SOLUTION OVERVIEW

AI-READY ENTERPRISE PLATFORM

Now with Red Hat OpenShift

Unleashing AI for Every Enterprise

Artificial intelligence (AI) is transforming every industry, whether it's by improving customer relationships in financial services, streamlining manufacturer supply chains, or helping doctors deliver better outcomes for patients. IDC projects that by 2024, 60 percent of global 2000 companies will expand the use of AI and machine learning (ML) across all business-critical horizontal functions, including marketing, legal, HR, procurement, and supply chain logistics.

Yet, while organizations know they need to invest in AI to secure their future, many struggle with finding the strategy and platform to enable success. Unlike traditional enterprise applications, AI apps are relatively new for enterprise IT departments. They're anchored in rapidly-evolving, open-source, bleeding-edge code and lack proven approaches that meet the rigors of scaled enterprise production settings.

AI-Ready Platform from NVIDIA and Red Hat

NVIDIA and Red Hat are working together to enable customers to run their businesses using familiar infrastructure on many footprints—from virtualized and private cloud deployments in corporate data centers, to massive-scale services deployed on public clouds, to the edge.

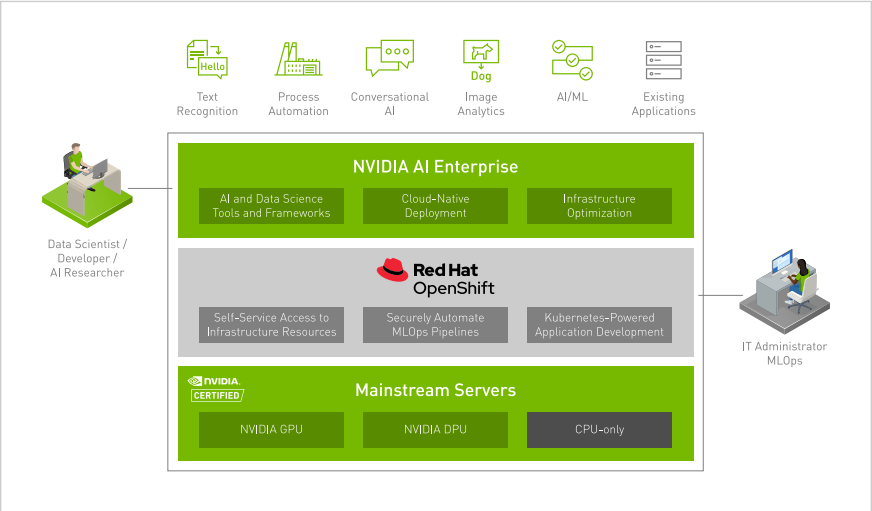
Combined, the NVIDIA AI Enterprise software suite, running on NVIDIA-Certified Systems™, and Red Hat® OpenShift®, the industry leading Kubernetes-powered hybrid cloud solution, offer a scalable platform that helps accelerate a diverse range of AI use cases. Enterprises can expedite deep learning (DL) and ML for every industry—from healthcare to financial services to manufacturing. NVIDIA AI Enterprise with Red Hat OpenShift also provides unparalleled flexibility, offering organizations the option to deploy in both bare-metal and virtualized environments.

KEY CHALLENGES

- **Complexity with deployment and scaling** Procuring and operationalizing an AI infrastructure takes months because of complexities with design, deployment, and scaling the end to end solution anywhere.
- **Lack of self-service access to AI tools and infrastructure** Data scientists, ML engineers, and developers have a constant dependency on IT for AI development tools, and compute resources to get their work done.
- **Siloed Operations** Manual, siloed operations across teams (data scientists, developers, etc.) results in slow ML modeling and intelligent applications development

BENEFITS

- **Ease of deployment and scaling** Leverage the jointly certified NVIDIA and Red Hat solution to deploy and manage AI workloads in containers or VMs with optimized software.
- **Accelerate Compute-Intensive ML** Modeling and Inferencing Jobs Maximize performance for ML modeling and inference
- **Streamline Delivery of Intelligent** Applications with Integrated MLOps Enable collaboration across teams so ML models can be quickly integrated into the development of intelligent applications and deployed into production.



NVIDIA and Red Hat AI-Ready Platform

Key Enabling Technologies

The NVIDIA and Red Hat solution is comprised of the following building blocks:

NVIDIA AI Enterprise

NVIDIA AI Enterprise, certified, licensed, and supported by NVIDIA, is an end-to-end, cloud-native suite of AI software, certified to run on Red Hat Openshift with NVIDIA-Certified Systems. It includes key enabling technologies and software from NVIDIA for the rapid deployment, management, and scaling of AI workloads in the bare metal and virtualized data center.

Red Hat OpenShift

Red Hat OpenShift is the industry leading Kubernetes-powered hybrid cloud solution that enables a cloud-like experience wherever it's deployed, and helps accelerate secure applications development and delivery. Whether it's in the public cloud, on-premises, or at the edge, Red Hat OpenShift gives enterprises the ability to choose where to build, deploy, and run applications through a consistent experience. Red Hat OpenShift's full-stack automated operations and self-service provisioning allows teams to work together more efficiently, and quickly and securely move ideas from development to production. NVIDIA GPU integration via the GPU Operator also helps accelerate the modeling and inferencing of tasks.

Deploying AI workloads in production is an iterative process that extends beyond simply creating AI/ML models. Red Hat OpenShift's built-in DevOps and GitOps capabilities enable MLOps, and streamline the development of intelligent, AI/ML-based applications.

NVIDIA-Certified Systems

The NVIDIA AI Enterprise Suite is certified to run on NVIDIA-Certified Systems, which include the following:

- NVIDIA Ampere architecture-based GPUs, the tensor core technology included in the Ampere architecture, deliver dramatic speedups to AI operations, reduce training times from weeks to hours and provide massive inference acceleration.
- NVIDIA® ConnectX® SmartNICs and the NVIDIA BlueField® data processing unit (DPU) provide a host of software-defined hardware engines for accelerating networking and security. These enable the best of both worlds: best-in-class AI training and inference performance with all the necessary levels of enterprise data privacy, integrity, and reliability.

NVIDIA AI Enterprise Trial Programs

NVIDIA offers the following trial programs that enable enterprise customers to evaluate products for free, based on their existing infrastructure:

- **NVIDIA LaunchPad**
  The NVIDIA LaunchPad program provides worldwide enterprises and organizations with immediate, short-term access to the NVIDIA AI Enterprise software suite running on private accelerated computing infrastructure. It also includes a set of curated, hands-on labs for AI practitioners and IT staff on how to orchestrate Kubernetes with Red Hat OpenShift.
- **Evaluation Software**
  NVIDIA AI Enterprise evaluation software is available for customers with existing NVIDIA-Certified Systems who are ready to start a proof-of-concept (POC) project for deployment at scale.