International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com
Volume 4, Issue 9, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i9.1047

Cloud as a Platform for AI/ML: Democratization, Services and Architectures

Raghuvar Karthik Durga[1]

Publication Date: 2025/09/30

Abstract

The rapid development of Artificial Intelligence (AI) and Machine Learning (ML) systems necessitates substantial computational resources, specialized tools, and scalable storage solutions. This research examines how cloud computing systems function as an essential infrastructure that supports the current AI/ML revolution. The paper discusses three crucial elements of cloud infrastructure that support AI/ML operations through GPU and TPU acceleration, data lake scalability, AWS SageMaker, Azure Machine Learning, and Google Vertex AI managed services. The paper examines current architectural designs and MLOps life cycles that support automated model development and deployment at scale while ensuring reproducibility. Cloud-based AI systems create a democratizing effect, which enables organizations of every size to access AI technology through cost-effective solutions that eliminate hardware purchase requirements. The paper examines essential problems, including managing costs, protecting data security, and minimizing vendor dependence. The paper delivers a complete assessment of cloud services and their advantages and disadvantages to demonstrate how cloud technology drives AI innovation and accessibility.

Keywords: *Cloud Computing, Artificial Intelligence (AI), Machine Learning (ML), MLOps, Democratization, AWS SageMaker, Azure Machine Learning, Google Vertex AI, GPU, TPU, Scalability.*

I. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) with deep learning have evolved from theoretical research to become a leading force for industrial innovation. The transition to this new era depends on more than just algorithmic progress. The fundamental transformation of computational requirements stands as the basis for this development. Modern AI/ML workloads need substantial resources because they consume computing resources at an extreme rate (Spjuth, 2021). The process of training advanced deep learning models needs enormous computational resources, which standard CPUs cannot handle, thus requiring specialized hardware, including Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs). The performance of these models depends on both the quantity and quality of available data, which demands both access to extensive datasets and effective data management systems. The need for scalable storage solutions emerges from the requirement to handle large datasets (Borra, 2024). The complete process from data preparation through experimentation, training, deployment, and monitoring needs an adaptable infrastructure that provides both strength and unified integration. Most organizations face insurmountable challenges when trying to establish and sustain their own specialized data center infrastructure because it proves too costly and complicated to build. The key challenges of an on-premises approach, as summarized in Figure 1, include high upfront hardware costs, significant ongoing maintenance expenses, and the burden of managing security and backups (Rane, 2024).

Cloud computing emerges as the transformative solution to these problems, as clearly shown in Figure 1. Cloud platforms resolve these issues through their ability to deliver flexible infrastructure that scales automatically while receiving managed services. The cloud platform provides users with immediate access to optimized AI hardware, unlimited storage capacity, and a growing collection of specialized services that handle system intricacies. The new paradigm removes the need for expensive on-premises cluster construction so researchers and businesses can concentrate on innovation instead of managing infrastructure. The main argument of this research demonstrates that cloud computing technology has made AI/ML development accessible to all users. The cloud serves as the primary driver of AI innovation through its comprehensive pay-per-use platform, which enables organizations at every level to contribute to the development of artificial intelligence (Hassan, 2022).

Durga, R. K. (2025). Cloud as a Platform for AI/ML: Democratization, Services and Architectures. *International Journal of Scientific Research and Modern Technology*, *4*(9), 257–265. https://doi.org/10.38124/ijsrmt.v4i9.1047

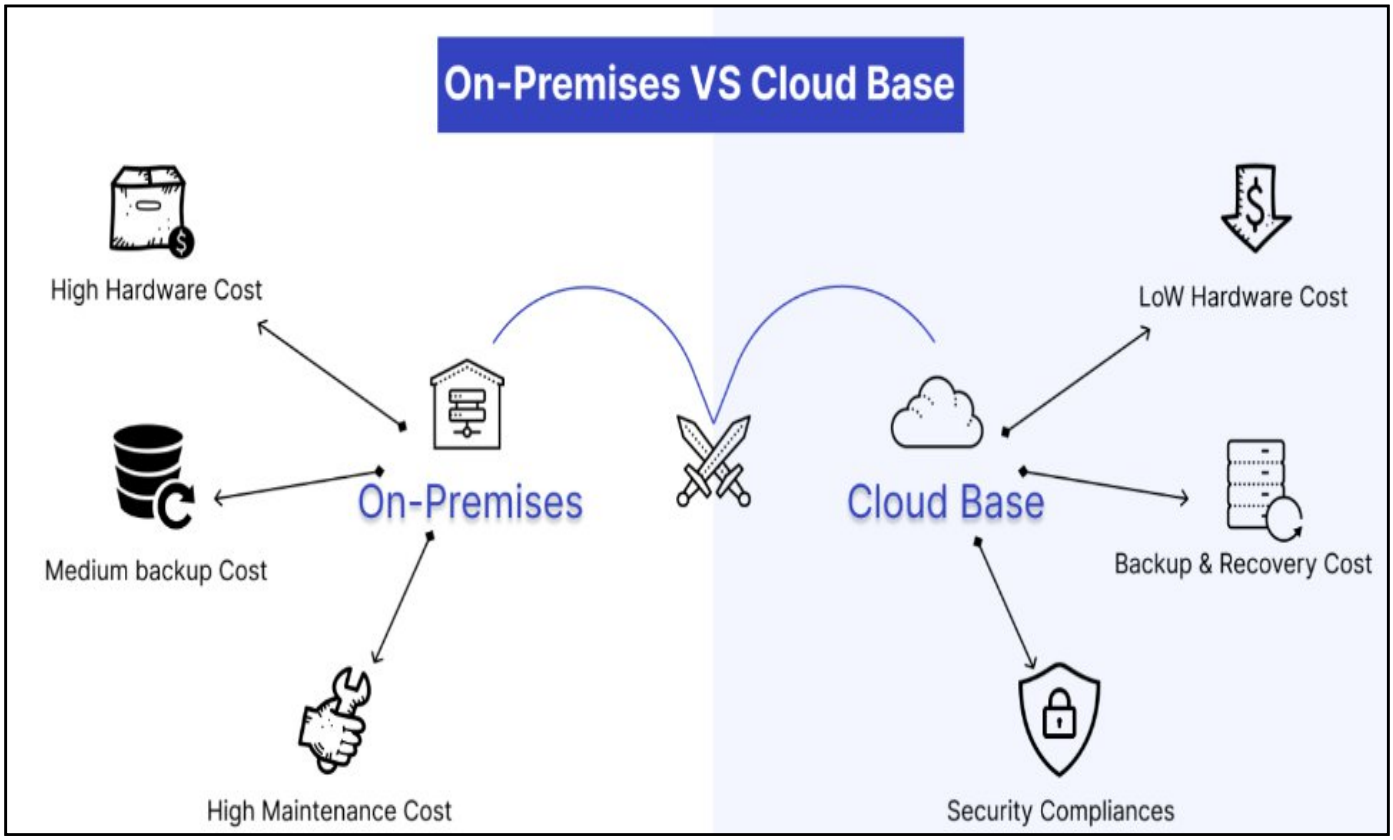

Fig 1 The comparison Between On-Premises Infrastructure Challenges and Cloud-Based Solution Advantages Shows How These Systems Differ in Terms of Management and Cost Requirements.

The following research will provide a complete analysis of this statement. The paper will examine cloud AI/ML platforms through a detailed analysis of their fundamental components, existing architectural designs, and MLOps methods, as well as their advantages for democratization and the associated management and security issues.

II. THE PILLARS OF CLOUD AI/ML PLATFORMS

Cloud providers offer on-demand instances of powerful virtual machines with NVIDIA GPUs (A100 and H100), as well as Google's TPUs and AWS's custom ASICs, all of which are available for immediate deployment. The availability of high-end AI infrastructure through on-demand instances enables organizations to access advanced computing resources without incurring capital expenditures on infrastructure purchases.

➢ *Compute Infrastructure for AI*

- *GPU & ASIC Instances*

The three major cloud providers, AWS and Google Cloud, as well as specialized GPU clouds, provide users with instant access to A100 and H100 GPU instances from NVIDIA for executing machine learning operations. The Multi-Instance GPU (MIG) technology enables users to divide their GPUs into separate instances, which supports efficient resource distribution. The H100 GPU costs more than the H100 but delivers better performance for running extensive AI models. Users can access these resources through web interfaces and APIs for on-demand provisioning (Nurvitadhi, 2016) (Wang, 2023).

Users can access Cloud TPUs from Google through their Google Cloud Platform interface, while researchers receive restricted free access to these resources. The deep learning training and inference capabilities of AWS Trainium and Inferentia chips are available through EC2 instances, which support major ML frameworks and containerized environments (Wang, 2023).

- *High-Performance Computing (HPC) for Distributed Training*

Cloud providers provide HPC clusters that enable distributed training operations across multiple GPUs and ASICs through their specialized infrastructure. The AWS P4d instances with A100 GPUs deliver superior networking capabilities and storage solutions that optimize multi-node training operations. The clusters utilize NVSwitch, EFA, and GPUDirect RDMA technologies to enhance data transfer speeds. The global availability of H100 and A100 GPU clusters from Lambda and HOSTKEY enables developers to access high-performance infrastructure for AI development through cloud services without needing to purchase hardware. Table 1 presents a summary of essential accelerator options together with their deployment locations and main features. The on-demand solutions provide researchers, developers, and independent practitioners from academia and industry with unrestricted access to perform large-scale research and development, eliminating the need to invest in physical infrastructure (Kechriniotis, 2024).

Table 1 Key Cloud AI Accelerators

Accelerator	Cloud Providers	On-Demand Access	Distributed Training	Cost-Effectiveness
NVIDIA A100/H100	AWS, Lambda, Gcore, Google	Yes	Yes (multi-GPU/MIG, NVSwitch)	High (A100 cost-effective, H100 premium)
Google TPU (v4+)	Google Cloud, Colab/TRC	Yes	Yes (TPU clusters)	High (especially spot/preemptible)
AWS Trainium	AWS EC2/SageMaker	Yes	Yes (Ray, tensor parallelism)	Very High (optimized for LLMs)
AWS Inferentia	AWS EC2/SageMaker	Yes	Yes (multi-core, distributed)	Very High (lowest inferencing cost)

➢ *Scalable Data Storage and Management:*

The management of AI data requires AWS S3 and Azure Data Lake Storage for unstructured training data storage, while cloud providers maintain SQL and NoSQL databases for operational structured data management. AWS S3 serves as the foundation for AI data lakes, enabling the storage of petabyte-scale, heterogeneous data. The platform allows for AI service integration while separating compute operations from storage functions, achieving cost optimization through storage tiering and implementing data protection through encryption and access control systems. The platform allows organizations to merge different data sources into a single platform for machine learning operations. Azure Data Lake Storage provides big data analytics and AI workflow support through its scalable cloud storage platform, which accepts various data formats and works with Hadoop systems. The platform maintains data privacy through security and integrates with Azure governance to meet compliance requirements (Mauch, 2013).

Table 2 Data Storage Solutions for Scalable AI

Solution	Example Providers	Use Case	AI Integration	Scalability
Data Lake (Object Store)	AWS S3, Azure ADLS	Raw, massive unstructured training data	Yes (native integration)	Exabyte-scale
SQL Database	Cloud SQL, Amazon RDS	Structured app & feature data, experiment logs	Yes (feature stores)	High
NoSQL Database	Bigtable, DynamoDB, Cosmos DB	Feature stores, real-time analytics, semi-structured data	Yes (vector & feature stores)	Global, low latency

The managed SQL databases, Google Cloud SQL and Amazon RDS, maintain structured data for feature engineering and application logic operations while providing secure high-availability features for scaling purposes. The distributed storage of semi-structured data relies on NoSQL databases, which include Cloud Bigtable, Firestore, DynamoDB, and Cosmos DB. Cloud NoSQL databases function as feature stores because they provide low-latency versioning and automatic sharding and ACID compliance for distributed database operations (Mauch, 2013).

➢ *Core AI/ML Managed Services*

Cloud-based managed services for AI/ML operate as platforms that handle infrastructure management so data scientists can concentrate on developing and deploying their models. The abstraction layer provides operational relief through its built-in tools, automatic scaling, and protection features (Konda, 2025) (Rane N. L., 2024).

• *Value of Managed AI Services*

Managed AI Services provide organizations with essential value through their ability to handle resource provisioning and monitoring operations in the cloud. Benefits include:

✓ *Resource Optimization:*

The platform offers instant access to scalable computing resources, eliminating the need for infrastructure management.

✓ *Operational Efficiency:*

The system provides automated maintenance services, continuous system checks, and fault correction operations.

✓ *Accelerated Workflows:*

The integrated development environments within these platforms enable data scientists to concentrate on model development instead of handling infrastructure tasks (Wang Z. L.-L., XAIport: A Service Framework for the Early Adoption of XAI in AI Model Development, 2024).

• *Amazon SageMaker*

Amazon SageMaker functions as a complete ML platform that supports all stages of the ML development process through its built-in components.

✓ *Built-in Notebooks:*

SageMaker Studio Lab enables users to access pre-configured notebooks that include standard data science packages and GPU/CPU support for development and exploration purposes.

✓ *Automated Model Training (AutoML):*

SageMaker Autopilot performs automatic data preprocessing, algorithm selection, model training, tuning, and model performance optimization through its automated process.

✓ *Hyperparameter Tuning:*

The system enables users to perform extensive distributed hyperparameter searches across multiple compute instances through its managed experiment functionality.

✓ *Model Deployment:*

The platform allows users to deploy models through SageMaker endpoints, which provide automatic scaling, version control, and real-time monitoring features.

SageMaker stands as a mature enterprise-grade ML solution because it integrates data lakes with labeling workflows and model governance features (Nigenda, 2022) (Joshi, 2019).

- *Google Vertex AI*

Google Vertex AI provides users with a unified interface that brings together all machine learning tools from Google:

✓ *Built-in Notebooks:*

The platform features Vertex AI Workbench, which provides managed JupyterLab instances that connect to storage and BigQuery for collaborative development needs.

✓ *AutoML:*

The platform enables users to work with custom code and code-free workflows through Vertex AI, which trains and optimizes models from tabular, image, text, and video datasets.

✓ *Hyperparameter Tuning:*

The system performs extensive search to optimize model hyperparameters, which leads to better accuracy results.

✓ *Model Deployment:*

The system enables users to deploy models for real-time or batch prediction through its built-in endpoint features, which include automatic version control, performance monitoring, and drift detection capabilities.

Vertex AI offers Model Garden, which includes pre-trained and open-source models, including multimodal Gemini, along with monitoring tools, metadata management, and feature stores for big collaborative teams (Wang Z. L.-L., 2024).

- *Azure Machine Learning*

Azure Machine Learning delivers complete cloud-based machine learning operations through its secure and scalable platform:

✓ *Built-in Notebooks and Workspaces:*

The platform includes managed compute instances, shared workspaces, and pre-configured environments, which enable fast model development.

✓ *AutoML:*

The portal enables users to run AutoML for automatic model selection and pipeline generation on tabular, image, and text data.

✓ *Hyperparameter Tuning:*

The system supports distributed hyperparameter tuning and parallel experimentation through native functionality in AutoML and custom pipelines.

✓ *Model Deployment:*

The platform allows users to deploy models through web APIs, Kubernetes, and Edge systems while providing endpoint scaling, model monitoring, and Azure Application Insights integration.

The platform delivers complete MLOps functionality and supports continuous integration and deployment through CI/CD pipelines and secure asset management through Azure Key Vault and Resource Manager (Tirupati, 2022).

The platforms differ from each other through their individual strengths and connection capabilities (SageMaker operates with AWS data services and Vertex AI works with BigQuery and Azure ML functions with Synapse Analytics), yet they all deliver abstraction as their fundamental benefit. The platforms offer a standardized, automated process that connects data to deployment, minimizing operational costs, expertise requirements, and deployment times. The cloud reaches its highest potential as an enabler of AI innovation through this development (Wang Z. L.-L., XAIport: A Service Framework for the Early Adoption of XAI in AI Model Development, 2024) (Rane J. K., 2024).

III. KEY ARCHITECTURAL PATTERNS AND PRACTICES

Cloud service optimization depends on knowledge of contemporary architectural patterns that support scalable, reliable, and efficient machine learning operations. The practices of MLOps unite experimental model development with operationalized AI, generating business value through their combined approach.

➢ *The MLOps Lifecycle on the Cloud*

The ML workflow operates as a recurring process that follows a circular pattern (Figure 2). Cloud platforms offer built-in tools that enable users to automate and control all stages of the process while maintaining reproducibility, scalability, and monitoring capabilities (Barrak, 2022).

- *Data Ingestion:*

The ingestion process requires managed tools to extract data from various sources, including data lakes, streaming services, and databases (Barrak, 2022).

- *Preprocessing:*

The data preprocessing and transformation process happens on Spark clusters, cloud notebooks, and pipeline

orchestration tools, which provide scalable computing capabilities (Das, 2023).

- *Model Training:*

The managed ML platforms Amazon SageMaker, Azure ML, and Vertex AI enable model training through their scalable GPU/TPU cluster infrastructure (Makinen, 2021).

- *Evaluation:*

The automated workflow system conducts validation tests, cross-validation operations, and metric logging to store results for future reproducibility evaluation.

- *Deployment:*

The deployment of validated models occurs through cloud-managed registries, which enable scalable endpoint deployment (SageMaker endpoints and Vertex AI Predictions and Azure ML Endpoints) (Barrak, 2022).

- *Monitoring:*

Cloud services enable continuous model monitoring through their provision of performance metrics, logging functions, and alert systems for detecting model drift and anomalies.

- *Retrain:*

The system utilizes feedback loops and schedulers to perform automated retraining with new data, thereby completing the lifecycle (Barrak, 2022).

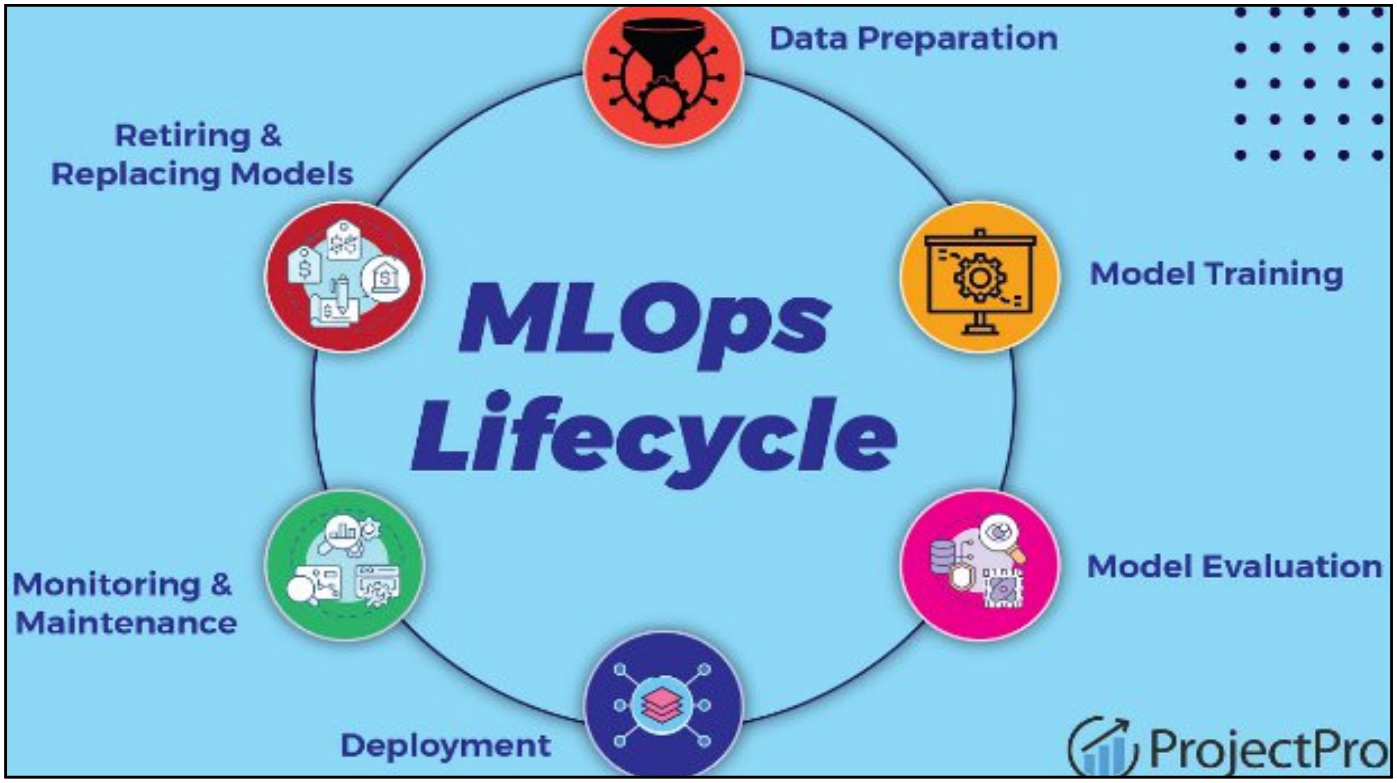

Fig 2 MLOps Lifecycle.

➢ *DevOps for ML: CI/CD Pipelines*

CI/CD automation stands as the core foundation for deploying ML systems in contemporary environments.

- *Code & Model Versioning:*

The system uses GitHub for source control and branch management and tagging functions to track code and pipeline modifications.

- *Automated Testing:*

The testing framework integrates with the system to execute unit tests, integration tests, and data quality tests on model code before deployment.

- *Pipeline Automation:*

The combination of GitHub Actions, AWS Code Pipeline, Google Cloud Build, and Azure DevOps enables automated pipeline operations for model and infrastructure deployment, which produces reproducible results while reducing human involvement.

- *Deployment Gates:*

The deployment process of models between development, staging, and production environments is automated through CI/CD pipelines, which implement approval workflows and integrate model registries.

- *Rollback & Recovery:*

The deployment versioning system enables quick model reversions when system failures occur.

The deployment practices create a system that provides both reliable and traceable ML model deployment in cloud environments (Manolov, 2025) (Bagai, 2024).

➢ *Serverless Architectures for Inference*

Serverless inference provides organizations with affordable, scalable endpoints that require minimal management efforts:

- *Function-as-a-Service:*

The three Function-as-a-Service platforms, AWS Lambda, Azure Functions, and Google Cloud Functions, enable users to create HTTP REST endpoints for real-time prediction by wrapping trained models.

- *Auto-Scaling:*

The system automatically scales its horizontal capacity in response to workload demands. At the same time, it turns off all resources when there is no activity to reduce operational costs and maintenance requirements.

- *Integration:*

Serverless functions enable the connection of feature stores, logging systems, and monitoring dashboards to create flexible event-driven prediction pipelines.

- *Update & Rollout:*

The deployment of code updates and new models happens instantly through CI/CD, while versioned releases enable auditing capabilities.

The architecture design works best for event-based intermittent workloads that include chatbot interactions, fraud detection, and scheduled batch predictions while achieving maximum resource efficiency (Ashutosh, 2020) (Lin, 2018).

IV. DEMOCRATIZING EFFECT OF CLOUD AI

Cloud AI enables democratization through its ability to reduce entry costs while speeding up innovation and delivering sophisticated technologies to users at all levels, including startups, individual researchers, and small organizations. The platforms provide access to top-tier tools and infrastructure, which used to be available only to large enterprises, thus reshaping the way AI research and development operate.

➢ *Lowering Barriers to Entry*

The combination of strong computational resources and advanced ML platforms through cloud-based pay-as-you-go pricing allows startups and independent researchers to use them without needing ample initial financial resources. The pay-as-you-go pricing system of cloud services eliminates hardware expenses, allowing worldwide innovators to create, test, and deploy AI solutions regardless of their organization's size or financial resources. AI development through drag-and-drop interfaces and no-code/low-code tools and user-friendly APIs enables non-experts to participate in AI development, thus expanding industry and skill level participation (Borra, The Evolution and Impact of Google Cloud Platform in Machine Learning and AI, 2024).

➢ *Accelerating Experimentation*

The instant resource deployment and retirement capabilities of Cloud AI platforms enable quick hypothesis testing and multiple development cycles. The following capabilities enable researchers and teams to work on their projects:

- Users can create virtual machines, data lakes, and AutoML workflows through a process that takes only a few minutes.
- The ability to scale up or down infrastructure based on experimental results allows users to test new algorithms and models.
- The pay-per-use model enables users to conduct more experiments because they only pay for what they actually use, thereby minimizing waste and risk.
- The quick deployment and shutdown of resources through this system enables organizations to achieve faster market entry while developing a continuous improvement mindset (Robertson, 2022).

➢ *Access to Advanced Tools and Managed Services*

The ability to access advanced tools and managed services stands as a fundamental aspect of democratization, extending beyond the availability of raw computing power. The software infrastructure needed for distributed training, automated model tuning, and production deployment requires MLOps expertise, which both costs a lot and remains scarce. Cloud providers transform their specialized knowledge into managed services that customers can access. The distributed training libraries of SageMaker and the Pipelines of Vertex AI and Azure's Automated ML simplify complex operations. A small team without DevOps experience can achieve complex hyperparameter tuning and model deployment with canary rollout through the use of managed services. The platform enables small teams to achieve operational excellence and follow best practices, which typically require extensive specialized engineering resources. The platform allows data scientists to concentrate on their main work of solving problems through data analysis instead of handling infrastructure management (Panda, 2024).

V. CHALLENGES AND CONSIDERATIONS

Cloud computing systems require careful management of costs, data governance, vendor independence, and network speed optimization to operate effectively.

➢ *Cost Management:*

Cloud computing systems face significant challenges in cost optimization because their flexible nature leads to unpredictable expenses when not adequately controlled. The flexible nature of cloud services creates unpredictable expenses when organizations fail to implement proper cost management strategies. The pricing structure of pay-as-you-go and subscription-based services affects both system efficiency and cost-effectiveness. Organizations can achieve better resource management through the implementation of spot instances and auto-shutdown features to optimize their resource utilization. The strategies help organizations achieve competitive pricing while maintaining operational efficiency, according to Zhou et al. (2024) and Luong et al. (2017).

➢ *Data Governance and Security:*

The multi-tenant structure of cloud computing requires organizations to prioritize data governance and security measures. The shared infrastructure of cloud computing systems creates challenges for protecting sensitive information while maintaining data privacy and residency standards. Organizations need to implement strong security protocols, including encryption and access management systems, to protect their data while following all applicable regulatory requirements (Patel et al., 2020; Shariati et al., 2015). Cloud platforms require governance systems that provide users with data visibility, trust, and control functions to maintain data integrity and availability (Guo & Song, 2010).

➢ *Vendor Lock-in:*

Cloud-based AI system development faces a significant risk from vendor lock-in, which restricts system flexibility when using proprietary services and APIs from a single provider. The dependency on proprietary services and APIs from one provider creates a situation where switching providers becomes expensive and causes operational problems. The risk of vendor lock-in can be reduced through three strategies, which involve using standardized formats and open-source solutions and deploying systems across multiple cloud platforms for platform-independent operation (Opara-Martins et al., 2016; Petcu, 2011).

➢ *Network Latency:*

Real-time data processing faces challenges because of network latency, which represents the time it takes for data to travel from user devices to cloud infrastructure. Technology faces special challenges when dealing with applications that need fast responses, such as gaming and real-time analytics. The implementation of Mobile Edge Computing (MEC) and fog computing systems helps reduce data transmission distances, thereby decreasing latency and enhancing system performance, according to Ren et al. (2019) and Kumaran and Sasikala (2021).

The existing problems require comprehensive solutions that address cost management, data governance, vendor lock-in, and network latency to achieve optimal utilization of cloud computing resources.

VI. FUTURE TRENDS

The development of TPUs and Trainium chips represents a fundamental step for cloud computing to advance into its future. The neural network processing capabilities of TPUs developed by Google outperform those of standard chip designs in both cloud and edge computing environments, according to Carrión et al. (2024). The specialized chips play a fundamental role in improving AI application performance because they optimize data processing operations, which form the core of machine learning models in distributed systems (Banerjee, 2024). The combination of cloud and edge computing has emerged as a vital development because AI processing now happens at the network edge instead of traditional data center locations. The combination of Edge AI technology enables real-time decision-making on limited resources, making it essential for applications that require quick responses and privacy protection, such as autonomous vehicles and smart home systems (Singh & Gill, 2023; Banjanović-Mehmedović & Husaković, 2023). The integration of these systems enables quicker system responses while reducing the bandwidth-related and latency problems that affect cloud-based processing (Rong et al., 2021).

Cloud operations have become increasingly automated, resulting in improved operational efficiency and lower costs. AI-based resource management systems use machine learning algorithms to perform dynamic resource allocation, performance optimization, predictive maintenance, and anomaly detection for cost reduction (Banerjee, 2024). The management of complex cloud infrastructure depends on AI and automation because these technologies enable better resource utilization and scaling according to demand forecasts, which results in operation optimization and improved system resilience (Mageshkumar et al., 2024). Currently, cloud computing will evolve into a system that uses specialized hardware, edge computing, and automated management to create an environment that supports AI development through enhanced performance and efficiency.

VII. CONCLUSION

The paper demonstrates how cloud computing functions as the essential base for the current AI revolution. The research shows that cloud computing serves as a comprehensive system, enabling efficient AI system development through its specialized computing capabilities, automated services, and architectural frameworks (Joshi, 2019). The cloud provides GPU and ASIC accelerators on demand, streamlining MLOps operations through SageMaker, Vertex AI, and Azure ML, and enables cost-efficient serverless inference deployment to eliminate traditional high costs and operational complexities. The fundamental change in the industry creates a leveling effect that benefits all participants. Cloud platforms create equal opportunities for startups, academic researchers, and enterprises of every business size to join advanced technological development (Makinen, 2021). The pay-as-you-go pricing model, combined with instant resource allocation and user-friendly managed tools with hidden complexity, enables financial accessibility and speeds up development cycles and simplifies complex operations. The AI industry is now showing increased diversity due to this development, leading to enhanced market competition. Cloud computing and artificial intelligence will develop a more interconnected relationship, which will become the dominant pattern in future development. The increasing complexity of AI models requires advanced cloud infrastructure systems that provide enhanced capabilities. AI technology will evolve into the primary system for optimizing and managing cloud operations, resulting in self-operating systems that perform maintenance tasks, security functions, and optimize performance. The continuous positive feedback loop between these

technologies will drive technological progress while establishing the cloud as an essential intelligent system for building future technological frameworks (Mauch, 2013).

REFERENCES

[1]. Shariati, S. M., Ahmadzadegan, M. H., & Abouzarjomehri, A. (2015). Challenges and security issues in cloud computing from two perspectives: Data security and privacy protection. 1078–1082. https://doi.org/10.1109/kbei.2015.7436196

[2]. Zhou, S., Xu, K., Yuan, B., Zhang, M., & Zheng, W. (2024). The Impact of Pricing Schemes On Cloud Computing And Distributed Systems. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (Online), 3(3), 193–205. https://doi.org/10.60087/jklst.v3.n3.p206-224

[3]. Kumaran, K., & Sasikala, E. (2021). Learning based Latency Minimization Techniques in Mobile Edge Computing (MEC) systems: A Comprehensive Survey. 1–6. https://doi.org/10.1109/icscan53069.2021.9526410

[4]. Patel, A., Nayak, A., Ramoliya, D., & Shah, N. (2020). A detailed review of Cloud Security: Issues, Threats & Attacks. 10 3, 758–764. https://doi.org/10.1109/iceca49313.2020.9297572

[5]. Ren, J., He, Y., Li, G. Y., & Yu, G. (2019). Collaborative Cloud and Edge Computing for Latency Minimization. IEEE Transactions on Vehicular Technology, 68(5), 5031–5044. https://doi.org/10.1109/tvt.2019.2904244

[6]. Luong, N. C., Wen, Y., Niyato, D., Wang, P., & Han, Z. (2017). Resource Management in Cloud Networking Using Economic Analysis and Pricing Models: A Survey. IEEE Communications Surveys & Tutorials, 19(2), 954–1001. https://doi.org/10.1109/comst.2017.2647981

[7]. Guo, Z., & Song, M. (2010, August 1). Notice of Retraction: A Governance Model for Cloud Computing. https://doi.org/10.1109/icmss.2010.5576281

[8]. Opara-Martins, J., Tian, F., & Sahandi, R. (2016). Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. Journal of Cloud Computing, 5(1). https://doi.org/10.1186/s13677-016-0054-z

[9]. Petcu, D. (2011). Portability and Interoperability between Clouds: Challenges and Case Study (pp. 62–74). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24755-2_6

[10]. Sanmartín Carrión, D., Prohaska, V., & Diez, O. (2024). Exploration of TPUs for AI Applications (p. 559). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56950-0_47

[11]. Banjanović-Mehmedović, L., & Husaković, A. (2023). Edge AI: Reshaping the Future of Edge Computing with Artificial Intelligence. 133–160. https://doi.org/10.5644/pi2023.209.07

[12]. Mageshkumar, N. V., Mohanraj, A., Viji, C., & Rajkumar, N. (2024). AI-powered financial operation strategy for cloud computing cost optimization for future. Salud, Ciencia y Tecnología - Serie de Conferencias, 3, 694. https://doi.org/10.56294/sctconf2024694

[13]. Banerjee, S. (2024). Intelligent Cloud Systems: AI-Driven Enhancements in Scalability and Predictive Resource Management. International Journal of Advanced Research in Science, Communication and Technology, 266–276. https://doi.org/10.48175/ijarsct-22840

[14]. Singh, R., & Gill, S. S. (2023). Edge AI: A survey. Internet of Things and Cyber-Physical Systems, 3, 71–92. https://doi.org/10.1016/j.iotcps.2023.02.004

[15]. Rong, G., Tong, X., Xu, Y., & Fan, H. (2021). An edge-cloud collaborative computing platform for building AIoT applications efficiently. Journal of Cloud Computing, 10(1). https://doi.org/10.1186/s13677-021-00250-w

[16]. Spjuth, O., Frid, J., & Hellander, A. (2021). The machine learning life cycle and the cloud: implications for drug discovery. Expert Opinion on Drug Discovery, 16(9), 1071–1079. https://doi.org/10.1080/17460441.2021.1932812

[17]. Borra, P. (2024). The Evolution and Impact of Google Cloud Platform in Machine Learning and AI. International Journal of Advanced Research in Science, Communication and Technology, 72–77. https://doi.org/10.48175/ijarsct-18908

[18]. Rane, J., Kaya, Ö., Rane, N. L., & Mallick, S. K. (2024). Artificial intelligence, machine learning, and deep learning in cloud, edge, and quantum computing: A review of trends, challenges, and future directions. Deep Science. https://doi.org/10.70593/978-81-981271-0-5_1

[19]. Hassan, M. B., Mokhtar, R. A., Ali, E. S., Saeed, R. A., Hashim, A. A., & Khalifa, O. (2022). Green Machine Learning for Green Cloud Energy Efficiency. 0, 288–294. https://doi.org/10.1109/mi-sta54861.2022.9837531

[20]. Nurvitadhi, E., Sheffield, D., Sim, J., Mishra, A., Venkatesh, G., & Marr, D. (2016, December). Accelerating binarized neural networks: Comparison of FPGA, CPU, GPU, and ASIC. In 2016 International Conference on Field-Programmable Technology (FPT) (pp. 77-84). IEEE.

[21]. Wang, Y. (2023). Artificial-intelligence integrated circuits: Comparison of gpu, fpga and asic. Applied and Computational Engineering, 4, 99-104.

[22]. Kechriniotis, P. I. (2024). Adaptive Distributed processing on HPC.

[23]. Mauch, V., Kunze, M., & Hillenbrand, M. (2013). High performance cloud computing. Future Generation Computer Systems, 29(6), 1408-1416.

[24]. Konda, S. D. Cloud-Based Ai/Ml Model Deployment: A Comparative Analysis Of Managed and Self-Managed Platforms. Journal ID, 5751, 5249.

[25]. Spjuth, O., Frid, J., & Hellander, A. (2021). The machine learning life cycle and the cloud: implications for drug discovery. Expert Opinion on Drug Discovery, 16(9), 1071–1079. https://doi.org/10.1080/17460441.2021.1932812

[26]. Joshi, A. V. (2019). Amazon's Machine Learning Toolkit: Sagemaker (pp. 233–243). Springer. https://doi.org/10.1007/978-3-030-26622-6_24

[27]. Hardt, M., Vasist, K., Gelman, J., Tsai, E., Liu, X., Cheng, X., Hill, T., Gollaprolu, S., Larroy, P., Chen, X., Kenthapadi, K., Haas, K., Donini, M., Zafar, M. B., Mccarthy, N., He, J., Yilmaz, P., Rathi, A., Das, S., … Siva, A. (2021). Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud. 2974–2983. https://doi.org/10.1145/3447548.3467177

[28]. Wang, Z., Liu, Y., Hamou-Lhadj, A., & Arumugam Thiruselvi, A. (2024). XAIport: A Service Framework for the Early Adoption of XAI in AI Model Development. 30, 67–71. https://doi.org/10.1145/3639476.3639759

[29]. Tirupati, K., Jain, Prof. (Dr.), Khair, M., Goel, O., & Mahadik, S. (2022). Optimizing Machine Learning Models for Predictive Analytics in Cloud Environments. International Journal for Research Publication and Seminar, 13(5), 611–642. https://doi.org/10.36676/jrps.v13.i5.1530

[30]. Rane, N. L., Mallick, S. K., Rane, J., & Kaya, Ö. (2024). Tools and frameworks for machine learning and deep learning: A review. Deep Science. https://doi.org/10.70593/978-81-981271-4-3_4

[31]. Nigenda, D., Tan, A., Ramesha, R., Zafar, M. B., Karnin, Z., Kenthapadi, K., & Donini, M. (2022). Amazon SageMaker Model Monitor: A System for Real-Time Insights into Deployed Machine Learning Models. 104, 3671–3681. https://doi.org/10.1145/3534678.3539145

[32]. Barrak, A., Jaafar, F., & Petrillo, F. (2022). Serverless on Machine Learning: A Systematic Mapping Study. IEEE Access, 10, 99337–99352. https://doi.org/10.1109/access.2022.3206366

[33]. Das, S. D., & Bala, P. K. (2023). What drives MLOps adoption? An analysis using the TOE framework. Journal of Decision Systems, 33(3), 376–412. https://doi.org/10.1080/12460125.2023.2214306

[34]. Makinen, S., Laaksonen, E., Mikkonen, T., & Skogstrom, H. (2021). Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? 109–112. https://doi.org/10.1109/wain52551.2021.00024

[35]. Manolov, V., Gotseva, D., & Hinov, N. (2025). Practical Comparison Between the CI/CD Platforms Azure DevOps and GitHub. Future Internet, 17(4), 153. https://doi.org/10.3390/fi17040153

[36]. Bagai, R., Masrani, A., Ranjan, P., & Najana, M. (2024). Implementing Continuous Integration and Deployment (CI/CD) for Machine Learning Models on AWS. International Journal of Global Innovations and Solutions (IJGIS). https://doi.org/10.21428/e90189c8.9cb39c55

[37]. Ashutosh, T. (2020). AWS SERVERLESS MESSAGING USING SQS. International Journal of Innovative Research in Advanced Engineering, 07(11), 391–393. https://doi.org/10.26562/ijirae.2020.v0711.003

[38]. Lin, W.-T., Xu, W., Krintz, C., Li, T., Wolski, R., Zhang, M., & Cai, X. (2018). Tracking Causal Order in AWS Lambda Applications. 50–60. https://doi.org/10.1109/ic2e.2018.00027

[39]. Borra, P. (2024). The Evolution and Impact of Google Cloud Platform in Machine Learning and AI. International Journal of Advanced Research in Science, Communication and Technology, 72–77. https://doi.org/10.48175/ijarsct-18908

[40]. Robertson, J., Bennett, K., & Fossaceca, J. (2022). A Cloud-Based Computing Framework for Artificial Intelligence Innovation in Support of Multidomain Operations. IEEE Transactions on Engineering Management, 69(6), 3913–3922. https://doi.org/10.1109/tem.2021.3088382

[41]. Panda, D. K., Plale, B., Sadayappan, P., Ramnath, R., Chaudhary, V., Savardekar, N., Tomko, K., Machiraju, R., Fosler-Lussier, E., & Majumdar, A. (2024). Creating intelligent cyberinfrastructure for democratizing AI. AI Magazine, 45(1), 22–28. https://doi.org/10.1002/aaai.12166