International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

AI-Driven Optimization Strategies InCloud Computing: A Comprehensive Review

Atharva Subandha[1], Sahil Pathan[2], Pratiksha Parkhe[3]

[1, 2, 3]Student
[1, 2, 3]Department of Computer Science, Haribhai V. Dessai College, Pune – 1

Abstract

Cloud computing delivers adaptable and scalable infrastructure, yet managing resources efficiently remains a persistent challenge, often leading to elevated operational costs and underutilized assets. This study explores the use of artificial intelligence (AI) techniques to enhance cloud resource management through methods such as dynamic scaling, automated provisioning, and intelligent workload distribution.

It examines the hurdles associated with integrating AI into cloud systems and presents a detailed feasibility analysis of these techniques. The research concludes with recommendations for future advancements aimed at sustainable and efficient cloud resource utilization.

1. Introduction and Problem Identification

With the widespread adoption of cloud technologies by organizations aiming for scalable and cost-effective computing environments, the importance of efficient resource allocation has significantly increased. Despite the promise of flexibility, many enterprises continue to over-allocate resources, resulting in financial inefficiencies and wasted computational power.

Although modern cloud platforms provide options for dynamic scaling, manual intervention and configuration often fall short when it comes to handling unpredictable workloads. This report investigates the potential of AI-based solutions to streamline cloud resource management, reduce costs, and optimize overall infrastructure performance.

2. Objectives of the Research

- To investigate various AI-based approaches currently employed in cloud resource management.
- To classify these techniques and assess their efficiency and reliability.
- To evaluate the practicality of implementing AI in diverse cloud-based applications.
- To identify research gaps and propose future directions for AI-enhanced optimization in cloud environments.

3. Literature Review / Study

- Overview of Previous Work

Researchers have developed a range of AI methodologies to improve cloud optimization. Common approaches include:

- PredictiveScaling: Techniques utilizing historical usage patterns to forecast resource demands, thereby reducing excess provisioning.
  - For example, Bousselmi& Abbes (2020) demonstrated machine learning models that predict future workloads.
  - Kaur &Kaushik (2023) explored prediction-driven strategies to manage dynamic environments effectively.
- Auto-Scaling: Real-time analytics allow systems to allocate or withdraw resources in response to demand fluctuations.
  - Tang & Pan (2021), as well as Wang & Li (2023), emphasized the effectiveness of adaptive auto-scaling algorithms.
- Workload Distribution: AI techniques are used to intelligently balance workloads across servers to prevent congestion and inefficiency.
  - Lee & Choi (2021) reviewed machine learning algorithms for equitable workload distribution in virtual environments.

- Research Value

  Despite these advances, many existing models are generalized and may not meet the specific requirements of industry sectors like finance, healthcare, or e-commerce. Tailoring AI solutions for such sectors offers a valuable opportunity for impactful research.

4. Feasibility Study

- Technical Aspect

  AI models require extensive datasets and computational resources for training and deployment. Predictive algorithms, in particular, demand large-scale data ingestion pipelines and scalable model architectures. Furthermore, sector-specific customization increases implementation complexity.

- Operational Considerations

  Maintaining AI models in cloud environments necessitates periodic retraining, tuning, and system integration. Organizations must possess not only the infrastructure but also the human expertise to manage AI-driven systems effectively.

- Economic Factors

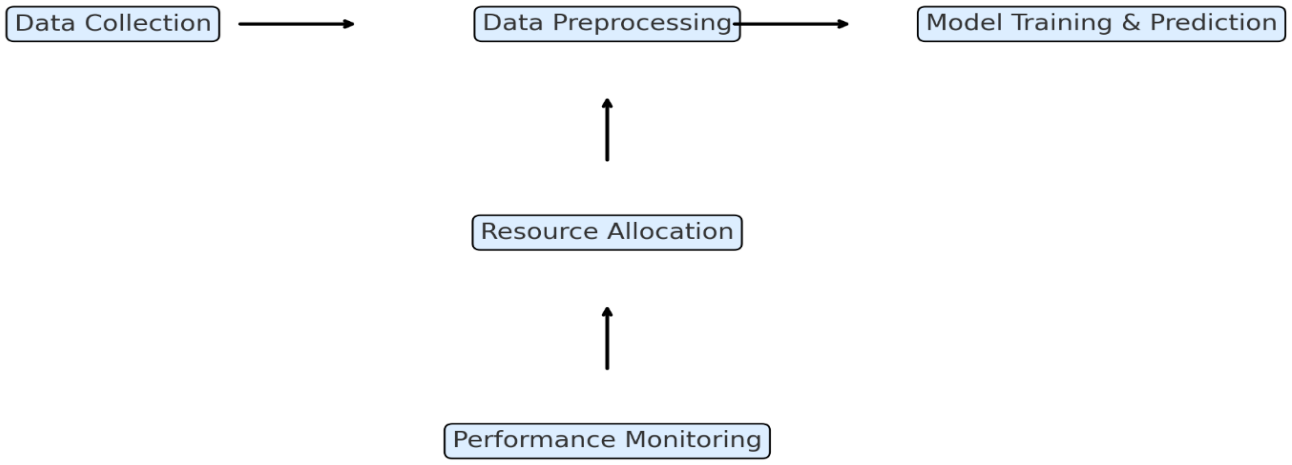
  Though the initial setup and integration of AI systems may involve significant expenses, they can result in long-term cost savings by optimizing resource consumption. For smaller enterprises, however, affordability remains a barrier, making simpler AI models more practical.

5. Design

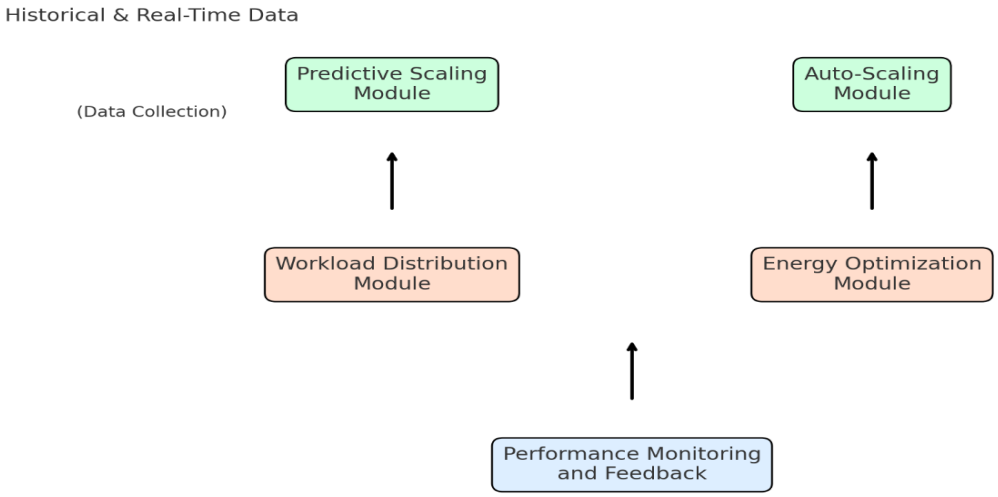
The proposed system architecture includes the following components:

- Database Modules
  - Usage History Repository: Stores traffic and workload patterns for model training.
  - Performance Metrics Store: Tracks response time, uptime, and resource usage to inform ongoing optimization.
  - Energy Consumption Logs: Used for power-aware resource scheduling.

- System Flow
  - Data Collection: Aggregation of real-time and historical data.
  - Pre-processing: Cleaning and transformation of data for model compatibility.
  - Model Training: AI algorithms are trained using pre-processed datasets.
  - Prediction and Allocation: System forecasts usage and adjusts resources dynamically.
  - Monitoring and Feedback: Continuous performance tracking and retraining to maintain optimization.



- Modules
  - Predictive Engine: Forecasts future resource needs.
  - Auto-Scaling Controller: Reacts to real-time changes.
  - Workload Balancer: Distributes tasks evenly across nodes.
  - Sustainability Manager: Oversees power usage and environmental compliance.



6. Results

Through a comprehensive review, it is evident that AI-based solutions can greatly enhance the efficiency of cloud resource management. Predictive models and real-time allocation techniques significantly lower operational costs and improve responsiveness. However, limitations like computational demands, dependency on data, and lack of generalizability still pose implementation challenges.

7. Future Scope and Limitations

- Future Research Directions
  - Developing AI models tailored for specific sectors.
  - Creating algorithms capable of optimizing multi-cloud deployments.
  - Emphasizing environmental sustainability in AI design for cloud platforms.
- Limitations
  - High resource requirements for training and deployment of AI models.
  - Dependency on comprehensive and high-quality historical data.
  - Customization challenges due to varied application requirements.

8. References

1. Bousselmi M., & Abbes T. (2020). A machine learning-based approach for cloud resource optimization. IEEE Access, 8, 107911-107923.
2. Tang X., & Pan X. (2021). AI-driven cloud resource allocation for dynamic workloads. Journal of Cloud Computing, 10(1), 25-35.
3. Kaur S., &Kaushik N. (2023). Predictive scaling and optimization in cloud environments using AI. International Journal of Computer Applications, 178(20), 45-52.
4. LeeJ., & Choi D. (2021). Intelligent resource management in cloud computing: A review of machine learning methods. Journal of Cloud Technology and Innovation, 4(3), 205-218.
5. Chen X., & Gupta R. (2021). AI for energy-efficient resource scheduling in cloud platforms. Green Computing Review, 5(2), 112-120.
6. Patil S., & More A. (2022). Power optimization using AI in cloud data centers. Sustainable IT Journal, 3(1), 55-61.
7. Nerella H., et al. (2024). AI-Driven Cloud Optimization: A Comprehensive Literature Review. International Journal of Computer Trends and Technology (IJCTT), 72(5), 177-181.
8. Joloudari J. H., et al. (2022). Resource Allocation Optimization Using Artificial Intelligence Methods in Various Computing Paradigms: A Review. arXiv preprint, arXiv:2203.12315.