



Red Hat



NVIDIA

6 reasons the Global Public Sector chooses Red Hat AI Factory with NVIDIA

Scaling AI in global public sector and government organizations has a different bar to clear than other industries do: classified data, air-gapped networks, and outputs that may need to be defended under oversight scrutiny. Here's how Red Hat AI Factory with NVIDIA clears the bar to take AI from pilot to production in a repeatable, scalable, and safeguarded factory process.

1 Keep sensitive data under sovereign control

Public sector and government entities can't let classified data near a public cloud. Red Hat AI Factory with NVIDIA keeps sensitive data, models, and infrastructure inside the organization's own environment, including air-gapped and fully disconnected deployments.

Red Hat's hardened operating system features, like SELinux and comprehensive security certifications, combined with NVIDIA's highly secured container environments, protect your digital sovereignty and support compliance with local security standards.

2 Run AI workloads wherever the mission demands

You need the flexibility to deploy AI where and when your mission it demands it. A workload that runs in a data center today might need to run at the disconnected mission edge tomorrow, at a different classification level and on a lot less bandwidth. Rebuilding infrastructure and retraining teams as needs change isn't sustainable. Or secure.

With Red Hat AI Factory with NVIDIA, you won't need to rebuild or retrain when workloads need to move across environments. The platform delivers a consistent architecture across data centers, hybrid clouds, and disconnected tactical edge environments, regardless of classification level. Red Hat OpenShift AI handles orchestration and access control, and NVIDIA accelerated computing runs consistently across every architecture from the data center to the edge.

3 Get AI output that's traceable and defensible

Teams in public sector organizations can't rely on plausible answers from LLMs connected to the open internet. They need output that is verifiable and defensible. With Red Hat AI Factory with NVIDIA, teams can use Retrieval-Augmented Generation (RAG) to connect AI models to their own approved data, like national archives, policy documents, regional regulatory filings, and secured internal databases.

Every AI output is traceable back to its source, whether that's a specific document or a timestamp in video monitoring footage, to reduce hallucination risk. Get the confidence to make AI-assisted decisions that hold up under oversight scrutiny.

4 Scale access without shadow AI

When different agencies, departments, and teams are eager to use AI and adoption expands, unsanctioned tools and models can spread with little visibility into who's using them and how. This proliferation of unauthorized AI puts your organizations at risk for data leakage, regional regulatory compliance failures such as sovereign data privacy acts, and security vulnerabilities.

Red Hat AI Factory with NVIDIA heads off shadow AI with a Models-as-a-Service (MaaS) architecture that centrally hosts and manages models on infrastructure under your organization's control. Teams get access to governed models through standardized APIs instead of each one standing up its own AI infrastructure from scratch. Role-based access controls, secure multitenancy, and usage tracking are built in. Your organization gets reusable AI capabilities across teams with the benefits of centralized visibility and control.

5 Do more with constrained compute at the edge

AI in remote enterprise and public sector environments often has to run in challenging places where computing power and electricity are strictly limited. This includes remote edge deployments, disconnected facilities, and older hardware that cannot easily be upgraded. In these environments, getting the absolute most out of your existing hardware isn't just a nice-to-have—it is the difference between running a critical AI application or having it fail entirely.

With Red Hat AI Factory with NVIDIA, global public sector agencies can deploy high-performing AI directly to the field with confidence. Red Hat and NVIDIA make it possible to run smart, low-latency AI models on smaller, edge-constrained hardware through three key features:

- ▶ **Intelligent Model Compression:** Utilizing built-in model compression tools (powered by Neural Magic technology), Red Hat drastically reduces the size and memory footprint of large AI models. This allows them to run at high speeds on smaller, lightweight hardware without losing their accuracy.
- ▶ **Smart Distributed Serving with `llm-d`:** For larger, more complex AI models that require multiple small devices to work together, `llm-d` acts as an intelligent coordinator. It routes requests to the devices that can process them fastest and avoids wasting energy or computing power.
- ▶ **Efficient Resource Allocation:** Red Hat ensures that precious GPU and hardware resources are divided dynamically across your different tasks, preventing one heavy workload from blocking other critical field operations.

6 Move from pilot to mission-ready production

Building AI infrastructure from scratch for every mission or classification level doesn't scale. You need a repeatable path from prototype to production.

Red Hat AI Factory with NVIDIA provides validated blueprints and co-engineered AgentOps capabilities so you so you're not rebuilding architecture with each deployment. NVIDIA NIM handles optimized inference, and Red Hat OpenShift AI enables unified prototyping, training, and deployment in a single, secured environment. Your organization moves from ad hoc AI experiments to a scalable, safeguarded process that clears the bar for your sovereignty and governance requirements.

Discover more about how to scale AI

Learn about the capabilities of Red Hat AI Factory with NVIDIA [here](#).



About Red Hat

Red Hat helps customers standardize across environments, develop cloud-native applications, and integrate, automate, secure, and manage complex environments with [award-winning](#) support, training, and consulting services.

f facebook.com/redhat
X x.com/RedHat
in linkedin.com/company/red-hat

redhat.com
#3662723_0326

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