Journal of Engineering and Technological Advances (JETA)
https://jeta.segi.edu.my/index.php/segi
2023 Vol. 8 No. 2

A REVIEW ON THE ROLES OF AI AND MACHINE LEARNING IN OPTIMIZING RESOURCE ALLOCATION, FORECASTING WORKLOAD DEMANDS, AND ENHANCING SECURITY MEASURES IN THE CLOUD

Yossef G.*, Abd Rashid, A.M.

[1,2]School of Information Technology, Faculty of Engineering, Built Environment, and Information Technology, SEGi University, 47810 Petaling Jaya, Selangor, Malaysia.

*Corresponding Author: yo.galal28@gmail.com TEL: (+20) 111780 5122

Abstract: The symbiotic link between artificial intelligence (AI) and cloud computing emerges as a revolutionary force in the continuously evolving world of digital infrastructure. This paper intends to highlight the complicated roles that artificial intelligence (AI) and machine learning have played in reshaping the landscape of cloud computing. In particular, the inquiry tackles three essential areas: cloud resource allocation optimization, workload demand forecast accuracy, and security measure fortification. A fundamental problem as businesses move more and more toward cloud systems is the dynamic resource allocation to manage changing workloads. Underutilization and overutilization lead to inefficiencies that are detrimental for both overall performance and cost-effectiveness.

Keywords: Artificial Intelligent; Cloud Computing; Machine Learning; Resource Allocations; Workload; Security; AWS

1. Introduction

Artificial Intelligence (AI) and machine learning work together symbiotically to revolutionize resource allocation paradigms in the dynamic area of cloud computing. This paper analyses the many roles that AI plays in optimal cloud resource distribution, investigates the complicated dance that cloud environments and AI algorithms conduct and breaking down how these systems dynamically allocate resources in order to reach the greatest possible balance. In addition, one of the most typical difficulties that cloud-based organizations experience is the difficulty of properly estimating future workload needs. Ineffective capacity planning puts operational effectiveness at risk and opens a route for potential service disruptions.

2. Literature Review

The strategic deployment of AI goes beyond theory to become a catalyst for organizations to not only meet the challenges posed by dynamic workloads and security threats, but also to emerge as leaders in the constantly changing cloud computing landscape (Minerva et al., 2017). Benefits from this deployment range from cost-efficiency and improved performance to proactive capacity planning and strengthened security.

2.1 Cloud Resource Allocation Optimization: Amazon Web Services (AWS) provides an interesting case study that exemplifies how AI can optimize resource allocation. AWS dynamically adjusts server instances in response to user behaviour and demand patterns through machine learning methods (Amazon Web Services, 2019). The feature analysed the historical usage data and identifies opportunities to downsize or terminate the underutilised EC2 instances thus this led to cost reductions and achieve significant cost savings.

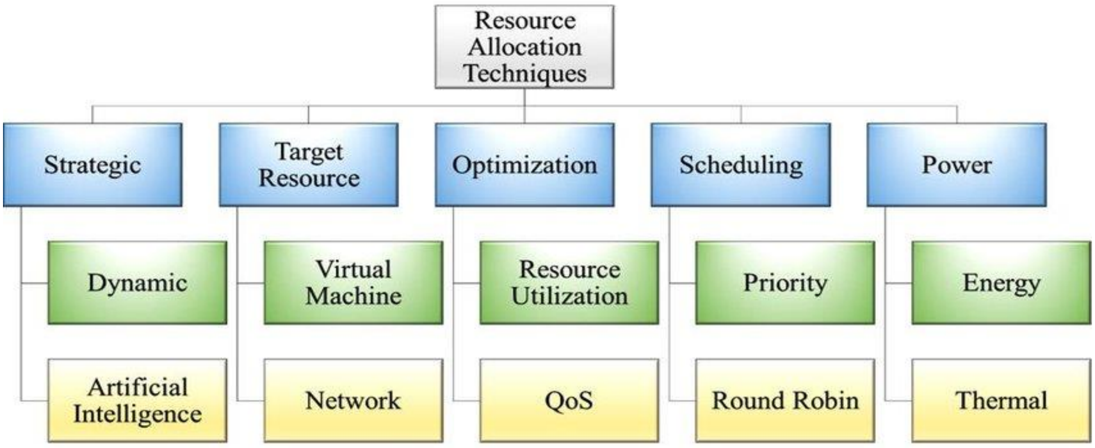

Figure 1. Key components and processes involved in Cloud Resource Allocation Optimization

Referring to **Figure 1** illustrated the five (5) key components and processes involved in Cloud Resource Allocation Optimization. The components include strategic workloads which represent various applications and services running in the cloud. These workloads have different resources requirements for example the CPU, memory, storage where all these depends on their complexity and functionality. The second component is the target cloud resource where it includes the use of virtual machines, containers, storage in Elastic Block Store (EBS) volume, databases and other cloud services.

In the process, the data would be collected continuously to the resource utilization which include the CPU, memory, disk input output and the network traffic. It followed by the resource analysis where the data collected is analysed to identify the patterns and trends in resource usage. The process would help to identify the under-utilised and over provisioned resources.

In optimization procedure, it continued to utilise various algorithms and techniques to recommend optimal resource configurations for the workloads. This included scaling the resources up and down according to the needs of the tasks, changing instances types and utilizing auto scaling features. The process moved forward to scheduling where the resource management executes the recommended changes or actions to optimise resource allocation. This involved provisioning or terminating resources, adjusting configurations or migrating the workloads to different instances.

Lastly in cost optimization, it would analyse the impact of resource optimization on costs. This would help the business to identify the areas to further cost savings and track the overall effectiveness of the optimization efforts.

To recognize the benefits of cloud resource allocation optimization, the process would further reduce the operational expenses by eliminating unnecessary consumption of the resource and optimizing resource utilization. Furthermore, the process would help to avoid the performance bottlenecks, increase the agility with the changing of workload and demand and at the same time reduce the environmental impact of cloud computing by minimizing the resource waste.

2.2 Workload Demand Forecast Accuracy

According to Aibin (2020), workload demand forecasting involves predicting the future resource requirements of a system or task (Aibin, 2020). To have an accurate forecast is very essential for efficient resource allocation, capacity planning, thus ensuring the optimum service quality (Chen, 2020). The AWS services leverage the machine learning algorithms and advanced data analysis techniques to generate the accurate forecast. Accurate workload demand prediction enables organizations to efficiently schedule tasks, ensure appropriate load balancing of cloud resources in a computing architecture, and minimize energy consumption (Karim et al., 2017).

According to McKinsey and Company (2023), machine learning and artificial intelligence technologies could significantly improve demand forecasting accuracy, potentially exceeding traditional methods by 10% to 20% (McKinsey and Company,2023). This could lead to 5% reduction in inventory cost and 2-3% increase in revenue. However, model selection and data quality could significantly impact accuracy (Johnson et al., 2020). Amazon SageMaker used the deep learning model to compute large amount of data and able to produce high accuracy with lower latency and faster training. (Baeldung, 2022).

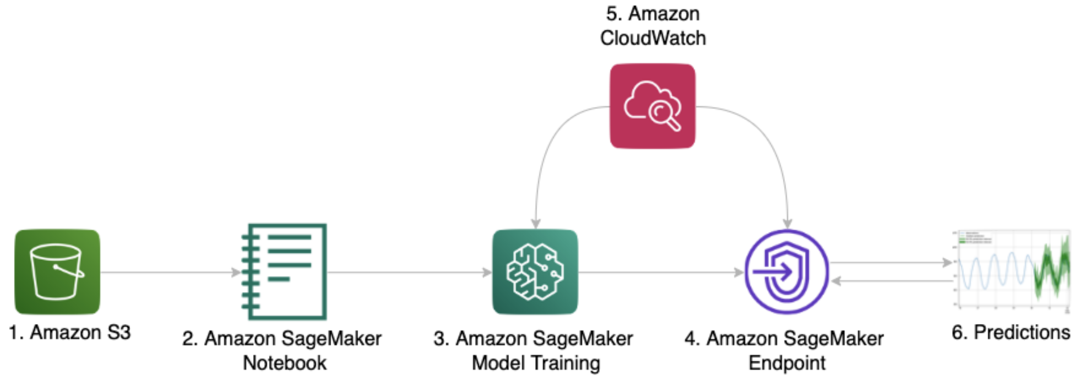
5. Amazon CloudWatch
1. Amazon S3
2. Amazon SageMaker Notebook
3. Amazon SageMaker Model Training
4. Amazon SageMaker Endpoint
6. Predictions

Figure 2. Demand forecasting system using Amazon SageMaker

Based on **Figure 2**, the demand forecasting system consists of the following components such as Amazon S3 to store the historical data and forecasting results. Then the data from Amazon S3 in loaded into Amazon SageMaker Notebook where these data is prepared for building and training the forecasting models that include cleaning, prepossessing and feature engineering and evaluate forecasting results. The Amazon SageMaker Training forecast models at scale and deployed trained forecasting models to production. And the lastly the Amazon SageMaker is used to monitor the system and provide the insights into the forecast performance.

2.3 Security Measure Fortification in the Cloud

With risks ranging from illegal access to data breaches, cloud security is a critical issue. Smith (2021) warns that conventional security measures may not change fast enough to counter new threats. Because of the constantly changing threat landscape, secure cloud infrastructures need quick and proactive reaction techniques (Smith, 2021). AI-powered security systems could identify anomalies, detect and mitigate threats in real-time, and provide intelligent incident response mechanisms to safeguard cloud infrastructure and data.

Unauthorized access and data breaches are two major hazards to cloud security. It's feasible that standard security measures won't keep up with these threats. Real-time threat detection, anomaly identification, and adaptive responses are given by AI-powered security systems (Brown, 2019). AWS has implemented robust security measures to protect the data. Using the Identity and Access Management (IAM), it allowed the users to control who has the access to cloud resources. Data encryption that could encrypt data at both rest and transit thus making it illegible to the unauthorized users. The use of security monitoring and logging to monitor the cloud resources for suspicious activities and lastly the vulnerability scanning and patching that helps to identify and fix vulnerabilities in the cloud resources.

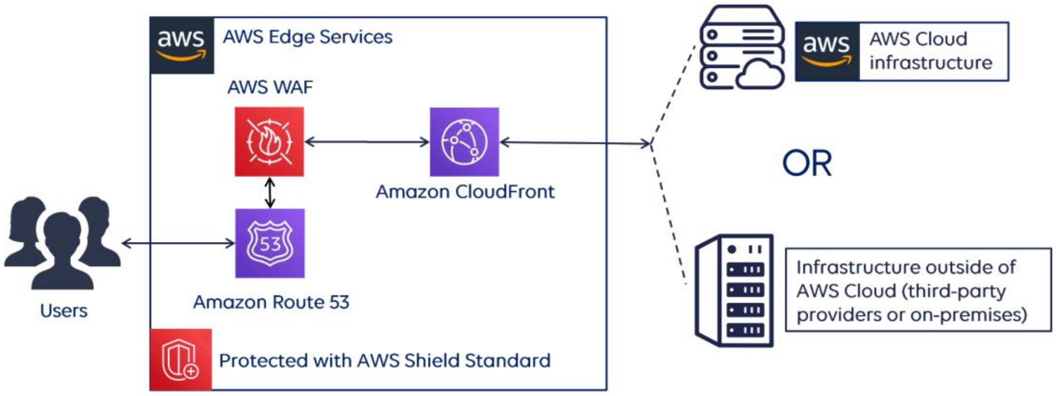
aws
AWS Edge Services
AWS WAF
Amazon CloudFront
Users
53
Amazon Route 53
Protected with AWS Shield Standard
aws
AWS Cloud infrastructure
OR
Infrastructure outside of AWS Cloud (third-party providers or on-premises)

Figure 3. AWS cloud infrastructure with AWS shield for DDoS protection

Based on the **Figure 3** showed AWS Shield could be used to protect AWS cloud infrastructure from DDoS attacks. The AWS Route 53 is highly scalable and available domain name system web service. It could route the internet traffic to any web applications by just translating any domain names into IP address of the underlying web servers. The Route 53 could also be used to create health checks that monitor the availability of the web application. If the checking is failed, Route 53 would automatically re-route the traffic away from the unavailable application.

The Cloud AWF is a web application firewall that helps protect the web applications from the common exploits and attacks. It could be configured to filter out malicious traffic based on the criteria of the source IP address, type of request and the content of the request.

3. Methodologies used by Netflix

Netflix is one compelling case study that has successfully used Artificial Intelligent (AI) and Machine Learning (ML) in optimizing the cloud resources, workload and security. The company delivers video over the Internet and depends heavily in the cloud services where it had to rely on AWS that provides AI powered auto scaling for cost efficiency while meeting with the fluctuating viewership. The auto scalers are able to predicts the demand and traffic patterns and automatically adjusts the usage of virtual machine in real time. As the overprovisions of the resources resulted underutilised resources and excessive costs, while the under provisioning compromised the performances and user experiences. Johnson et al. (2020)

has identified four (4) key integrated auto scaling components which are as follows (Johnson et al., 2020):

i. Data Collection and Integration: The system gathers historical viewership data, real time streaming metrics that includes time of the days, geographic locations and content popularity.

ii. Predictive Modelling: The systems generates forecasts of future viewership patterns by utilizing machine learning algorithms that includes time series analysis, regression techniques and deep learning algorithms

iii. Auto scaling algorithms: The algorithms analyse the predictions and initiates scaling actions either adding or removing virtual machines to ensure sufficient capacity and resources

iv. Continuous Monitoring: The system continually monitors performance and viewership trends thus refining its prediction models and auto scaling decisions.

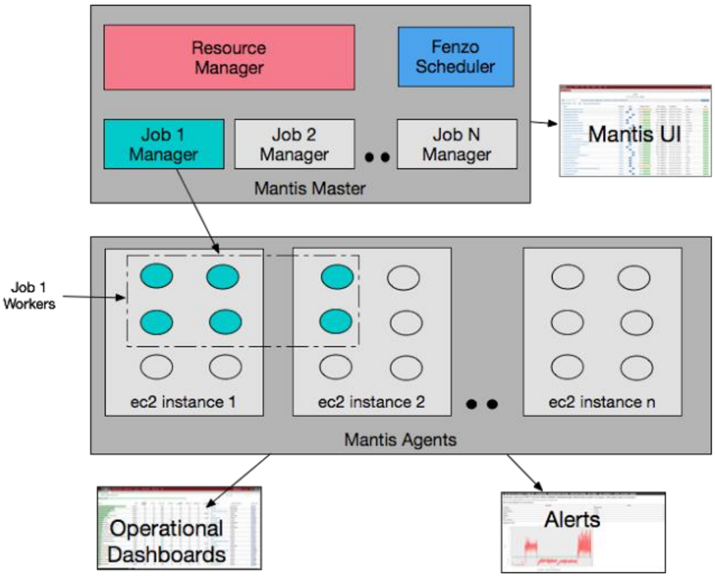

Figure 4. Netflix's distributed resource allocation

Based on the **Figure 4**, it shows in resource allocation employed by Netflix where the resource manager is the key component that is responsible in managing the pool of the resources available that includes the virtual machines, storage and networking. It will keep track of all the available resources and assigns jobs as needed. The Frenzo scheduler is being tasked as containerized workloads that helps in ensuring the jobs are placed on the most appropriate resources based on their requirements and constraints. Mantis Master is used as central components that is responsible to track the status of all the jobs, mana the dependencies between jobs.

While the Mantis UI is the web based interface that allows the users to submit and monitor job and operational dashboards and alerts. The Job Managers are role as interacting with the resource managers to acquire resources and with the Mantis Master to track progress and report any issues and the workers would be used as individual machines or containers that run the tasks to make up the job while being managed by the Mantis Agents. Operational Dashboards are used to provide the real time insights into the performance of the systems, overseeing the resource utilization, job execution times and any errors that have occurred. The alerts would then notify users of any critical issues that might arise such as job failure or in the events the resources are becoming not available.

4. Conclusion

Businesses industry in particularly is heading toward large-scale, intricate automated resource management as a consequence of strategic deployment of AI that goes beyond theory. Machine learning and artificial intelligence technologies substantially accurately allocate resource, increase the accuracy of demand forecasting and enhance cloud systems, making them more effective and versatile while also enhancing their defences against a threat environment.

Leveraging AI's capabilities within cloud computing becomes a crucial instrument for fostering innovation, efficiency, and security in the area of digital infrastructure as enterprises navigate the ever-changing digital world. Cloud platforms could autonomously manage workloads, dynamically scaling resources up and down based on real time needs. This ensures the optimal performance, identify and respond to security threats with exceptional accuracy, thus provides business with more confidence to operate in today’s increasingly complex digital environment.

In essence, the strategic utilization of AI and machine learning empowers the business to achieve unprecedented levels of efficiency, agility and resilience (Kumar & Kumar, 2019). As the businesses continue to explore and harness the power of this synergistic union, we would expect to witness the transformative shifts across industries, propelling the business word into a new era of prosperity and innovation.

Acknowledgement

The authors would like to thank SEGi University for accepting and publishing our paper into the journal.

Conflicts of Interest

"The authors declare no conflict of interest".

References

Aibin, M. (2020). LSTM for Cloud Data Centres Resource Allocation in Software-Defined Optical Networks. In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, New York, pp. 0162–0167

Amazon Web Services. (2019). Elastic Compute Cloud (EC2) Cloud Server & Hosting AWS. Retrieved from https://aws.amazon.com/ec2.

Baeldung. (2022). A Guide to DeepLearning4J. [Online] Retrieved from: https://www.baeldung.com/deeplearning4j.

Brown, L. (2019). AI-powered security measures for cloud infrastructure. *Journal of Cybersecurity*, 25(2), pp.87-102.

Chen, Y. (2020). IoT, cloud, big data and AI in interdisciplinary domains, Simulation Modelling Practice and Theory,102, pp.102070. https://doi.org/10.1016/j.simpat.2020.102070.

Johnson, A., et al. (2020). Predicting workload demands using machine learning in cloud environments. *Proceedings of the International Conference on Artificial Intelligence*, pp.45-57.

Karim, Ahmad & Siddiqa, Aisha & Safdar, Zanab & Razzaq, Maham & Gillani, Syeda & Tahir, Huma & Kiran, Sana & Ahmed, Ejaz & Imran, Muhammad. (2017). Big data management in participatory sensing: Issues, trends and future directions. *Future Generation Computer Systems*. 107. 10.1016/j.future.2017.10.007.

Kumar, P., & Kumar, R. (2019). Issues and challenges of load balancing techniques in cloud computing: A survey. *ACM Computer Survey (CSUR),* 51(6), pp.1–35.

McKinsey & Company. (2023). AI-driven operations forecasting in data-light environments. McKinsey & Company. https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments

Minerva, R., Esposito, M., & Nardone, R. (2017). Cloud-based IoT solutions: Opportunities and challenges. *IEEE Cloud Computing*, 4(1), pp.32-37.

Smith, J. (2021). AI-driven resource allocation in the cloud. *Journal of Cloud Computing,* 15(3), pp.123-135.