theCUBE research

# Orchestrating the AI Future and How Rafay Powers Enterprise-Grade GPU Infrastructure

## A CIO's Guide to Scalable, Compliant, and Developer-Ready AI Deployment

ANALYST WHITE PAPER

Paul Nashawaty, Principal Analyst
and Samantha Weston,
Industry Analyst

May 2025

# Executive Summary

Enterprise adoption of artificial intelligence is now operational and not just aspirational. As organizations race to embed AI into products, processes, and decision-making, the infrastructure required to support AI workloads must evolve just as rapidly. Modern AI initiatives demand more than raw compute power. They need integrated, scalable platforms capable of managing complex GPU environments, empowering developers, and meeting stringent regulatory demands.

GPU Platform-as-a-Service (GPU PaaS) has emerged as a foundational strategy in this evolution. Enterprises need a unified platform that orchestrates CPU and GPU workloads, enabling seamless deployment of AI models across hybrid and sovereign environments. Dynamic resource allocation, AI-native CI/CD pipelines, integrated MLOps, and policy-driven automation are critical to delivering performant and compliant AI systems.

Rafay System's (Rafay) platform meets these needs by providing centralized orchestration that bridges infrastructure and innovation. It enables developer velocity through built-in automation, supports regulatory compliance through sovereign AI capabilities, and delivers operational agility across multi-cloud and hybrid deployments. Backed by strategic partnerships with leaders like Nvidia, Rafay helps enterprises deploy AI at scale, without compromising speed, security, or governance.

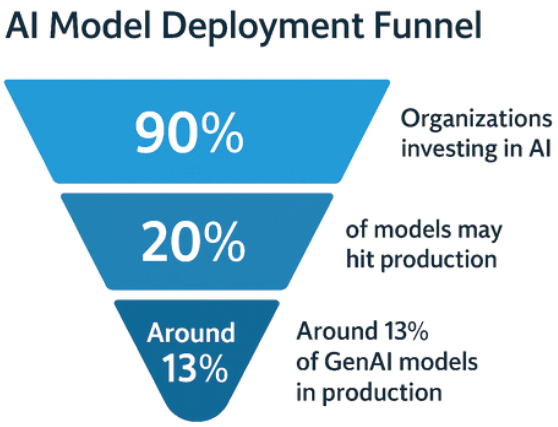
As AI maturity deepens across industries, CIOs seek infrastructure strategies that keep pace and provide competitive differentiation. Organizations that manage complexity, maintain compliance, and bring AI to market faster will define the next wave of digital transformation. Rafay provides the platform to lead in that future.

# The AI Infrastructure Challenge

Enterprise demand for AI is surging, with theCUBE Research finding that over 90% of organizations are actively investing in AI initiatives. However, the journey from experimentation to production remains filled with obstacles. Roughly 80% of AI models never progress beyond the prototype phase, primarily due to deployment challenges rather than technical inefficiencies.

Even though Kubernetes has become the standard for container orchestration, it wasn't designed with GPU workloads in mind. Managing AI-specific pipelines introduces complexities such as multi-architecture support, distributed training, inference scaling, and hardware-level optimization. Most teams lack the automation to effectively provision, monitor, and scale GPU clusters across environments.

The challenges extend to generative AI implementations as well. A recent survey revealed that 68% of organizations have moved less than 30% of their generative AI experiments into production. An additional survey found that 86% of organizations require tech stack upgrades to deploy AI agents effectively.




Organizations are under increasing pressure to demonstrate real-time value from AI investments. AI has quickly transitioned from a research project to a core business enabler. CIOs are tasked with shortening development cycles, empowering developers, and scaling applications that leverage large language models (LLMs) and predictive algorithms. Meeting these expectations requires a shift in infrastructure mindset with a strong focus on scalable, efficient, and compliant AI deployment strategies.

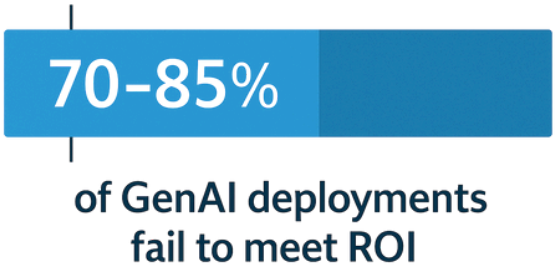
## Developer Velocity and Integration

The infrastructure is essential, but the people building on it are equally important. A consistent challenge in enterprise AI is the failure to operationalize models. Research suggests that nearly 80 percent of AI models never reach deployment, often due to a lack of MLOps maturity and disconnected workflows between development and operations. Supporting this, a recent report indicates that between 70% and 85% of generative AI deployment efforts fail to meet their desired ROI. This data shows there remains a critical need for effective operational strategies.

Rafay addresses this disconnect with native support for AI-focused CI/CD pipelines and MLOps integration. Through Rafay, developers are provided with prebuilt

workflows, security guardrails, and policy-based automation that accelerate deployment without introducing risk. This approach significantly improves time-to-value while minimizing the rework often required to productionize models.

The developer experience is paramount. By reducing friction across the build, train, and deploy phases, organizations can transform AI from an R&D experiment into a reliable and repeatable process that integrates directly into broader DevOps ecosystems.

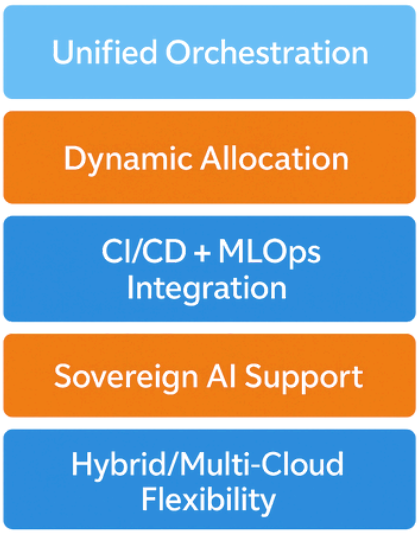



## GPU PaaS and the Rise of Cloud-Native AI Infrastructure

To meet the demands of modern AI workloads, enterprises are turning to GPU Platform-as-a-Service (GPU PaaS) solutions as a strategic enabler. Unlike traditional infrastructure models, GPU PaaS abstracts the complexity of managing GPU hardware, allowing development and operations teams to focus on accelerating AI deployment rather than troubleshooting orchestration or provisioning delays.

GPU PaaS delivers several strategic advantages:

- Unified orchestration across CPU and GPU environments for consistent workload management.
- Dynamic resource allocation allows enterprises to optimize performance and cost across training and inference workloads.
- Developer-first tooling through self-service deployment environments, integrated CI/CD pipelines, and GPU-aware scheduling policies.



As organizations scale AI models, the need for agile infrastructure becomes mission-critical. GPU PaaS enables that agility, delivering the performance of dedicated GPU infrastructure with the elasticity of the cloud. For organizations navigating hybrid or multi-cloud strategies, this abstraction is key to consistency, scalability, and developer productivity.

## Compliance, Sovereignty, and Strategic AI Expansion

The need for strong governance and regulation also expands as AI capabilities expand. To date, more than 50 countries have enacted or proposed legislation impacting data privacy, AI ethics, and model governance. The emergence of "Sovereign AI" is a response to this global regulatory pressure.

Sovereign AI allows organizations to deploy private AI clouds within national borders, ensuring compliance with data residency and security requirements. But sovereignty demands infrastructure flexibility. Organizations must be able to deploy GPU-enabled environments across private data centers, regional cloud providers, and edge locations without compromising orchestration or security.

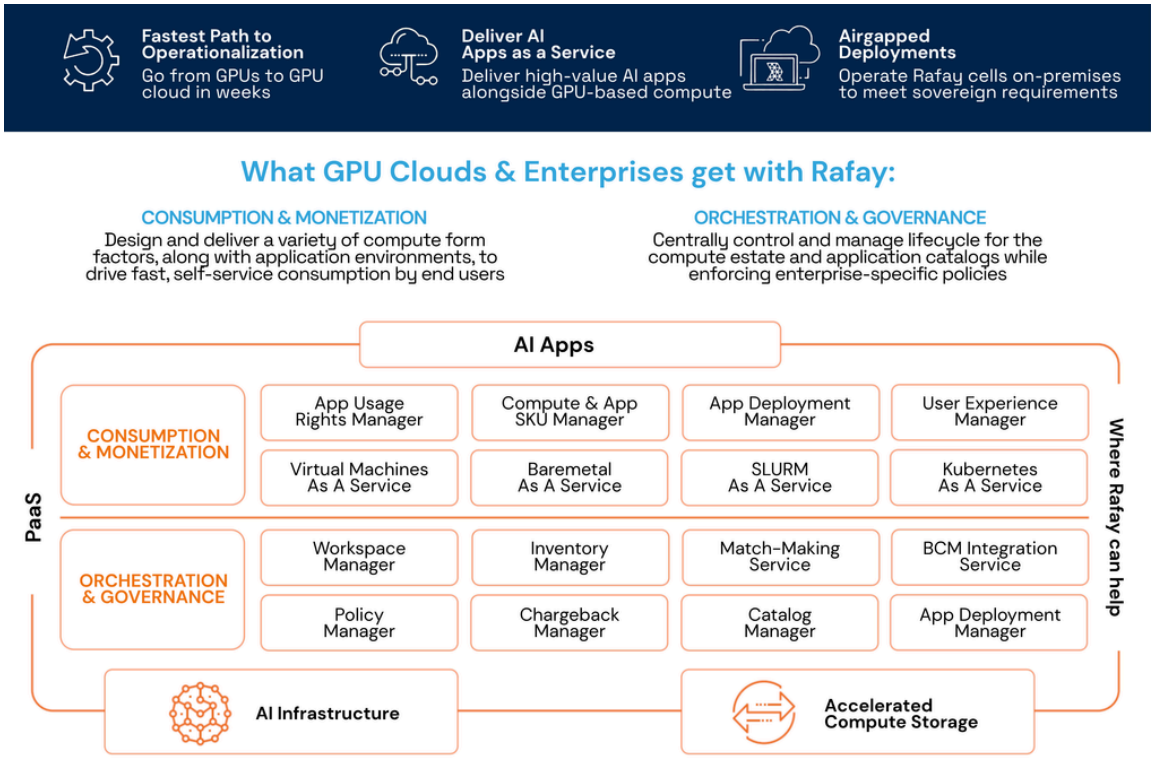
Rafay enables this level of control. Its platform supports hybrid and multi-cloud deployments, including sovereign environments. This allows CIOs to align their AI strategies with national regulatory frameworks while maintaining global scalability and agility.

## Rafay's Platform in Practice

Rafay's cloud-native orchestration platform is designed to manage the full lifecycle of enterprise AI workloads. The solution is engineered for scalability, developer usability, and compliance while delivering on the following core capabilities:

- Seamless GPU PaaS Management: Rafay provides a unified platform for orchestrating AI workloads across CPU and GPU environments, simplifying deployment and scalability. By delivering GPU Platform-as-a-Service (GPU PaaS) capabilities, Rafay abstracts the complexity of GPU infrastructure, allowing enterprises to dynamically allocate resources, optimize cost-performance, and scale AI models without manual provisioning. This enables teams to deploy AI applications with the speed and flexibility of cloud-native environments, regardless of underlying hardware.

- Accelerated AI Application Deployment: By automating Kubernetes-based AI infrastructure management, Rafay reduces operational complexity and enables faster time-to-market for AI models.
- Developer-Friendly Integration: The platform offers the ability to package and deliver a self-service experience as needed.
- Hybrid and Multi-Cloud Flexibility: Rafay supports diverse cloud environments, enabling enterprises to deploy AI workloads seamlessly across public, private, and sovereign cloud infrastructures.
- Regulatory Compliance and Data Sovereignty: The platform helps organizations adhere to AI-related regulations by facilitating sovereign AI cloud deployments within national borders.
- Strategic Partnerships and Ecosystem Expansion: Collaborations with industry leaders like Nvidia enhance Rafay's ability to deliver optimized GPU orchestration solutions.
- Enterprise Use Cases and Success Stories: Rafay has demonstrated proven AI transformation results across multiple industries, reinforcing its leadership in cloud-native AI infrastructure.

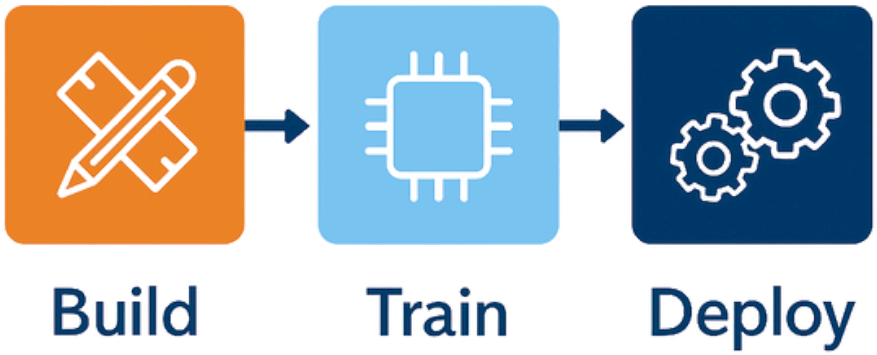


In practice, Rafay's platform has powered AI transformations in industries ranging from financial services to healthcare. These implementations demonstrate measurable gains in deployment speed, developer productivity, and infrastructure efficiency.

## Future Outlook for AI and GPU Orchestration

The supporting infrastructure must evolve in lockstep as enterprise AI becomes more advanced. There are several key trends shaping this evolution:

- AI Infrastructure Will Outgrow CPU-Only Models: GPUs, and increasingly, specialized accelerators, will be essential for training and inferencing complex models. Unified orchestration tools will be critical to managing heterogeneous environments.
- Compliance Will Drive Deployment Decisions: The need to comply with data protection and AI governance laws will influence how and where organizations deploy AI workloads. Sovereign AI is not optional.
- Developer Enablement Will Be a Market Differentiator: Organizations that streamline AI development and deployment will outperform those that silo experimentation. MLOps, CI/CD, and policy automation are not nice-to-haves; they're competitive necessities.




CIOs should prioritize platforms that simplify orchestration, enable compliance, and empower development teams. Investments in cloud-native, GPU-ready infrastructure will be foundational.

# A Strategic Path Forward

Organizations face unprecedented levels of opportunity and pressure as AI adoption accelerates. Although challenges remain, especially in scaling infrastructure and complying with regulations, solutions like Rafay's cloud-native orchestration platform are closing the gap between AI ambition and AI execution.

## Key Takeaways

- Infrastructure is the Backbone of AI Success: Organizations must align hardware and orchestration strategies to support evolving AI workloads.
- GPU PaaS Enables Flexibility and Scale: Unified platforms simplify operations across CPU and GPU environments.
- Compliance Is Driving Sovereign AI Demand: Organizations need infrastructure that adapts to geopolitical and regulatory realities.
- Developer Experience Determines Time-to-Value: Integrated MLOps and CI/CD workflows reduce friction and accelerate production.
- Rafay Delivers a Strategic Advantage: With hybrid support, automation, and strong partnerships, Rafay is positioned to lead in the AI infrastructure market.

Organizations must prioritize platforms that enable agility, compliance, and performance to remain competitive. Rafay offers a future-ready foundation for enterprise AI success.



## Contact


**Silicon Valley**
989 Commercial Street
Palo Alto, CA 94303

**Boston Metro**
95 Mount Royal Avenue
Marlborough, MA 01752

**David Butler**
david.butler@siliconangle.com
774-463-3400