

Enterprise Application Performance Optimization in Distributed Cloud Environments

Sagar Kesarpu

Expert Application Engineer Leading Financial Institution, USA

ARTICLE INFO

Received: 05 May 2023

Revised: 17 June 2023

Accepted: 28 June 2023

ABSTRACT

Enterprise apps now work in distributed cloud environments combining public, private, hybrid, and multi-cloud infrastructures. The above infrastructure provides an opportunity to boost the scalability, availability, and business continuity of your app while catering to globally distributed end-users. But performance management in such cloud infrastructures faces challenges like latency, utilisation of computing resources, complexity of networking, and load management, among others. Therefore, performance optimisation is key to achieving the best from your enterprise apps. Performance optimisation in distributed cloud infrastructure requires consistent monitoring, proper resource management, and effective workload management. Auto-scaling automatically adjusts computing resources based on application demand. This ensures consistent performance even in peak hours and at the lowest cost when there is low demand. Load balancing distributes the incoming traffic over several servers in different cloud regions. Moreover, AI-based analytics help increase application performance through the processing of telemetry information such as logs, metrics, traces, and events to find anomalies and identify the points of application performance breakdowns in real-time. Predictive models make it possible for companies to engage in proactive capacity planning and thus to address infrastructure challenges even before they impact end-users. The automation of optimisation processes will reduce the need for human intervention and speed up the reaction to incidents. Efficiency is achieved through containerization and microservices because these technologies enable the deployment and scaling of applications independently. This technology benefits from edge computing by performing sensitive operations in proximity to users and minimizing the network load. By using continuous observability and DevOps, complete visibility of the state of applications, infrastructure, and service level objectives can be obtained.

Keywords: Enterprise application performance, distributed cloud environments, multi-cloud computing, auto-scaling, load balancing, AI-driven analytics, Kubernetes orchestration, cloud observability.

INTRODUCTION

Performance tuning of enterprise applications is critical for ensuring that the services provided are reliable and scalable and also cost-effective in distributed cloud environments. Today's enterprise applications are run in several cloud regions, edge locations, and in hybrid environments. Optimum workloads placement, elasticity in resource management, load balancing and observability are the ways to achieve high availability and low latency for applications. The above-mentioned concepts will help in optimization of resource utilization as well as effective workloads distribution. Technologies like Kubernetes, container orchestration, caching systems, and real-time monitoring are some of the examples of cloud-native technologies that allow firms to tune their applications for improved performance based on varying workload conditions. Monitoring and distributed tracing provide continuous observability and help pinpoint any bottlenecks (Agarwal, 2022). In addition to that, AI provides significant support for cloud optimization by allowing predictive scaling, automation of infrastructure, and fault detection. As reported by McKinsey, firms are increasingly incorporating artificial intelligence into their operations.

METHODS

Research Design

This study utilized a qualitative research methodology for the analysis of performance optimization of enterprise applications in distributed cloud environments (Tomaszewski, Zarestky & Gonzalez, 2020). Qualitative research methodology was chosen since the research involved analysis of concepts, methods, approaches, and challenges related to the topic based on literature review and not numbers. In the research, an in-depth understanding of the process of cloud optimization in enterprises using workload placement, elastic scaling, load balancing, caching, and observability was sought.

Research Philosophy and Approach

An interpretivist paradigm was used in the study, as it was meant to discover the meaning, practice, and perspective provided in the already existing body of knowledge concerning cloud performance optimization (Dewi, 2021). It allows the exploration of complicated organizational and technological processes by way of interpretation. An inductive methodology was also chosen to facilitate the emergence of certain patterns from the collected literature sources. There were no hypotheses being tested, and conclusions were drawn from the interpretation of the evidence found in academic sources and cloud computing research.

Data Collection Method

The data collection process in this study was done using the secondary data collection approach. The sources of the data include peer-reviewed journal articles, papers from conferences, industry reports, cloud computing resources, and credible organizational literature.

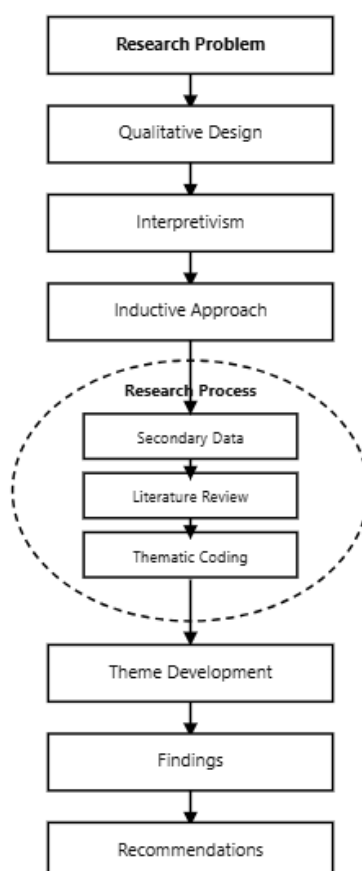


Figure 1: Research Workflow Framework for Enterprise Application Performance Optimization

(Source: Self-created)

These sources included distributed cloud technologies, enterprise application performance, Kubernetes, artificial intelligence, observability, and cloud resource optimization. The secondary data provided credible and comprehensive information that could be used in analyzing current practices and optimization processes within distributed cloud environments (Shu et al., 2020).

Data Analysis Technique

Thematic analysis was used to analyze the obtained data. Thematic analysis entails conducting literature review, determining the recurring themes, coding similar themes, and categorizing them into themes (Kiger & Varpio, 2020). The identified themes were workload placement, elasticity of resources management, load balancing, caching approaches, observability, artificial intelligence, governance, security, and efficiency of operations. Through thematic analysis, the study was able to compare findings from various sources.

Ethical Considerations

The research adhered to all ethical standards of research throughout the research process. Publicly available and credible secondary sources have been used as a means of collecting data. Ideas, findings, and evidence used in this research have been appropriately referenced using APA 7th Edition references in order to avoid any kind of plagiarism. The research upholds academic integrity by ensuring accuracy of information without manipulation or misrepresentation of evidence published. No human subjects have been used.

RESULTS

Performance Bottlenecks in Distributed Cloud Environments

Latency issues within distributed cloud computing environments stem from network latency, contention for resources, dependency between services, and the lack of observability in multiple cloud nodes (Liu & Buyya, 2020). The latency problem is further compounded by cross-regional communication, whereby contention for resources such as CPU, memory, and bandwidth leads to the degradation of performance of applications in multi-tenant architecture environments.

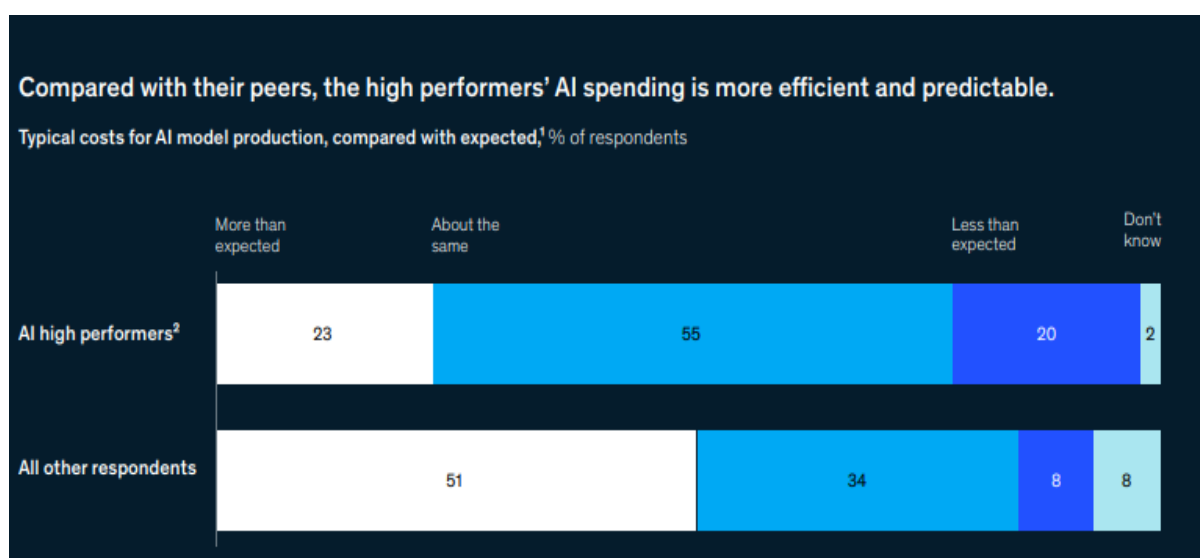


Figure 2: Compared with their peers, the high performers' AI spending is more efficient and predictable

(Source: McKinsey & Company, 2021)

Latency is a critical problem in microservice architectures because any failure in one service affects other services. AIOps has emerged as one of the key solutions for dealing with such bottlenecks using intelligent monitoring, event correlation, and automation.

Resource Allocation and Workload Optimization Strategies

Efficient resource allocation and workload optimization lead to the increased efficiency of cloud computing due to proper allocation of computing resources depending on workloads with minimal wastage. As opposed to performance optimization that seeks to optimize cloud performance, resource allocation and workload optimization is concerned with workload optimization and minimizing wastage of infrastructure. Dynamic provisioning and autoscaling ensure that computing resources are provisioned and de-provisioned respectively in response to real-time workloads (Panwar & Supriya, 2022). Thus, this minimizes over-provisioning and costs of infrastructure. Rightsizing enhances resource optimization by ensuring that cloud computing instances and capacities match application needs.

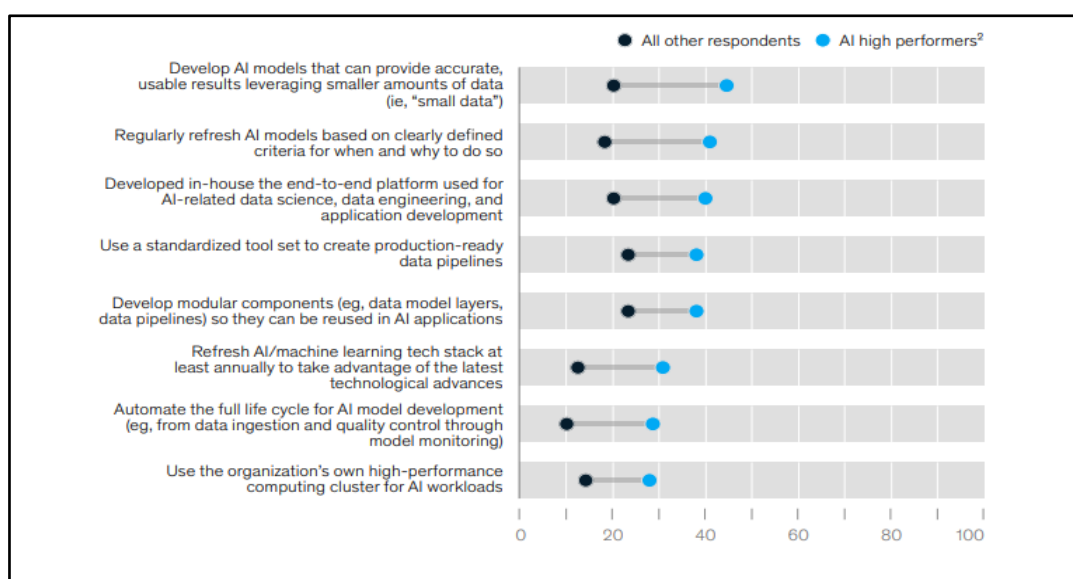


Figure 3: Models, tools, and technology practices adopted by AI high-performing organizations compared with other organizations

(Source: McKinsey & Company, 2022)

Resource allocation and workload optimization applies artificial intelligence-enabled scheduling, automation, modularity, scalable data pipelines and infrastructure to maximize resource utilization, cut costs, optimize the scalability of workloads. Based on the results of the 2022 global AI survey by McKinsey, 24% of companies apply AI for service operations optimization, thus making it the leading use of AI (Mechouche et al., 2022). Also, 63% of the surveyed companies plan to make further investments in AI solutions in the coming three years. That is due to rising use of AI-enabled infrastructure management and workload optimization. Besides, AI brings operational benefits to other organizational areas, and respondents state the greatest cost reduction in supply chain management (52%), and the revenue growth in marketing and sales (70%) and product and service development (70%).

Impact of Auto-Scaling and Load Balancing on Application Performance

Improved Application Scalability: The auto-scaling technique increases or decreases computing resources based on actual time-based demand on the system. This technique helps applications cope with any spike in the traffic without disruption to the service. Dynamic scaling makes efficient use of the computing resources, guarantees application availability, and consistent performance.

Reduced Response Time and Latency: Load balancing involves distributing the requests received to several servers or containers so as to prevent resource congestion in one particular node (Rabiu, Yong & Mohamad, 2022). The balanced workload enhances the performance by reducing the response time, minimizing latency, and increasing the efficiency of the user experience.

Enhanced System Reliability and Availability: Auto-scaling and load balancing increase the reliability of applications through the prevention of overloading servers and automatic transfer of traffic even in the event of failure of the underlying infrastructure. In case of unavailability of one of the nodes, the traffic is shifted to other available nodes (Baladari, 2022).

Technique	Function	Benefit	Outcome
Auto-scaling	Provisioning	Reliability	Availability
Load-balancing	Distribution	Stability	Continuity
Failover	Redirection	Resilience	Recovery
Monitoring	Detection	Prevention	Uptime

Table 1: Enhanced System Reliability and Availability Through Auto-Scaling and Load Balancing

Better Resource Utilization and Cost Efficiency: Autoscaling ensures that resources are not wasted because they are provided based on the need (Rzadca et al., 2020). Through load balancing, we are able to avoid under or overloading our servers. This ensures optimization of CPUs, memory, and network while reducing costs and improving cloud infrastructure efficiency.

Observability and Real-Time Performance Monitoring for Distributed Applications

Observability helps improve the performance of distributed applications by real-time monitoring of infrastructure, applications, and customer health metrics in cloud ecosystems. An instance can be seen in the transformation of Azure Support provided by Infosys, where observability was utilized to combine data from 350+ SharePoint destinations to a single monitoring platform (Ahmed et al., 2021). Observability helped in continuous monitoring of customer health metrics, automated service provision, and root cause analysis with the help of AI. As a result, performance was improved. Cycle time for root cause analysis was reduced by 75% from four weeks to one week.

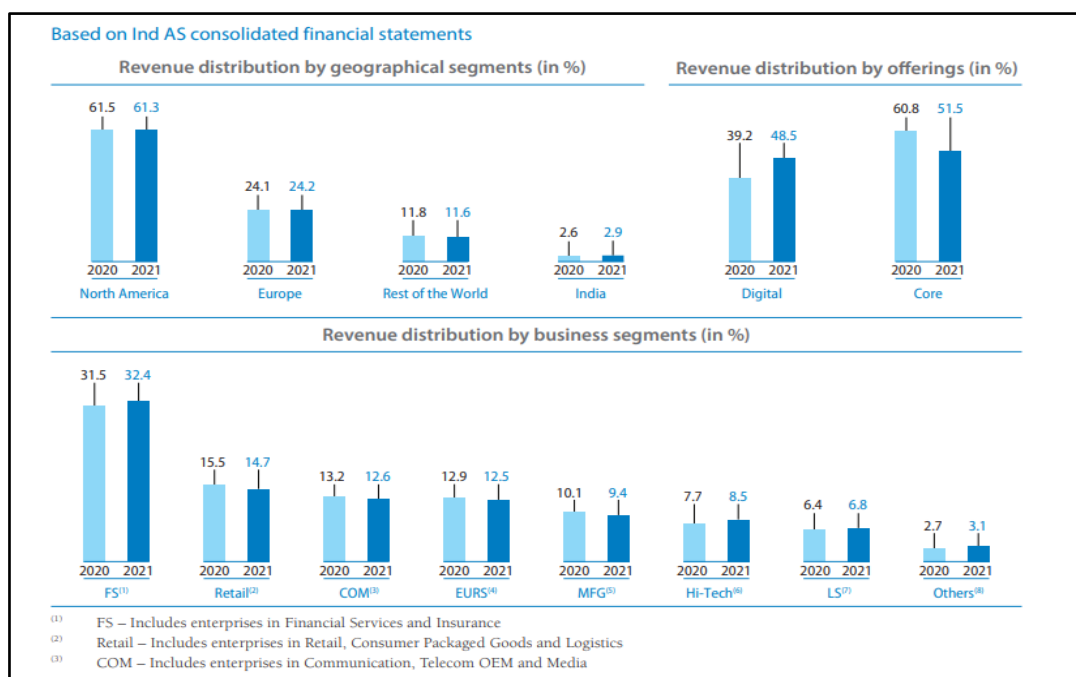


Figure 4: Revenue distribution by offerings (Digital vs Core), FY2020–FY2021

(Source: Infosys, 2021)

From the Infosys Annual Report 2020–21, we can see that digital revenue grew from 39.2% in FY2020 to 48.5% in FY2021, while core revenue fell from 60.8% to 51.5%. The growth of 9.3 percentage points is due to the rising use of cloud digital services such as artificial intelligence, automation, observability, and real-time performance monitoring (Ricca, Marchetto & Stocco, 2021). The customer onboarding process was streamlined using the automated onboarding portal in less than 42 days. Moreover, over 200 premium customers were onboarded using the new automated support system, contributing to better scalability and monitoring uniformity. In addition, the Microsoft company announced an improvement in its customer satisfaction index following the implementation of real-time monitoring and automated incident management (Korteling *et al.*, 2021). The outcomes prove that the centralization of observability, along with automated monitoring and AI-powered analysis, help improve incident management, increase service reliability, and enhance application performance.

Challenges and Best Practices for Enterprise Performance Optimization

Challenges

Operational Complexity: Enterprise settings include multi-cloud architecture, edge computing, software-defined networking, secure access service edge, and Zero Trust Network Access. Coordinating distributed applications in such an environment is challenging and results in increased complexity in terms of infrastructure management and workload orchestration.

Cybersecurity Threats: A distributed cloud infrastructure exposes enterprises to potential cyber threats due to connections made by users, applications, and other devices over several networks (Berente *et al.*, 2021). Weakness in access management, insecure APIs, and security in networks may interfere with the operation of enterprise applications. Enterprises need Zero Trust security, encryption, monitoring, and segmentation of the network.

Cost and Performance Trade-offs: There needs to be a careful balancing between the performance of applications and the cost of cloud infrastructure resources. Provisions that are too many increase the cost of operation, but those that are too few lead to latency and service bottlenecks (Perron *et al.*, 2020).

Best Practices

AI-Driven Infrastructure Automation: Artificial intelligence facilitates the automation of resource allocation, scheduling, anomaly detection, and predictive scaling within cloud computing systems. Intelligent automation eliminates manual interference, optimizes the use of infrastructure, and ensures proactive performance management (Ricca, Marchetto & Stocco, 2021). Automation of decision making is also important for improving the dependability of applications.

Unified Data Governance: Enterprise data will be consistent, accurate, and easily available within all departments and cloud environments through a data governance framework that is centralized. Data models that are standardized improve monitoring, analysis, and work load optimization. Accurate data will also improve the AI model and observability (Ricca, Marchetto & Stocco, 2021).

Software-Defined Networking and Zero Trust: Software-defined networking optimizes traffic through segregation of control and data planes, while Zero Trust constantly authenticates users and devices before allowing access. These innovations not only optimize the functioning of networks but also increase their security and reduce any unauthorized access (Ricca, Marchetto & Stocco, 2021).

DISCUSSION

In this research, the optimization of enterprise application performance in distributed cloud computing environments was analyzed through secondary sources drawn from academic literature, industry research, and business cases. The analysis revealed that the optimization of enterprise performance is driven by intelligent resource allocation, auto-scaling, load balancing, observability in real time, and AI-based automation (Sarker, 2022). Such innovations enhance application scalability, minimize latency, and ensure effective operation in distributed cloud environments. The results imply that artificial intelligence has become an influential element that drives workload optimization and operational effectiveness in organizations. AI has been used increasingly for the purpose of optimizing service operations, predicting resource needs, automation of monitoring processes, and

detecting anomalies. As seen from the Microsoft Azure case study, the use of automation for monitoring, along with root cause analysis enabled through artificial intelligence, cut the analysis cycle time by over 75% and helped speed up the process of onboarding customers using automated workflow processes (Infosys, 2021). However, the results indicate some critical challenges that need to be considered for implementation purposes. First of all, the complexity of operations associated with multiple clouds and edge computing leads to an increased level of management difficulty, with cybersecurity issues necessitating Zero Trust security, continuous monitoring, and governance. Costs versus performance and skills gaps can further contribute to implementing enterprise optimization strategies.

CONCLUSION

The importance of optimizing application performance within a scalable, reliable, and cost-effective distributed cloud system cannot be understated. The research has revealed that AI-based resource allocation, auto-scaling, load balancing, and observability contribute to better performance and efficiency. In addition, cloud native technologies increase the effectiveness of workload management and infrastructure usage. Nevertheless, effective implementation of these optimization approaches will require governance, cybersecurity measures, proper staffing, and continuous monitoring.

REFERENCE LIST

- [1] Agarwal, S. (2022). Observability in Microservices: From Traditional Monitoring to Distributed System Intelligence. *International Journal of Computer Technology and Electronics Communication*, 5(6), 16220-16226. <https://ijctec.com/index.php/IJCTEC/article/view/497>
- [2] Ahmed, T., Qureshi, H. A., Kaleem, M., & Nazir, S. (2021, April). Event monitoring and observability for industrial systems on azure cloud. In 6th International Electrical Engineering Conference. The Institution of Engineers Pakistan. <https://researchonline.gcu.ac.uk/en/publications/event-monitoring-and-observability-for-industrial-systems-on-azur/>
- [3] Baladari, V. (2022). Cloud resiliency engineering: Best practices for ensuring high availability in Multi-Cloud Architectures. *International Journal of Science and Research (IJSR)*, 11(6), 2062-2067. https://www.academia.edu/download/122471383/Cloud_Resiliency_Engineering_Best_Practices_for_Ensuring_High_Availability_in_Multi_Cloud_Architectures.pdf
- [4] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS quarterly*, 45(3), 1433-1450. <https://misq.umn.edu/misq/article/45/3/1433/1890>
- [5] Dewi, I. G. A. A. O. (2021). Understanding data collection methods in qualitative research: The perspective of interpretive accounting research. *Journal of Tourism Economics and Policy*, 1(1), 23-34. <https://journalkeberlanjutan.com/index.php/jtep/article/view/105>
- [6] 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>
- [7] Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical teacher*, 42(8), 846-854. <https://www.tandfonline.com/doi/abs/10.1080/0142159X.2020.1755030>
- [8] Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, 4, 622364. <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.622364/full>
- [9] 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/>
- [10] Liu, X., & Buyya, R. (2020). Resource management and scheduling in distributed stream processing systems: a taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 53(3), 1-41. <https://dl.acm.org/doi/abs/10.1145/3355399>
- [11] 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/Global-survey-The-state-of-AI-in-2021.pdf>

- [12] 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%202022%20and%20a%20half%20decade%20in%20review/the-state-of-ai-in-2022-and-a-half-decade-in-review.pdf>
- [13] Mechouche, J., Touihri, R., Sellami, M., & Gaaloul, W. (2022). Conformance checking for autonomous multi-cloud SLA management and adaptation: J. Mechouche et al. The Journal of Supercomputing, 78(11), 13004-13039. <https://link.springer.com/article/10.1007/s11227-022-04363-0>
- [14] Panwar, R., & Supriya, M. (2022). Dynamic resource provisioning for service-based cloud applications: A Bayesian learning approach. Journal of Parallel and Distributed Computing, 168, 90-107. <https://www.sciencedirect.com/science/article/pii/S0743731522001307>
- [15] Perron, M., Castro Fernandez, R., DeWitt, D., & Madden, S. (2020, June). Starling: A scalable query engine on cloud functions. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 131-141). <https://dl.acm.org/doi/abs/10.1145/3318464.3380609>
- [16] Rabiou, S., Yong, C. H., & Mohamad, S. M. S. (2022). A cloud-based container microservices: A review on load-balancing and auto-scaling issues. International Journal of Data Science, 3(2), 80-92. <https://www.ijods.org/index.php/ds/article/view/45>
- [17] Ricca, F., Marchetto, A., & Stocco, A. (2021, April). Ai-based test automation: A grey literature analysis. In 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW) (pp. 263-270). IEEE. <https://ieeexplore.ieee.org/abstract/document/9440153/>
- [18] Rzadca, K., Findeisen, P., Swiderski, J., Zych, P., Broniek, P., Kusmieriek, J., ... & Wilkes, J. (2020, April). Autopilot: workload autoscaling at google. In proceedings of the fifteenth european conference on computer systems (pp. 1-16). <https://dl.acm.org/doi/abs/10.1145/3342195.3387524>
- [19] Sarker, I. H. (2022). AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN computer science, 3(2), 158. <https://link.springer.com/article/10.1007/s42979-022-01043-x>
- [20] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data, 8(3), 171-188. <https://journals.sagepub.com/doi/abs/10.1089/big.2020.0062>
- [21] 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>