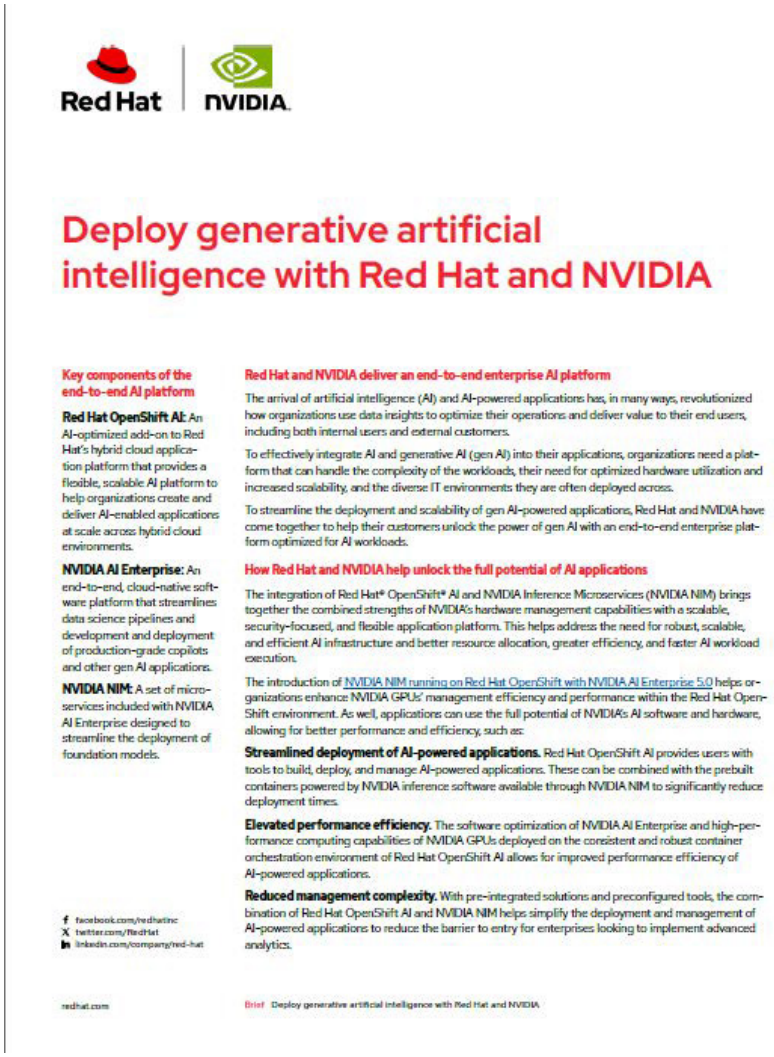
Red Hat

Deploy generative artificial intelligence with Red Hat and NVIDIA



Red Hat | NVIDIA

Deploy generative artificial intelligence with Red Hat and NVIDIA

Key components of the end-to-end AI platform

Red Hat OpenShift AI: An AI-optimized add-on to Red Hat's hybrid cloud application platform that provides a flexible, scalable AI platform to help organizations create and deliver AI-enabled applications at scale across hybrid cloud environments.

NVIDIA AI Enterprise: An end-to-end, cloud-native software platform that streamlines data science pipelines and development and deployment of production-grade copilots and other gen AI applications.

NVIDIA NIM: A set of microservices included with NVIDIA AI Enterprise designed to streamline the deployment of foundation models.

facebook.com/redhatinc
twitter.com/RedHat
linkedin.com/company/red-hat

redhat.com

Red Hat and NVIDIA deliver an end-to-end enterprise AI platform

The arrival of artificial intelligence (AI) and AI-powered applications has, in many ways, revolutionized how organizations use data insights to optimize their operations and deliver value to their end users, including both internal users and external customers.

To effectively integrate AI and generative AI (gen AI) into their applications, organizations need a platform that can handle the complexity of the workloads, their need for optimized hardware utilization and increased scalability, and the diverse IT environments they are often deployed across.

To streamline the deployment and scalability of gen AI-powered applications, Red Hat and NVIDIA have come together to help their customers unlock the power of gen AI with an end-to-end enterprise platform optimized for AI workloads.

How Red Hat and NVIDIA help unlock the full potential of AI applications

The integration of Red Hat® OpenShift® AI and NVIDIA Inference Microservices (NVIDIA NIM) brings together the combined strengths of NVIDIA's hardware management capabilities with a scalable, security-focused, and flexible application platform. This helps address the need for robust, scalable, and efficient AI infrastructure and better resource allocation, greater efficiency, and faster AI workload execution.

The introduction of NVIDIA NIM running on Red Hat OpenShift with NVIDIA AI Enterprise 5.0 helps organizations enhance NVIDIA GPUs' management efficiency and performance within the Red Hat OpenShift environment. As well, applications can use the full potential of NVIDIA's AI software and hardware, allowing for better performance and efficiency, such as:

Streamlined deployment of AI-powered applications. Red Hat OpenShift AI provides users with tools to build, deploy, and manage AI-powered applications. These can be combined with the prebuilt containers powered by NVIDIA inference software available through NVIDIA NIM to significantly reduce deployment times.

Elevated performance efficiency. The software optimization of NVIDIA AI Enterprise and high-performance computing capabilities of NVIDIA GPUs deployed on the consistent and robust container orchestration environment of Red Hat OpenShift AI allows for improved performance efficiency of AI-powered applications.

Reduced management complexity. With pre-integrated solutions and preconfigured tools, the combination of Red Hat OpenShift AI and NVIDIA NIM helps simplify the deployment and management of AI-powered applications to reduce the barrier to entry for enterprises looking to implement advanced analytics.

Brief Deploy generative artificial intelligence with Red Hat and NVIDIA

Deploy generative artificial intelligence with Red Hat and NVIDIA

Key components of the end-to-end AI platform

Red Hat OpenShift AI: An AI-optimized add-on to Red Hat's hybrid cloud application platform that provides a flexible, scalable AI platform to help organizations create and deliver AI-enabled applications at scale across hybrid cloud environments.

NVIDIA AI Enterprise: An end-to-end, cloud-native software platform that streamlines data science pipelines and development and deployment of production-grade copilots and other gen AI applications.

NVIDIA NIM: A set of micro-services included with NVIDIA AI Enterprise designed to streamline the deployment of foundation models.

Red Hat and NVIDIA deliver an end-to-end enterprise AI platform

The arrival of artificial intelligence (AI) and AI-powered applications has, in many ways, revolutionized how organizations use data insights to optimize their operations and deliver value to their end users, including both internal users and external customers.

To effectively integrate AI and generative AI (gen AI) into their applications, organizations need a platform that can handle the complexity of the workloads, their need for optimized hardware utilization and increased scalability, and the diverse IT environments they are often deployed across.

To streamline the deployment and scalability of gen AI-powered applications, Red Hat and NVIDIA have come together to help their customers unlock the power of gen AI with an end-to-end enterprise platform optimized for AI workloads.

How Red Hat and NVIDIA help unlock the full potential of AI applications

The integration of Red Hat® OpenShift® AI and NVIDIA Inference Microservices (NVIDIA NIM) brings together the combined strengths of NVIDIA's hardware management capabilities with a scalable, security-focused, and flexible application platform. This helps address the need for robust, scalable, and efficient AI infrastructure and better resource allocation, greater efficiency, and faster AI workload execution.

The introduction of NVIDIA NIM running on Red Hat OpenShift with NVIDIA AI Enterprise 5.0 helps organizations enhance NVIDIA GPUs' management efficiency and performance within the Red Hat Open-Shift environment. As well, applications can use the full potential of NVIDIA's AI software and hardware, allowing for better performance and efficiency, such as:

Streamlined deployment of AI-powered applications. Red Hat OpenShift AI provides users with tools to build, deploy, and manage AI-powered applications. These can be combined with the prebuilt containers powered by NVIDIA inference software available through NVIDIA NIM to significantly reduce deployment times.

Elevated performance efficiency. The software optimization of NVIDIA AI Enterprise and high-performance computing capabilities of NVIDIA GPUs deployed on the consistent and robust container orchestration environment of Red Hat OpenShift AI allows for improved performance efficiency of AI-powered applications.

Reduced management complexity. With pre-integrated solutions and preconfigured tools, the combination of Red Hat OpenShift AI and NVIDIA NIM helps simplify the deployment and management of AI-powered applications to reduce the barrier to entry for enterprises looking to implement advanced analytics.

facebook.com/redhatinc
twitter.com/RedHat
linkedin.com/company/red-hat

Improved scalability and flexibility. The combined capabilities of NVIDIA's full-stack solution and Red Hat's platform helps support scalable and flexible deployment options that can adapt to any organization's changing business needs and data volume.

Enhanced focus on security and compliance. The security-focused computing environments and security features provided by this joint solution helps improve the focus on compliance and security of data and AI models in all IT environments.

Streamline AI innovation with a consistent and flexible end-to-end platform

Whether your organization is trying to integrate AI into its operations to enhance data analytics, streamline business operations, or find ways to innovate at the edge of your network, Red Hat and NVIDIA can help you turn data into actionable insights and new business opportunities.

Unlock advanced inferencing for AI models of all types

With NVIDIA NIM running on Red Hat OpenShift, organizations can integrate advanced AI inference into their applications, benefit from real-time decision-making, and improve workload efficiency. This allows organizations to benefit from flexible inference runtime and support for a variety of AI models, including open source community models, NVIDIA AI foundation models, or their own bespoke customized models.

Automate GPU provisioning with NVIDIA software

Deploying NVIDIA AI Enterprise (and subsequently NVIDIA NIM) on Red Hat OpenShift AI allows NVIDIA GPU Operator to use the operator framework within Kubernetes to automate the management of all NVIDIA software components needed to provision GPUs, including NVIDIA drivers, Kubernetes device plug-ins for GPUs, the NVIDIA Container Runtime, automatic node labeling, DCGM-based monitoring, and others.

Develop, deploy, and scale AI applications with flexibility and consistency

Red Hat OpenShift AI provides a complete developer toolset and application portability, that when combined with NVIDIA NIM, allows developers to streamline how they build gen AI-powered applications. This allows for a consistent developer experience and management of NVIDIA NIM across all IT environments, as well as a consistent application programming interface (API) based on OpenAI standards for generative models.

Start optimizing gen AI development with Red Hat and NVIDIA

Contact a Red Hatter to learn more about how the combination of Red Hat's platform with NVIDIA's AI architecture and microservices can help you streamline the development and deployment of gen AI-powered applications, with consistency, flexibility, scalability, and an improved focus on security and compliance.

North America	Europe, Middle East, and Africa	Asia Pacific	Latin America
1 888 REDHAT1 www.redhat.com	00800 7334 2835 europe@redhat.com	+65 6490 4200 apac@redhat.com	+54 11 4329 7300 info-latam@redhat.com

facebook.com/redhatinc
twitter.com/RedHat
linkedin.com/company/red-hat

redhat.com

Thank you for downloading this Red Hat resource! Carahsoft is the Master GSA and SLSA Dealer and Distributor for Red Hat Enterprise Open Source solutions available via GSA, SLSA, ITES-SW2, The Quilt and other contract vehicles.

To learn how to take the next step toward acquiring Red Hat's solutions, please check out the following resources and information:

For additional resources:
carah.io/RedHatResources

For upcoming events:
carah.io/RedHatEvents

For additional Red Hat solutions:
carah.io/RedHatSolutions

For additional Open Source solutions:
carah.io/OpenSourceSolutions

To set up a meeting:
redhat@carahsoft.com
877-RHAT-GOV

To purchase, check out the contract vehicles available for procurement:
carah.io/RedHatContracts