

Automated Infrastructure Management and Resource Optimization in Cloud Environments

Sagar Kesarpu

Submitted: 07/02/2022

Revised: 15/03/2022

Accepted: 25/03/2022

Abstract: Cloud computing has revolutionized enterprise infrastructures through scalability, flexibility, and provision of resources on demand. However, manually managing cloud infrastructures has become difficult due to the dynamism of the workloads, deployment in multiple clouds and changing resource demands. Infrastructure automation involves intelligent automation that is used to monitor, allocate, configure and optimise cloud infrastructure in real-time. Resource optimisation involves automation combined with artificial intelligence, machine learning and Infrastructure as Code to optimise infrastructure and improve its reliability. The system constantly monitors and analyses the usage patterns and behaviour of the workload along with the performance metrics in order to predict the capacity needs and accordingly optimise the computing, storage, and networking resources. Moreover, automation helps in making cloud computing operations better through smart allocation, configuration management, and self-healing. The automation process is capable of detecting any anomalies within the system and fixing the problems accordingly. This results in increased service availability, reduced downtime, and overall enhanced resiliency of the cloud environment in hybrid and multi-cloud settings. Continuous monitoring and observability also ensure real-time tracking of infrastructure performance, thus allowing for proactive improvement of its efficiency and fast troubleshooting. Integration of automation with DevOps and AIOps allows organisations to automate deployment processes, perform predictive maintenance of resources and implement governance policies for the resources. Therefore, consistency, improved scalability, and effective utilisation of the infrastructure are ensured.

Keywords: *Automated infrastructure management, resource optimization, cloud automation, Infrastructure as Code, AI-driven optimization, AIOps, self-healing infrastructure, cloud observability.*

Introduction

It is now essential for an organization to automate cloud infrastructure management and optimization because there is now the need for the distribution of computing power in order to offer digital services. There are dynamic workloads created in the cloud environment that require constant provision, management, and optimization of the cloud infrastructure to ensure optimum performance and availability. Manual infrastructure management proves to be inefficient in responding to the changing demands in the cloud infrastructure. Therefore, companies are adopting various strategies such as IaC, AI, and automation to enhance their cloud environment. These features help enhance the use of the infrastructure in question while lowering the cost of operations and

avoiding any disruption of services. The current study looks at the topic of automated infrastructure management and resource optimization in the cloud environment via secondary qualitative research.

Related work

Infrastructure management using automation technology has been researched more intensively due to the emergence of cloud computing. Previous research was dedicated to virtualization and rule-based automation. These technologies provided a decrease in manual involvement; however, they faced challenges associated with dynamic nature of loads and complex distributed systems. Murphy (2022) states that that Infrastructure as Code (IaC), which is considered a basis of automated cloud infrastructure management. It allows for consistent provisioning, configuration and deployment of the infrastructure within different cloud environments. The research reveals that tools like Terraform and AWS CloudFormation enhance the process of deployment and eliminate possible mistakes in configurations. Resource optimization has turned

Expert Application Engineer
Leading Financial Institution, USA
sagark546@gmail.com
ORCID: 0009-0003-6272-3968

into one of the main issues due to continuous changing of cloud workloads. Various techniques related to autoscaling have been studied. Lins *et al.*, (2021) provides that automation of the cloud infrastructure has been increased by artificial intelligence and machine learning capabilities. The predictions of the historical behavior of the workload help determine future resource needs. Some researchers have claimed that the allocation of resources using AI helps decrease the number of idle resources. The observability and monitoring techniques are tightly related to the management of the infrastructure. Prior studies have explained the role of the real-time metrics, logging, and tracing in the identification of performance problems prior to the disruption of the services. There are still many research gaps in this area. Firstly, multi-cloud environments raise the issue of the interoperations and governance in the process of automation of resource allocation.

Method

Qualitative research design is applied in the study since it involves examining the knowledge about the automated infrastructure management and optimization of resources in cloud systems. This research design enables one to expound more on the concept, technology, practice, and experience. Qualitative research design is relevant in this case since it seeks to examine patterns and insights instead of numerical relationships (Tomaszewski, Zarestky & Gonzalez, 2020). In this study, the research employs secondary data from peer-reviewed journal articles, conference presentations, industry reports, and documentation provided by the cloud providers. Secondary data gives current information on the technology of automation in the cloud without using surveys or interviews. This data helps the researcher compare different sources and establish patterns.

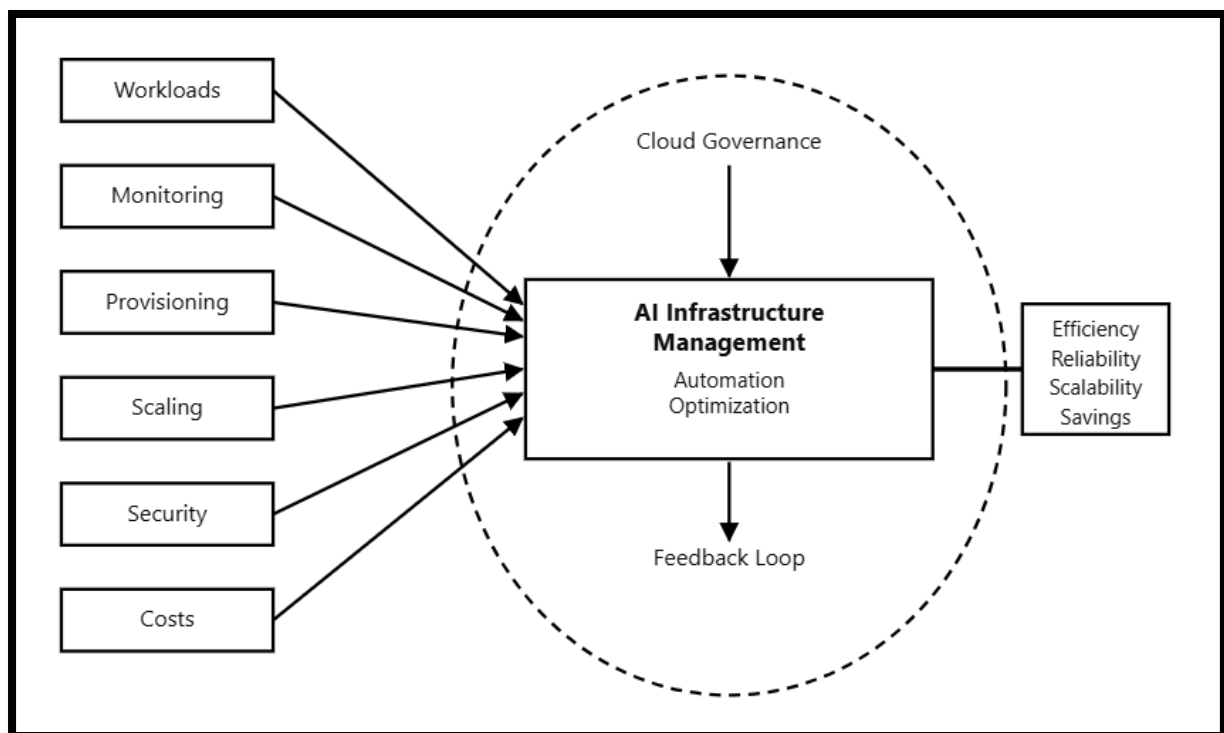


Figure 1: Conceptual Framework for Automated Infrastructure Management and Resource Optimization in Cloud Environments

(Source: Self-created)

The research follows an inductive approach since it generates its themes and conclusions from the existing literature rather than testing a hypothesis (Teru, Denis & Hamilton, 2020). This research employs thematic analysis in analyzing the secondary data that was collected. Thematic

analysis entails the identification of recurring themes, as well as similarities and differences in published literature. Thematic analysis aids in organizing the findings in categories in relation to the theme of automation, resource allocation, performance optimization, governance, and

implementation issues (Prakash & Das, 2020). This research adheres to ethical research principles through the use of credible and accessible academic materials from public sources. All sources are appropriately cited according to APA 7th edition citation rules.

Results

Automation of Infrastructure Provisioning and Configuration Management

Automated provisioning and configuration management have come to be a central approach to enhancing cloud performance and enterprise

productivity. Companies now prefer automated provisioning and configuration management instead of manually configuring their infrastructures through the use of IaC (Infrastructure as Code). AI high performers report stronger governance and operational processes that directly support automated provisioning. 60% have well-defined capability-building programs for AI technology personnel, compared with 46% of other organizations (Teru, Denis & Hamilton, 2020). Similarly, 57% follow standardized AI development and deployment protocols, while only 43% of other organizations do.

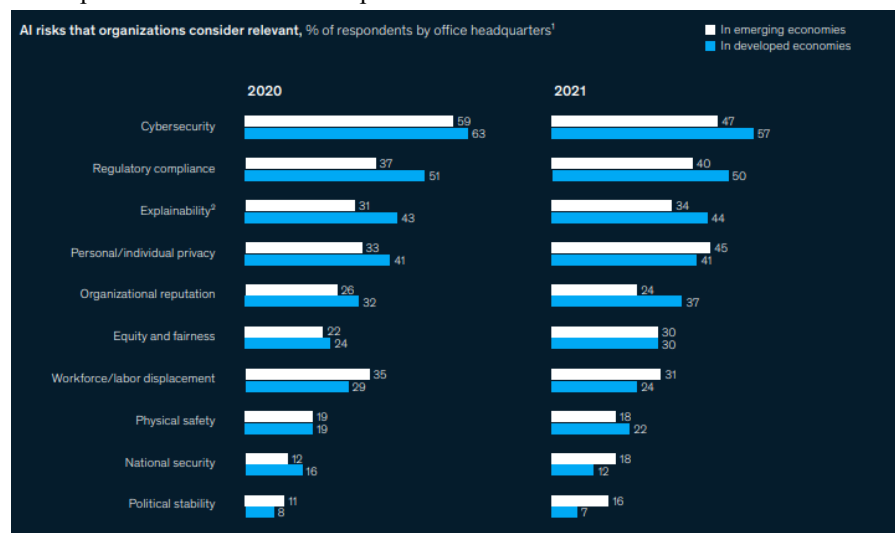


Figure 2: AI Risks Considered Relevant and Mitigated by Organizations in Developed and Emerging Economies

(Source: McKinsey, 2021)

The use of practices for data governance is more prominent as 40% of AI high-performers practice the implementation of a robust data governance practice, which is 20% less than other organizations. Moreover, 38% of AI high-performing organizations monitor their AI models performance once they get deployed, while only 20% of other organizations perform this practice. Adoption of cloud infrastructure further helps automate the configuration management process. 64% of AI high-performers' workload runs on a public or hybrid cloud infrastructure, while 44% of other organizations have similar practices (Patchamatla, 2022). This helps in improving consistency, scalability, efficiency, predictability of costs, reducing configuration errors, and cloud modernization. The previous studies demonstrate that the use of Generative and Agentic AI together will help enhance application resilience, minimize

technical debt, improve efficiency, and innovate continually (Soundappan, 2022).

AI-Driven Resource Allocation and Workload Optimization

Resource allocation and workload optimization with AI in cloud infrastructure enhance the infrastructure by analyzing the workload, recognizing abnormalities and allocation of computing resources according to actual needs (Holowaychuk, 2022). With artificial intelligence, one can achieve predictive scaling, intelligent alerts, and workload scheduling, decreasing manual involvement and increasing reliability of the infrastructure. The following factors help businesses optimize the usage of cloud resources and keep the performance of applications. There is evidence that companies which perform better in AI use much effort in the recruitment of staff to create and manage their optimization system.

AI Capability	Cloud Function	AI High Performers	Other Organizations	Infrastructure Benefit	Operational Outcome
AI Data Scientists	Workload Prediction	60%	31%	Predictive Scaling	Higher Performance
AI Product Managers	Resource Planning	40%	21%	Intelligent Provisioning	Faster Deployment
AI Translators	System Integration	25%	8%	Workflow Automation	Better Coordination
Auto-Scaling Models	Demand Forecasting	Dynamic Scaling	Rule-Based Scaling	Reduced Over-Provisioning	Cost Optimization
AI Scheduling	Load Balancing	Real-Time Allocation	Manual Allocation	Balanced Workloads	Reliable Infrastructure

Table 1: AI-Driven Resource Allocation and Workload Optimization in Cloud Infrastructure

According to the global survey carried out by McKinsey, 60 percent of AI high-performing organizations hired AI data scientists against 31 percent of others. High achievers were also close to double the number of recruiting AI product managers (40% versus 21%) and over three times as many hiring AI translators (25% versus 8%). This is indicative of the need for specialized skills to build scalable AI operational systems. AI-driven optimization can make cloud computing systems efficient through the prediction of load changes and the automation of activities such as auto-scaling, load balancing, and provisioning of resources (Nagarajan, 2022). AI-driven optimization prevents over-provisioning, ensures minimal waste of

resources, and also guarantees application performance during load peaks. Companies with strong AI capabilities also automate data processes and use scalable AI architecture.

Auto-Scaling and Elastic Resource Management in Cloud Environments

These functionalities enable the cloud infrastructure to dynamically scale computing capacity depending on the load requirements of the system. Auto-scaling using AI technology makes use of characteristics of the workload, CPU usage, memory usage, and network traffic in provision of capacity to prevent performance problems (Suryadevara, 2021).

Formula 1: Auto-Scaling Efficiency

$$ASE = \frac{\text{Demand Met}}{\text{Resources Provisioned}} \times 100\%$$

This formula measures how effectively auto-scaling provisions resources to satisfy application demand.

Formula 2: Elastic Resource Utilization

$$ERU = \frac{\text{Resources Used}}{\text{Resources Allocated}} \times 100\%$$

This formula measures the percentage of allocated cloud resources that are actively utilized after elastic scaling.

These formulas directly support the section by quantifying scalability, resource efficiency, and AI-driven cloud performance optimization. The mentioned solution increases the availability of applications, reduces idle resources, and allows for

effective cloud operations. In addition to the above, the implementation of AIOps brings specific operational benefits that help manage the elastic resources effectively (Mohammed, 2022). Namely, 66.9% of the MSPs have experienced a decrease in the number of open incident tickets due to intelligent alerting and automation of remediation activities. Also, 56.7% of the organizations have

managed to automate the repetitive operational tasks, 53% – suppress alerts and correlate events, and 47% - reduce the mean time to detect and fix incidents.

Performance, Cost, and Operational Efficiency Improvements

Performance Improvements: The automation of infrastructure management helps to improve cloud performance by ensuring constant monitoring, predicting and allocating resources dynamically. The resources are auto-scaled depending on how the workload changes, thus lowering the latency and eliminating performance bottlenecks. Auto-fault detection and recovery systems make the infrastructure more reliable and available.

Cost Improvements: Through resource optimization, cloud cost savings can be realized by ensuring that resource provisioning is done based on the actual demand. Automatic scaling will help avoid situations where there is an overprovision of resources when the demand is minimal and will provide enough resources when the workload peaks (Nagarajan, 2022). Monitoring helps identify inefficient resources, thus avoiding costs.

Area	AI	Benefit	Outcome
Provisioning	Prediction	Savings	Efficiency
Scaling	Automation	Optimization	Availability
Monitoring	Detection	Reduction	Reliability
Balancing	Scheduling	Utilization	Performance

Table: Cost Improvements Through AI-Driven Resource Optimization

Poor security settings and insufficient access control may increase the likelihood of cyber attacks. To secure cloud infrastructure, it is necessary to enforce proper governance policies, identity management, encryption, and compliance monitoring.

Governance and Policy Enforcement: A governance policy is necessary for proper cloud automation management in relation to resource management, authorization of access, and standard operation procedures. The lack of consistent policy implementation can lead to resource sprawl, inconsistency in configuration, and violations (Liu et al., 2020).

Skills and Organizational Readiness: Automation in the cloud needs people who understand cloud computing technologies and automation systems.

Operational Efficiency Improvements:

Management of infrastructure through automation provides efficiency by minimizing the level of human intervention in the carrying out of administrative tasks (Berglund et al., 2020). The process of automation of provisioning, monitoring, maintenance, and fault management provides for efficient cloud services. This process gives real-time information about the performance of the infrastructure. This enables decision making and issue solving to be done faster.

Implementation Challenges and Governance for Automated Cloud Infrastructure

Integration with Legacy Systems: Some firms have legacy software applications that do not work with automation software in clouds. It becomes very challenging for the firm to integrate the legacy software with automation because it takes a lot of work. Compatibility problems can cause disruptions and delay the process of automation (Sarilo-Kankaanranta & Frank, 2021).

Security and Compliance Risks: The automation of infrastructure ensures the security of cloud resources and organizational information.

The lack of technical skills and resistance to change could slow down automation. Training, leadership support, and change management ensure that there is more preparedness for change and automation.

Discussion

The use of automated infrastructure management has become one of the ways in which cloud environments have been improved in order to enhance their performance, cut costs and improve efficiency. According to the results of the study, automation is what has made it possible for cloud computing systems to distribute computing resources depending on the demands of the workload. This makes the process more responsive, reduces latency, and ensures service availability during changes in demand. Automation of

monitoring and errors, on the other hand, makes the system more reliable through early detection of problems. The findings also indicate that resource optimization leads to cost savings in cloud computing. Through automated scaling, cloud computing systems prevent over-provisioning in times of low demand and ensure sufficient provisioning in times of peak demand (Baral, Ghimire & Basnet, 2022). Automation through infrastructure starts with intelligent provisioning, where AI provisions virtual machines, containerized applications, storage, and networking. Telemetry data for these variables are continually gathered. These are analyzed by the machine learning algorithms in order to predict workload changes and abnormalities, and also to provide recommendations on how to take corrective action (Mohammed, 2022). As a consequence, automation leads to increased financial efficiency and sustainability of cloud resource management. It contributes to enhanced decision making, faster problem-solving, and ensures uniformity of service delivery on the cloud. However, despite all the advantages, the results emphasize numerous difficulties in implementation. In addition, organizational readiness and competency of employees are vital in implementing automation successfully.

Conclusion

It concluded from this research that automation of infrastructure management enhances the performance, cost-efficiency, and effectiveness of clouds. Automation ensures dynamic allocation of resources, monitoring, and effective management of workload, ensuring efficient and scalable services. The optimization of resources cuts down infrastructure costs and enhances utilization of the system. The key success factor is effective governance, security controls, integration of the systems, and appropriate skills of the workforce. Organizations that leverage automation through proper governance and skill enhancement will succeed in their infrastructure management processes.

Reference List

- [1] Baral, B., Ghimire, B., & Basnet, D. R. (2022). Understanding policy coherence and interplay governing biodiversity conservation and associated livelihood practices in Karnali province, Nepal. *Nepal Public Policy Review*, 27-91.
<https://nepjol.info/index.php/nppr/article/view/48395>
- [2] Berglund, E. Z., Monroe, J. G., Ahmed, I., Noghabaei, M., Do, J., Pesantez, J. E., ... & Levis, J. (2020). Smart infrastructure: a vision for the role of the civil engineering profession in smart cities. *Journal of Infrastructure Systems*, 26(2), 03120001.
[https://ascelibrary.org/doi/abs/10.1061/\(ASCE\)IS.1943-555X.0000549](https://ascelibrary.org/doi/abs/10.1061/(ASCE)IS.1943-555X.0000549)
- [3] Holowaychuk, T. J. (2022). Adaptive Cloud Enterprise Framework for Secure Financial Systems AI Governance and Intelligent Workload Automation. *International Journal of Emerging Trends in Engineering and Management Research*, 7(5), 12537.
<http://ijetemr.com/index.php/ijetemr/article/view/62>
- [4] Infosys. (2021). Infosys Annual Report 2020–21: Cloud chaos to clarity. Infosys Limited. Infosys Annual Report 2020-21.
<https://www.infosys.com/investors/reports-filings/annual-report/annual/documents/infosys-ar-21.pdf>
- [5] Lakeside Team. (2019, June 18). Guide to AIOps tools in 2019: How to choose & key concepts.
<https://www.lakesidesoftware.com/blog/guide-aiops-tools-2019-how-choose-key-concepts/>
- [6] Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2021). Artificial intelligence as a service: classification and research directions. *Business & Information Systems Engineering*, 63(4), 441-456.
<https://link.springer.com/article/10.1007/s12599-021-00708-w>
- [7] Liu, T., Huang, D., Tan, X., & Kong, F. (2020). Planning consistency and implementation in urbanizing China: Comparing urban and land use plans in suburban Beijing. *Land Use Policy*, 94, 104498.
<https://www.sciencedirect.com/science/article/pii/S026483771931525X>
- [8] McKinsey, 2021. The state of AI in 2021.
<https://www.mckinsey.com/~media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Global%20survey%20The%20state%20of%20AI%20in%202021/Globa-l-survey-The-state-of-AI-in-2021.pdf>

- [9] McKinsey, 2022. The state of AI in 2022—and a half decade in review. McKinsey & Company. The state of AI in 2022. <https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai%20in%2022%20and%20a%20half%20decade%20in%20review/the-state-of-ai-in-2022-and-a-half-decade-in-review.pdf>
- [10] Mohammed, S. (2022). AI-driven IT operations and AIOps enablement across hybrid cloud platforms. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, 4(2), 4635-4639. <https://www.ijeetr.com/index.php/ijeetr/article/view/1013>
- [11] Murphy, O. (2022). Adoption of Infrastructure as Code (IaC) in Real World; lessons and practices from industry. <https://www.theseus.fi/handle/10024/786729>
- [12] Nagarajan, G. (2022). Optimizing project resource allocation through a caching-enhanced cloud AI decision support system. *International Journal of Computer Technology and Electronics Communication*, 5(2), 4812-4820. <https://ijetece.com/index.php/IJCTEC/article/view/306>
- [13] Patchamatla, P. S. S. (2022). A hybrid Infrastructure-as-Code strategy for scalable and automated AI/ML deployment in telecom clouds. *International Journal of Computer Technology and Electronics Communication*, 5(6), 6075-6083. <https://ijetece.com/index.php/IJCTEC/article/view/240>
- [14] Prakash, A. V., & Das, S. (2020). Intelligent conversational agents in mental healthcare services: a thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems*, 12(2), 1. <https://aisel.aisnet.org/pajais/vol12/iss2/1/>
- [15] Sarilo-Kankaanranta, H., & Frank, L. (2021, October). The slow adoption rate of software robotics in accounting and payroll services and the role of resistance to change in innovation-decision process. In *Conference of the Italian Chapter of AIS* (pp. 201-216). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-10902-7_14
- [16] Soundappan, S. J. (2022). Modernizing Enterprise Software Ecosystems through Artificial Intelligence Driven Integration and Cloud Transformation. *International Journal of Future Innovative Science and Technology (IJFIST)*, 5(5), 9253. <https://iadier-academy.org/index.php/IJFIST/article/view/477>
- [17] Suryadevara, S. S. K. (2021). AI-Driven Multi-Cloud Orchestration System for Enterprise Digital Experience Delivery. *American International Journal of Computer Science and Technology*, 3(1), 21-34. <http://aijcs.org/index.php/aijcs/article/view/205>
- [18] Teru, K., Denis, E., & Hamilton, W. (2020, November). Inductive relation prediction by subgraph reasoning. In *International conference on machine learning* (pp. 9448-9457). PMLR. <https://proceedings.mlr.press/v119/teru20a.html>
- [19] Tomaszewski, L. E., Zarestky, J., & Gonzalez, E. (2020). Planning qualitative research: Design and decision making for new researchers. *International journal of qualitative methods*, 19, 1609406920967174. <https://journals.sagepub.com/doi/abs/10.1177/1609406920967174>