# Cloud Platforms for AI: A Survey on Infrastructure, Optimization, and Governance”

**Nishanth S [1] Ranjan V [1] Sayyed Johar [1]**

[1] *Department of Artificial Intelligence and Machine Learning, JNNCE Shivamogga, Visvesvaraya Technological University, Belagavi – 590018, Karnataka, India*

## Abstract

Cloud platforms have increasingly matured into end-to-end ecosystems for artificial intelligence (AI), enabling the development, deployment, and governance of AI systems at scale. This survey paper reviews recent advancements across seven key domains: (i) hyperscaler AI infrastructure and hardware accelerators, (ii) managed foundation model (FM) platforms and agent tooling, (iii) AIOps and observability frameworks for AI workloads, (iv) vector databases and Retrieval-Augmented Generation (RAG) techniques, (v) privacy and security with a focus on confidential computing, (vi) financial operations and cost governance for AI (FinOps), and (vii) sustainability considerations in light of energy and power constraints. The survey consolidates state-of-the-art practices, identifies emerging trends, and discusses unresolved challenges, particularly in achieving multi-cloud portability, ensuring robust data governance, and addressing the economic and environmental externalities associated with large-scale AI adoption.

**Keywords—** Cloud computing, Artificial intelligence, Foundation models, AIOps, Vector databases, Retrieval-Augmented Generation (RAG), Confidential computing, FinOps, Sustainability, Multi-cloud portability, Data governance**1.**

## 1. Introduction

The AI surge has turned cloud platforms into de facto AI operating systems. Hyperscalers now combine specialized compute (GPUs/TPUs/custom silicon), orchestration layers, managed FM endpoints, agent frameworks, vector stores, and integrated observability. At the same time, enterprises face a second-order problem: how to operate, secure, and pay for AI workloads sustainably. This review focuses on evidence and announcements through August 28, 2025 (IST), highlighting widely-used services and independent market/ESG findings.

**Hyperscaler AI Infrastructure: Specialized Silicon and Capacity Ramps**

Cloud AI is constrained by compute, memory, and power. Providers expanded fleets of H100/H200-class GPUs, introduced “AI factories” (e.g., NVIDIA DGX Cloud) and signaled massive capex ramps. DGX Cloud positions a unified AI platform available on leading clouds; NVIDIA promotes DGX H200 and Blackwell-era SuperPODs as enterprise “AI factories.” NVIDIA+1.

Market reporting in 2025 shows continued hyperscaler capex acceleration to meet AI demand; analysts note AWS boosting AI/data-center build outs to catch up with Azure/Google on capacity—a barometer for anticipated gen-AI workloads. Investors GPU pricing and memory footprints matter for LLM scale: H200’s larger HBM3e memory enables hosting larger models per device, influencing cloud price/perf tradeoffs.

**Implication.** Capacity and silicon roadmaps shape feasibility and unit economics for training/fine-tuning and high-QPS inference. Organizations planning 2025–2027 AI programs must track region/model availability, accelerator types, and queue times.

**Managed Foundation-Model Platforms and Agent Tooling (Bedrock, Vertex AI, Azure OpenAI)**

Clouds now deliver FM access as managed endpoints plus orchestration, evals, and guardrails:

- **AWS Bedrock.** Offers a model garden (multiple providers), knowledge bases, evaluations, guardrails (with automated reasoning), and expanding regional coverage. AWS continues to refresh supported FMs and document lifecycle management.
- **Google Cloud Vertex AI.** Regular model updates (e.g., Gemini 2.5 family) and a formalized path for **AI agents** via Vertex AI Agent Builder (multi-agent orchestration). Release notes in mid-2025 detail Gemini 2.5 GA endpoints and rebranding of “Agent Builder” under “AI Applications.”

- **Azure OpenAI (Azure AI Foundry).** Continual model refresh/retirements and updated API lifecycles; 2025 notes highlight new image models and the latest GA API version. Independent reporting points to broader availability of next-gen GPT suites on Azure.

**Implication.** The platform "stack" has shifted from raw model access to governed agentic systems with evaluation, safety, and lifecycle tooling—crucial for enterprise rollout and compliance.

**2 AIOps, Observability, and SRE for AI Era**

As AI workloads scale, enterprises adopt **AIOps** to automate incident detection, capacity forecasting, and root-cause analysis across distributed systems. 2025 market analyses show rapid AIOps growth (20%+ CAGR projections). Thought leadership emphasizes predictive analytics, AI-driven observability, and edge/distributed monitoring—aligned with modern microservice and data pipeline sprawl.

**Trend.** Cloud providers and ISVs integrate LLMs into logging/metrics/tracing to summarize incidents, propose remediations, and auto-generate runbooks—reducing MTTR but raising governance/guardrail needs.

**2. Vector Databases, RAG, and Application Patterns**

RAG remains the leading pattern to ground models on enterprise data. Clouds now document **managed vector options** (e.g., OpenSearch, pgvector on managed PostgreSQL, Kendra) and partner solutions; the independent ecosystem evaluates performance, scale, and cost across specialized and multi-model stores.

Recent practitioner guides compare vector DBs for 2025 RAG workloads and emphasize embeddings throughput, filtering, hybrid search (sparse+dense), and operational cost.

**Emerging practice.** Hybrid search + structured context (via SQL/graph joins) and **chunking/ATC** (adaptive text chunking) improve retrieval quality; vector stores integrate with agent frameworks for tool-use and memory.

**3. Security & Privacy: Confidential Computing and Data-in-Use Protection**

With regulated and proprietary data entering AI workflows, **confidential computing**—hardware trusted execution environments (TEEs) for data-in-use protection—has moved into mainstream cloud SKUs (e.g., Azure confidential VMs, services). Microsoft documents hardware-enforced isolation for VMs/containers to mitigate cloud-operator access and hypervisor threats.

**Implication.** Confidential GPUs/accelerators and enclave-friendly toolchains are becoming table stakes for sensitive fine-tuning and inference, complementing encryption at rest/in transit and KMS integrations.

**4. Cost Governance: From Cloud FinOps to AI FinOps**

Organizations report exploding AI line items: model/API calls, vector-store I/O, GPU hours, agent pipelines, and data egress. The **State of FinOps 2025** emphasizes persistent priorities around workload optimization, waste reduction, and improved allocation/visibility, with growing attention on AI spend governance. Summaries from multiple practitioners echo the shift from one-off savings to continuous governance and show rising demand for granular AI cost attribution.

**Emerging capabilities:** Cloud-native recommendations increasingly use AI to detect idle/over-provisioned resources across serverless, containers, and databases, and now extend to AI services (e.g., managed LLM endpoints, vector DBs). FinOps Foundation

**Practical note.** Compute prices for advanced GPUs (e.g., H200) materially impact TCO and architecture choices (quantization, distillation, batching, LoRA vs. full fine-tune), reinforcing the need for AI-aware capacity/throughput planning. Jarvislabs.ai Docs

**Sustainability, Power, and Grid Constraints**

AI raises data-center energy and water scrutiny. Structure Research's 2025 ESG analysis finds renewable sourcing rising among hyperscalers, with estimates around ~90% renewable share for leading platforms, while overall operator averages lag; industry coverage spotlights increased total TWh use but improved carbon intensity. Deloitte forecasts data centers at ~2% of global electricity in 2025 (contextualizing public concern). Google reports a trailing-12-month fleet PUE of 1.09 at stable operations. News and policy coverage highlight grid constraints in hot markets and evolving utility mechanisms.

**Implication.** Site selection increasingly prioritizes power availability and carbon intensity; AI workload scheduling, liquid cooling, and on-site generation/storage (including batteries) are becoming strategic levers. Independent analyses also note storage/UPS evolution to meet AI load profiles. ZincFive | Nickel-Zinc Batteries

**5.Agentic and Multimodal Futures**

Platform roadmaps emphasize **agent frameworks** (multi-tool, multi-step workflows), multimodal generation/editing, and long-context models. In 2025, Google's Gemini 2.5 line reached GA in Vertex AI with expanded agent tooling; Azure updates surfaced new model capabilities and lifecycles; AWS added Bedrock evaluation frameworks and guardrails improvements. Collectively, these signal a normalized path to production for **agentic** enterprise apps (task orchestration, tool use, retrieval, workflow memory), integrated with security and observability stacks. Google Cloud+1Microsoft Learn+1AWS Documentation

**6. Data Platforms: Convergence with AI Workloads**

The data layer is being reshaped by AI demand (streaming for real-time features, lake house architectures, and cross-cloud pipelines). Investor and market coverage underscores how AI acceleration boosts demand for modern data platforms and cross-vendor competition (e.g., Snowflake, Databricks). Reuters

**Trend.** Expect continued convergence: vector-native features inside data warehouses and SQL engines; unified governance catalogs spanning files, tables, and embeddings; and tighter cost controls across data + AI pipelines.

## Open Challenges and Research Directions

1. **Portability and lock-in.** Model, agent, and vector APIs still vary across clouds; "model gardens" help, but migration costs remain high. Standardization across evals, safety policies, and agent protocols is nascent. (See vendor lifecycle docs and agent-platform updates.)
2. **Safety & governance.** Guardrails (input/output filtering, grounding checks, fact-verification) are improving; AWS notes automated reasoning checks in 2025 updates. Formal methods, audit trails, and reproducibility for agentic systems remain research targets. AWS Documentation
3. **Observability for LLMs.** Aligning SRE with model-quality metrics (hallucination, latency, cost/token), plus data drift detection for RAG indices, is still emerging practice. (AIOps sources forecast this convergence.)
4. **Cost & performance engineering.** Pricing for advanced accelerators, token/embedding costs, and vector-I/O create complex trade spaces; FinOps for AI is shifting from reporting to **continuous optimization**.
5. **Sustainability & siting.** Power constraints and water/cooling externalities will shape region choices and SLAs; expect more disclosures and policy linkage. ReutersData Center POST

**Table 1: Comparison of Recent Survey Papers on Cloud with Artificial Intelligence**

| No. | Title (Authors, Year) | Venue / Publication | Key Focus / Trends Reviewed | Reference |
|---|---|---|---|---|
| 1 | **Edge-Cloud Collaborative Computing on Distributed Intelligence and Model Optimization** (Liu, Wang, Zhang, & Chen, 2025) | arXiv preprint (2025) | Edge-cloud AI systems; model compression & optimization; orchestration; latency/energy tradeoffs; benchmarking | Liu et al. (2025), arXiv:2505.01821 |
| 2 | **AI-Driven Security in Cloud Computing** (Shaffi, Alazzawi, & Shuaib, 2025) | arXiv preprint (2025) | AI-based cloud security; automated detection/response; predictive analytics; resilience, bias, compliance | Shaffi et al. (2025), arXiv:2505.03945 |
| 3 | **AI-Driven Innovations in Modern Cloud Computing** (Kumar, 2024) | arXiv preprint (2024) | Intelligent resource management; predictive analytics; automated deployment/scaling; cost reduction; governance | Kumar (2024), arXiv:2410.15960 |
| 4 | **Survey on Cloud Computing Integrated with Artificial Intelligence** (Sadargari & Balaji, 2023) | Int. J. on Recent & Innovation Trends in Computing & Communication (IJRITCC) | Overview of AI-cloud integration; data management; parallel/distributed training; scaling; optimization; deployment | Sadargari & Balaji (2023), IJRITCC |
| 5 | **AI and Computing Horizons: Cloud and Edge in the Modern Era** (Editorial Team, 2024) | *Journal of Sensor and Actuator Networks*, 13(4) | Synergistic roles of cloud, edge, and AI in IoT; efficiency via edge intelligence; future cloud-edge AI design | JSAN Editorial Team (2024) |
| 6 | **An Overview on Generative AI at Scale with Edge-Cloud Computing** (Wang, Xue, Wei, & Kuo, 2023) | arXiv preprint (2023) | Edge-cloud collaboration for scalable generative AI; latency challenges; system design considerations | Wang et al. (2023), arXiv:2306.17170 |
| 7 | **A Survey of Machine Learning in Edge Computing: Techniques, Frameworks, Applications, Issues, and Research Directions** (Jouini, Sethom, Namoun, Alanazi, & Alanazi, 2024) | *Technologies*, 12(6) | ML deployment at edge; frameworks, hardware; applications in IoT; security, scalability, open challenges | Jouini et al. (2024), *Technologies* |
| 8 | **The Security and Privacy of Mobile Edge Computing: An Artificial Intelligence Perspective** (Wang, Yuan, Zhou, Xu, Li, & Wu, 2024) | arXiv preprint (2024) | AI-based security & privacy in MEC; ETSI MEC standard; SDN/NFV frameworks; attack detection & defense | Wang et al. (2024), arXiv:2401.01589 |
| 9 | **A Survey on Software-Defined Network-Enabled Edge Cloud Networks: Challenges and Future Research Directions** (Kazi, Islam, Siddiqui, & Jaseemuddin, 2025) | *Network*, 5(2) | SDN-enabled edge-cloud; programmability; orchestration; future research gaps in cloud networking | Kazi et al. (2025), *Network* |

## Conclusion

This survey has traced the evolution of cloud platforms from bespoke AI infrastructure toward integrated enterprise ecosystems capable of supporting the full lifecycle of artificial intelligence. The trajectory indicates that the next phase will be shaped by agentic systems, advances in confidential computing, convergence of vector and SQL data management, AIOps-driven observability, and FinOps-grade governance for both cost and sustainability. However, realizing this vision requires more than technological progress alone. Research and practice must jointly establish common standards for portability across multi-cloud environments, implement robust guardrails for security and compliance, and develop reliable frameworks for measuring efficiency, cost, and environmental impact. Only through such coordinated efforts can cloud platforms enable AI workloads that are not merely scalable, but also safe, trustworthy, affordable, and sustainable—thereby ensuring long-term value for enterprises, researchers, and society at large

## References


[1]. Liu, J., Wang, X., Zhang, Y., & Chen, H. (2025). *Edge-Cloud collaborative computing on distributed intelligence and model optimization: A survey*. arXiv preprint arXiv:2505.01821. https://arxiv.org/abs/2505.01821

[2]. Shaffi, S. M., Alazzawi, A. K., & Shuaib, K. (2025). *AI-driven security in cloud computing: A survey*. arXiv preprint arXiv:2505.03945. https://arxiv.org/abs/2505.03945

[3]. Kumar, A. (2024). *AI-driven innovations in modern cloud computing: A survey*. arXiv preprint arXiv:2410.15960. https://arxiv.org/abs/2410.15960

[4]. Sadargari, V., & Balaji, N. A. (2023). *Survey on cloud computing integrated with artificial intelligence*. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), 11*(9), 4440–4445. https://doi.org/10.17762/ijritcc.v11i9.9936

[5]. Sens Actuator Netw. Editorial Team. (2024). *AI and Computing Horizons: Cloud and Edge in the Modern Era*. *Journal of Sensor and Actuator Networks, 13*(4), 44. https://doi.org/10.3390/jsan13040044

[6]. Wang, C.-C. J., Xue, J., Wei, C., & Kuo, C.-C. J. (2023). *An overview on generative AI at scale with edge-cloud computing*. arXiv preprint arXiv:2306.17170.

[7]. Jouini, O., Sethom, K., Namoun, A., Alanazi, M. H., & Alanazi, M. N. (2024). *A survey of machine learning in edge computing: Techniques, frameworks, applications, issues, and research directions*. *Technologies, 12*(6), Article 81. https://doi.org/10.3390/technologies12060081

[8]. Wang, C., Yuan, Z., Zhou, P., Xu, Z., Li, R., & Wu, D. O. (2024). *The security and privacy of mobile edge computing: An artificial intelligence perspective*. arXiv preprint arXiv:2401.01589.

[9]. Kazi, B. U., Islam, M. K., Siddiqui, M. M. H., & Jaseemuddin, M. (2025). *A survey on software-defined network-enabled edge cloud networks: Challenges and future research directions*. *Network, 5*(2), 16. https://doi.org/10.3390/network5020016