

NVIDIA Blueprints-as-a-Service powered by Rafay

**Turn NVIDIA NIM Blueprints into differentiated, high-value services
with cloud-like agility and full control**

Deliver one-click, self-service deployment of NVIDIA AI Blueprints across on-premises or sovereign environments. Rafay's orchestration platform automates the end-to-end process of provisioning, configuring, and operationalizing NVIDIA Blueprints, enabling enterprises and service providers to deploy production-ready AI solutions in minutes instead of weeks.

From Bottleneck to Business Value

NVIDIA Blueprints provide modular, pre-integrated reference architectures that combine multiple NIM models, components, and workflows. Yet deploying these blueprints on-premises remains complex, requiring manual setup of resources, network policies, and access controls.

What users want is clear:

- A single-click, self-service experience for deploying NVIDIA AI Blueprints anywhere.
- Automated provisioning of all underlying NIM models, storage, and compute resources.
- Consistent governance, visibility, and compliance across environments.

Rafay's Blueprints-as-a-Service offering makes this possible. With Rafay, cloud providers and enterprises can transform NVIDIA's static blueprint definitions into fully operational, governed services. A single click deploys all required components : NIM models, pipelines, storage classes, and policies , orchestrated automatically with enterprise-grade compliance and lifecycle automation.

Designed For



Cloud & Service Providers: Offer NVIDIA AI Blueprints as self-service SKUs, enabling customers to deploy multi-model AI solutions instantly across GPU clouds.



Enterprises Accelerate AI adoption by simplifying the deployment of complex, multi-component blueprints within governed, on-premises environments.



Sovereign & Regional Clouds: Provide in-region, compliant, and fully automated NVIDIA AI Blueprint deployments that meet national data residency and security standards.

Key Capabilities

Capability	Description
One-Click Blueprint Deployment	Deploy NVIDIA Blueprints with a single action, including all associated NIM models, services, and dependencies.
Automated Infrastructure Orchestration	Rafay provisions and configures required GPU, storage, and network resources for each blueprint automatically.
NIM Model Integration	Supports individual NIM models as building blocks or as part of multi-model Blueprints.
Self-Service Portal	Expose blueprints as catalog items within Rafay's governed, multi-tenant interface.
Air-Gapped & On-Prem Ready	Deploy NVIDIA Blueprints securely within disconnected or sovereign environments.
Governance & Policy Controls	Enforce role-based access, audit logging, and compliance guardrails or all deployments.
Version & Lifecycle Management	Manage updates to NIM components, dependencies, and infrastructure without downtime.
Telemetry & Monitoring	Track resource usage, deployment health, and performance metrics in real time.

Business Outcomes

Outcome	Impact
Accelerated Time-to-Value	Reduce blueprint deployment time from weeks to minutes with a fully automated orchestration layer.
Operational Consistency	Deliver identical NVIDIA Blueprint deployments across cloud, on-prem, or sovereign environments with zero drift.
Enterprise & Sovereign Adoption	Enable enterprises and national clouds to deploy complex AI solutions compliantly, securely, and locally.