INTISARI

IMPLEMENTASI SISTEM ASISTEN DOKUMEN BERBASIS WEB MENGGUNAKAN *LARGE LANGUAGE MODEL* PADA AWS

Febri Eli Enyelika

22/493218/SV/20675

Pengelolaan dokumen PDF secara konvensional cenderung tidak efisien seiring meningkatnya volume dokumen. Penelitian ini mengimplementasikan dan mengevaluasi sistem DocJa, yaitu asisten dokumen berbasis *Large Language Model* (LLM) yang dibangun di atas infrastruktur *cloud computing* Amazon Web Services (AWS). Sistem mengintegrasikan EC2, Lambda, API Gateway, S3, DynamoDB, serta model Claude Haiku melalui Anthropic API untuk analisis dan tanya jawab berbasis konteks. Hasil pengujian *black box* menunjukkan tingkat keberhasilan 100% dari 23 *test case*. *Response time* berkisar antara 69 ms hingga 9.412 ms sesuai kompleksitas proses. Pada *load testing*, sistem mampu meningkatkan *throughput* dari 0,23 menjadi 27,58 *req/s* tanpa kegagalan (*error rate* mendekati 0%), menunjukkan efektivitas *auto-scaling* AWS Lambda. Pengujian *cold start* menunjukkan tambahan latensi sebesar 2,258 detik dibandingkan *warm start*. Selain itu, sistem mampu melakukan *text extraction*, *automatic summarization*, serta *question answering* baik pada dokumen tunggal maupun lintas dokumen secara relevan dan konsisten. Secara keseluruhan, arsitektur *hybrid* EC2 dan Lambda mampu mendukung kebutuhan sistem dalam menangani beban kerja yang dinamis secara efisien.

Kata Kunci: Cloud Computing, Serverless, Amazon Web Services (AWS), Large Language Model, Performance Evaluation

ABSTRACT

IMPLEMENTATION OF A WEB-BASED DOCUMENT ASSISTANT SYSTEM USING A LARGE LANGUAGE MODEL ON AWS

Febri Eli Enyelika

22/493218/SV/20675

Conventional PDF document management tends to be inefficient as document volume increases. This study implements and evaluates DocJa, an AI-powered document assistant based on a Large Language Model (LLM) deployed on Amazon Web Services (AWS) cloud computing infrastructure. The system integrates EC2, Lambda, API Gateway, S3, and DynamoDB, along with the Claude Haiku model via the Anthropic API for context-based document analysis and question answering. The black box testing results show a 100% success rate across 23 test cases. The system's response time ranges from 69 ms to 9,412 ms depending on process complexity. In load testing, the system improves throughput from 0.23 to 27.58 req/s with no failures (error rate close to 0%), demonstrating the effectiveness of AWS Lambda's auto-scaling. The cold start evaluation indicates an additional latency of 2.258 seconds compared to the warm start condition. Furthermore, the system is capable of performing text extraction, automatic summarization, and question answering across both single and multiple documents with relevant and consistent results. Overall, the hybrid EC2 and Lambda architecture effectively supports dynamic workloads in an efficient manner.

Keywords: *Cloud Computing, Serverless, Amazon Web Services (AWS), Large Language Model, Performance Evaluation*