

ABSTRACT

A system for recommending and optimizing artificial intelligence (AI) workload placement in a multi-vendor cloud environment is disclosed. The system utilizes a server with at least one processor to access input datasets stored in a datacentre associated with an AI workload and determine suitable processing units from various manufacturers based on predefined criteria. It calculates the required number of processing units and accesses a multi-vendor performance database. A deep learning model predicts infrastructure requirements for the AI workload, enabling the generation of recommendations for optimal processing unit configuration. The system automatically allocates processing unit resources from multiple manufacturers based on these recommendations. Finally, it generates data for a user interface dashboard, presenting information about manufacturers, processing unit types, recommended configurations, and real-time performance metrics of allocated resources. This system optimizes AI workload placement across diverse cloud environments and hardware options.

Fig. 1