliability, and responsiveness, neither architecture proved superior in all scenarios. Each architecture demonstrates strengths under specific conditions, providing valuable insights for system designers when selecting the appropriate architecture for AI-driven applications.

Keywords: Monolithic Architecture, Microservices Architecture, Chatbot, Generative AI, Load Testing, k6, Glances, Google Cloud Platform, Quality of Service (QoS), Non-parametric Regression