

Red Hat AI Factory with NVIDIA

Thank you for your interest
in exploring this content.

Carahsoft is the **Trusted Government IT Solutions Provider**® supporting a broad portfolio of industry-leading technologies through GSA, NASA SEWP V, ITES-SW2 and a wide range of other contract vehicles.

As the **Master Government Aggregator**®, Carahsoft connects government agencies, industry partners, and technology providers to deliver innovative, mission-focused solutions.

In partnership with Red Hat, we provide technology solutions that drive modernization, strengthen operations, and ensure compliance with evolving government standards.



To learn more about how Carahsoft can support your technology needs, please visit carahsoft.com



Explore More Resources:
carah.io/RedHatResources



Join Events & Webinars:
carah.io/RedHatEvents



Discover Technology Solutions:
carah.io/redhat



Learn About Procurement:
carah.io/RedHatContracts



Connect With Our Team:
RedHat@carahsoft.com
877-RHAT-GOV

Red Hat AI Factory with NVIDIA

AI quickstart examples

- Models-as-a-Service: Centrally host and serve approved models through API-based access.
- Enterprise RAG: Ground AI applications in proprietary enterprise data.
- Enterprise research: Use agentic workflows to search, synthesize, and distill internal content.
- Video search and summarization: Search, analyze, and summarize video content at scale.

Overview

As AI becomes instrumental to how organizations run and differentiate their business, they need greater control over applications, data, models, capacity planning, costs, and lifecycle management. Many teams are adopting AI through cloud services, disconnected tools, and fragmented infrastructure, making it harder to integrate AI into their own environments, apply proprietary data, manage usage, and govern AI consistently.

Red Hat® AI Factory with NVIDIA is a co-engineered, supported, enterprise-grade AI solution that brings together Red Hat's open hybrid cloud platform, AI engineering capabilities, and enterprise operations and NVIDIA's optimized AI software, models, microservices, networking, and accelerated infrastructure. It helps organizations run AI with greater control, consistency, privacy, and predictability across hybrid cloud environments.

Features and benefits

Table 1. Features and benefits of Red Hat AI Factory with NVIDIA

Feature	Benefit
Co-engineered and jointly supported enterprise AI solution	Combines Red Hat AI Enterprise and NVIDIA AI Enterprise in an integrated, enterprise AI stack with production support from both Red Hat and NVIDIA.
Unified platform for enterprise AI	Provides an integrated stack that remains open across models, tools, infrastructure, and environments, helping reduce fragmentation without forcing a closed appliance model.
Optimized high-performance inference	Helps platform teams meet model-serving service level objectives (SLOs) with optimized capabilities for throughput, latency, density, and use of NVIDIA accelerated infrastructure.
Scalable distributed serving	Supports large and mixed-model inference workloads with distributed serving for enterprise users, applications, and rack-scale infrastructure where supported.

Enterprise data as an intelligence asset	Helps teams apply proprietary data through retrieval-augmented generation (RAG), customization, evaluation, synthetic data generation, and AI pipelines to create business-specific intelligence.
Validated blueprints and AI quickstarts	Provides co-engineered prototype samples and tested starting points that show how to combine Red Hat and NVIDIA capabilities for common AI use cases.
Agentic AI operations	Helps teams identify frameworks and deploy, observe, control, and govern agents while connecting them to enterprise systems, tools, data sources.
Security and governance for autonomous workflows	Supports guardrails, observability, auditability, isolation, and least-privilege access patterns for AI workloads and agents that can act on users' behalf.
Predictable economics and infrastructure value	Supports planning with node-based pricing while helping organizations improve use, manage cost growth, and strengthen return on investment (ROI)/ total cost of ownership (TCO) arguments for AI infrastructure investments.

Technical specifications

Red Hat AI Factory with NVIDIA includes Red Hat AI Enterprise and NVIDIA AI Enterprise in a jointly supported offering for enterprise AI across hybrid cloud environments.

Key capabilities include:

- **Enterprise AI platform:** [Red Hat AI Enterprise](#) delivers a tested and supported full AI stack, including Red Hat OpenShift®, to support model development, customization, evaluation, deployment, inference, observability, governance, and lifecycle management. It also provides capabilities for AI gateways, synthetic data generation, AI pipelines, metrics, tracing, logging, and alerting.
- **Inference and serving:** Red Hat AI Inference, vLLM, llm-d, NVIDIA NIM microservices, NVIDIA Triton Inference Server, and NVIDIA TensorRT LLM support high-performance model serving and distributed inference where supported.
- **Models, agents, and customization:** This includes Red Hat validated and optimized models, NVIDIA optimized open models, NVIDIA NeMo, enterprise RAG, and AgentOps capabilities that help teams customize models, connect models to enterprise data and tools, and manage governed agent lifecycles.
- **NVIDIA infrastructure enablement:** NVIDIA CUDA-X libraries, NVIDIA GPU Operator, NVIDIA DOCA Operator, NVIDIA Network Operator, and NVIDIA NIM Operator help optimize and manage AI workloads on NVIDIA-accelerated computing and networking infrastructure.

Explore the foundation for enterprise AI

Discover how Red Hat AI Factory with NVIDIA helps organizations run AI with greater control, consistency, and performance across the hybrid cloud. Visit the [Red Hat AI Factory with NVIDIA](#) product page.

Start your AI Factory journey with trusted partners

Engage your preferred original equipment provider (OEM), solution provider, or distributor in the Red Hat and NVIDIA ecosystem to simplify procurement and map your unified AI platform strategy. Connect with a [certified partner](#) today.

Accelerate to production with Red Hat Services

Red Hat Consulting helps design initial AI architecture and onboard priority use cases, while Red Hat Training and Red Hat Technical Account Management help teams build GenAIOps skills and scale with confidence. Explore [Red Hat Services for AI Factory](#).



About Red Hat

Red Hat is the open hybrid cloud technology leader, delivering a trusted, consistent and comprehensive foundation for transformative IT innovation and AI applications. Its portfolio of cloud, developer, AI, Linux, automation and application platform technologies enables any application, anywhere—from the datacenter to the edge. As the world's leading provider of enterprise open source software solutions, Red Hat invests in open ecosystems and communities to solve tomorrow's IT challenges. Collaborating with partners and customers, Red Hat helps them build, connect, automate, secure, and manage their IT environments, supported by consulting services and [award-winning](#) training and certification offerings.

North America	888-REDHAT1	www.redhat.com	in @red-hat
Asia Pacific	+6564904200	apace@redhat.com	X @redhat
Latin America	+5443297300	info-latam@redhat.com	@redhat
Europe, Middle East, and Africa	+0080073342835	europe@redhat.com	@red_hat

Copyright © 2026 Red Hat. Red Hat, the Red Hat logo, Ansible, and OpenShift are trademarks or registered trademarks of Red Hat, LLC or its subsidiaries in the United States and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. The OPENSTACK logo and word mark are trademarks or registered trademarks of OpenInfra Foundation, used under license. All other trademarks are the property of their respective owners.