Build and Run AI Factories with Rafay

Turn accelerated compute infrastructure into a secure, self-service AI platform.

Organizations are investing heavily in GPU infrastructure to power AI innovation. But without the right platform, these environments become fragmented, underutilized, and difficult for AI teams to access.

The Rafay Platform transforms raw GPU infrastructure into a secure, multi-tenant AI factory platform, enabling self-service consumption of compute while maintaining centralized governance and operational control.

The Rafay Platform for AI Factories

Rafay provides a platform layer for accelerated compute infrastructure, enabling organizations to build and run AI factories across data centers, cloud, and hybrid environments.

Self-Service AI Infrastructure: Instant access to GPU resources and AI environments without manual provisioning or ticket queues.

Accelerated Infrastructure Utilization: Secure multi-tenancy and shared infrastructure pools maximize GPU efficiency.

AI Platform Enablement: Deliver ready-to-use AI environments to accelerate AI development from experimentation to production.

Business Outcomes for AI Factories

Outcome	Impact
Accelerate AI Innovation	Reduce the time required to provision infrastructure so AI teams can experiment, train models, and deploy applications faster.
Maximize GPU Utilization	Turn underutilized GPU clusters into shared infrastructure pools that support multiple teams and workloads.
Scale AI Infrastructure Efficiently	Operate AI factories across cloud and data centers with centralized governance and automation.
Enable Secure Multi-Team AI Platforms	Support multiple teams and tenants while enforcing enterprise-grade governance and compliance.

Built for AI Infrastructure & GPU Cloud Providers

For NeoClouds & GPU Cloud Providers

Launch differentiated AI infrastructure services and monetize GPU investments faster.

With Rafay, NeoClouds can:

- Launch GPU infrastructure services in weeks
- Deliver GPU VMs, bare metal, and AI platforms as services
- Increase GPU utilization and infrastructure monetization
- Offer enterprise-ready multi-tenant AI environments

For Enterprises Building AI Factories

Transform internal GPU infrastructure into a self-service AI platform.

With Rafay, enterprises can:

- Provide self-service GPU environments for AI teams
- Standardize infrastructure across cloud and data centers
- Improve utilization of high-value GPU resources
- Enforce enterprise governance and security policies

Rafay: The Platform Behind AI Factories

Rafay transforms GPU clusters and accelerated infrastructure into operational AI factories, enabling organizations to scale AI innovation securely and efficiently.