carahsoft

Red Hat

Accelerate the Path to Enterprise AI

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

Accelerate the Path to Enterprise AI

Red Hat AI Factory with NVIDIA

Solution Overview

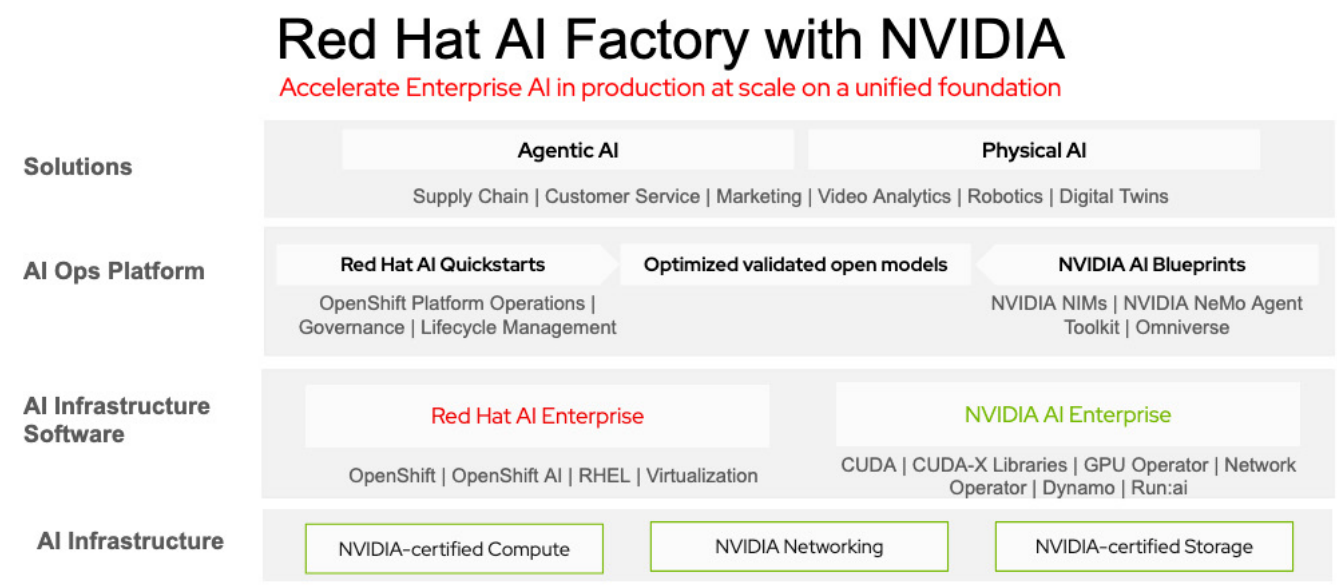
Red Hat® AI Factory with NVIDIA combines the integrated AI platform capabilities of Red Hat AI Enterprise with NVIDIA's accelerated computing, networking, and NVIDIA AI Enterprise software. Offered with simplified node-based pricing, this joint solution transforms the ad hoc creation, customization, and deployment of AI models into a repeatable, scalable, and safeguarded factory process across the hybrid cloud. Enterprises no longer need to assemble disconnected AI components piece-by-piece; instead, they gain a comprehensive, co-engineered foundation that bridges the gap between rapid infrastructure innovation and production stability.

By shifting from fragmented tools to a unified factory model, you can confidently accelerate your path from AI strategy to enterprise-wide production.

Business Value

While 95% of enterprises struggle to find measurable financial ROI from AI experiments, the AI Factory model transforms these isolated projects into a repeatable, scalable business capability. This joint platform delivers strategic value across three pillars:

- **Predictable Economics:** Stop paying a premium for cloud provider margins. Shift from an unpredictable pay-per-token model to a predictable cost structure, keeping efficiency gains directly on your own balance sheet.
- **Sustainable Autonomy & Privacy:** Maintain absolute control over your enterprise AI. Keep sensitive data, prompts, and usage strictly internal while ensuring a consistent, open platform that prevents vendor lock-in across hybrid environments.
- **Proven ROI:** Organizations are already realizing transformative, real-world impact across industries. For example, DenizBank achieved a 210% ROI over three years in financial services, healthcare providers have seen up to a 99.3% reduction in clinical report generation time, and Turkish Airlines captured $10 million in annual fuel savings.



Key benefits

- **Accelerated Time-to-Value:** Move from concept to production faster with streamlined workflows for agentic AI and model customization. Jumpstart development with instant access to validated, pre-configured models—including the indemnified IBM Granite family, NVIDIA Nemotron, and NVIDIA Cosmos—delivered as NVIDIA NIM microservices.
- **Optimized Operational Efficiency:** Maximize your infrastructure usage and bolster inference performance with a unified serving stack. The platform combines the Red Hat AI Inference Server (powered by vLLM) with NVIDIA NIM to deliver built-in observability and intelligent GPU orchestration, reducing your total cost of ownership (TCO).
- **Hardened Security & Mitigated Risk:** Deploy mission-critical AI workloads with confidence. The platform enforces strict workload isolation and a zero-trust architecture rooted in Red Hat's hardened OS features (SELinux, FIPS compliance) and NVIDIA DOCA microservices for real-time threat detection.
- **Hybrid Cloud Flexibility:** Deploy and scale your enterprise AI consistently across on-premises, edge, and public cloud environments while maintaining full architectural control. This unified platform eliminates fragmented silos, allowing you to bring AI directly to where your data securely lives. As a result, your governance policies, security standards, and operational workflows remain uniform whether you are running workloads on bare metal, virtual machines, or in the public cloud.

Common Use Cases

Model-as-a-Service	Centralize hosting for multiple teams and applications with robust orchestration, multi-tenancy, and API management capabilities.
Enterprise RAG	Ground AI in your proprietary data securely using NVIDIA NeMo Retriever while relying on Red Hat for full orchestration and governance control.
Enterprise Research & Agentic AI	Connect AI agents directly to enterprise data to distill complex materials with improved semantic accuracy, efficiency, and precision.
Video Search & Summarization	Deploy NVIDIA VLM microservices to search, summarize, and analyze video content at massive scale.

Call to action

Start your AI Factory journey today. Engage your preferred OEM, Value-Added Reseller (VAR), or Distributor to design, deliver, and scale your deployment. Explore these resources to get started:

- PR: Red Hat AI Factory with NVIDIA Accelerates the Path to Scalable Production AI
- Webpage: Red Hat AI Factory with NVIDIA
- Red Hat AI quickstarts

About NVIDIA

NVIDIA's architectural breakthroughs have made AI an imperative, proving that the accelerated computing stack will define the industry's future. NVIDIA pioneers the computing infrastructure, networking, and software that turns data into intelligence at scale, powering the next generation of agentic and physical AI applications.

About Red Hat

Red Hat is an open hybrid cloud technology leader, delivering a consistent, comprehensive foundation for transformative IT and artificial intelligence (AI) applications in the enterprise. As a trusted adviser to the Fortune 500, Red Hat offers cloud, developer, Linux, automation, and application platform technologies, as well as award-winning services.

North America
1–888–REDHAT1
www.redhat.com

Europe, Middle East, and Africa 00800 7334 2835
europe@redhat.com

Asia Pacific +65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com