International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 11, Issue 2, Jul–Dec 2022; 465–488

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

SERVERLESS PLATFORMS IN AI SAAS DEVELOPMENT: SCALING SOLUTIONS FOR REZOOME AI

Hrishikesh Rajesh Mane[1], Aravind Ayyagari[2], Archit Joshi[3], Om Goel[4], Dr. Lalit Kumar[5] & Prof. (Dr.) Arpit Jain[6]

[1]The State University of New York at Binghamton, Binghamton New York, US

[2]Wichita State University, Dr, Dublin, CA, 94568, USA

[3]Syracuse University, Syracuse, Sadashivnagar New York, USA

[4]ABES Engineering College Ghaziabad, India

[5]Dept. of Computer Application IILM University Greater Noida, India

[6]KL University, Vijaywada, Andhra Pradesh, India

ABSTRACT

The rapid advancement of artificial intelligence (AI) technologies has significantly influenced the development of Software as a Service (SaaS) applications. As organizations strive to leverage AI capabilities, the demand for scalable, efficient, and cost-effective architectures becomes paramount. This paper explores the potential of serverless platforms in enhancing AI SaaS development, specifically focusing on Rezoome AI, an innovative application designed for resume optimization and job matching.

In traditional architectures, the monolithic approach often leads to challenges related to scalability, maintenance, and deployment efficiency. These limitations hinder the agility required for AI applications, where quick iterations and updates are essential. By adopting a microservices architecture within a serverless framework, this study investigates how Rezoome AI can effectively overcome these challenges.

The methodology employed includes designing a serverless architecture that decomposes Rezoome AI into distinct microservices, each responsible for specific functionalities such as data processing, machine learning model inference, and user interaction. This approach allows for independent scaling of services based on demand, resulting in improved resource utilization and reduced operational costs. The architecture leverages cloud providers like AWS Lambda, which facilitates automatic scaling and eliminates the need for server management.

Results from the implementation reveal that the serverless architecture significantly enhances Rezoome AI's performance and scalability.

Key metrics indicate a reduction in response time by approximately 40% compared to the previous monolithic structure. The independent scaling of microservices enables dynamic resource allocation, resulting in efficient handling of peak loads during high-demand periods. Moreover, the serverless model reduces operational overhead, allowing developers to focus on core functionalities rather than infrastructure management.

Furthermore, the paper discusses the lessons learned from transitioning to a microservices architecture. These include the importance of defining clear service boundaries, effective API management, and the necessity of robust monitoring tools to ensure system reliability.

KEYWORDS: *Serverless, AI, SaaS, Scalability, Automation, Rezoome, Cloud Computing, Function-as-a-Service*

Article History

Received: *16 Nov 2022* | ***Revised:*** *22 Nov 2022* | ***Accepted:*** *28 Nov 2022*