HPE Machine Learning Inference Software

What's new

- Create a simplified path to scalable production model deployments for MLOps or ITOps, using an intuitive graphical interface removing the need for extensive Kubernetes experience.
- Streamlined integration with Hugging Face and NVIDIA Foundation Models offers a zero-coding deployment experience for large language models (LLMs) directly from Hugging Face and NVIDIA NGC.
- Seamless integration with NVIDIA AI Enterprise® includes NIM® microservices for enhanced inference on more than two dozen popular AI models from NVIDIA and partners.
- Facilitate support for pre-trainied and

Overview

Do you need to streamline the AI/ML deployment process? Do you need to support a diverse AI frameworks and scalable infrastructure in a cloud/hybrid environment that often requires customized data protection? The HPE Machine Learning Inference Software features user-friendly tools to update, monitor, and deploy models that will help you get value from AI/ML initiatives faster. Role-Based Access Controls (RBAC) and endpoint security provide additional protection for ML resources. Dramatically improve team efficiency by using consistent tooling and pre-trained models to focus more on model development and less on the complexities of getting models into production. By offering a product that handles the intricacies of deployment, routing, and real-time monitoring, HPE Machine Learning Inference Software provides the agility needed to ship ML models quickly, iterate on them based on feedback from the real-world, and maintain high-performance standards.

bespoke models built on popular frameworks such as TensorFlow, PyTorch, scikit-learn, and XGBoost.

- Benefit from integrated monitoring and logging for tracking model performance, usage metrics, and system health, facilitating proactive optimization.
- Offer adaptable deployment across varied infrastructures with compatibility for many Kubernetes environments, including HPE Ezmeral, HPE GreenLake, AWS, Azure, Google Cloud, and on-premise setups.

Features

Predictable, Dependable, Protected, and Monitored Deployment for Diverse Environments

HPE Machine Learning Inference Software can deploy models using an intuitive graphical interface and scale deployments based on load.

Customize performance with real-time monitoring of models and track predictions and statistics around deployment.

Whether in an existing Kubernetes cluster, a private cloud, or even a hybrid cloud, HPE Machine Learning Inference Software provides consistent tooling across continually modernizing systems to meet your needs.

Industry-standard Helm charts are used to deploy into any Kubernetes-compatible platform, e.g., OpenShift, Rancher, EKS, AKS, or GKS—any cloud can be leveraged consistently.

Out-of-box Support for NVIDIA Models and Tools

HPE Machine Learning Inference Software offers flexible, first-class support for Nvidia GPUs with architecture to easily add support for continually-modernizing systems.

Integration with NVIDIAs' AI Enterprise (NVAIE) software suite, NVIDIA Inference Microservice (NIM) (utilizing Triton, TensorRT-LLM) and other AI inferencing techniques offer enhanced performance.

Built-In Enterprise-Class Security

HPE Machine Learning Inference Software features execute workloads in your preferred environment, including cloud, hybrid, on-premise, or even air gaped—thus enabling models, code, and data to remain protected.

Use Role-Based Access Controls (RBAC) to authorize development and MLOps teams to collaborate and share ML resources and artifacts securely.

Protect deployment endpoints with enterprise-class security features that require advanced authentication, including OIDC and OAuth 2.0, to interact with models.

Broad Model Compatibility

HPE Machine Learning Inference Software offers streamlined integration for specific large language models (LLMs) directly from Hugging Face and NVIDIA Inference Server (NIM) while enabling development of models from most frameworks.

Achieve increased flexibility using models from diverse frameworks such as TensorFlow, PyTorch, Scikit-Learn, and XGBoost to accommodate a broad range of pre-trained and customer models.

Technical specifications

	HPE Machine Learning Inference Software
Supported hardware environment	We support NVIDIA GPUs from the Ampere generation and newer. Please check the HPE Machine Learning Inference Software QuickSpecs for supported GPUs.
Version	V1.0

For additional technical information, available models and options, please reference the QuickSpecs

HPE Services

No matter where you are in your transformation journey, you can count on HPE Services to deliver the expertise you need when, where and how you need it. From strategy and planning to deployment, ongoing operations and beyond, our experts can help you realize your digital ambitions.

Consulting services

Experts can help you map out your path to hybrid cloud and optimize your operations.

Managed services

HPE runs your IT operations, giving you unified control, so can focus on innovation.

Operational services

Optimize your entire IT environment and drive innovation. Manage day-to-day IT operational tasks while freeing up valuable time and resources.

- HPE Complete Care Service: a modular service designed to help optimize your entire IT environment and achieve agreed upon IT outcomes and business goals. All delivered by an assigned team of HPE experts.
- HPE Tech Care Service: the operational service experience for HPE products. The service provides access to product specific experts, an AI driven digital experience, and general technical guidance to help reduce risk and search for ways to do things better.

Lifecycle Services

Address your specific IT deployment project needs with tailored project management and deployment services.

HPE Education Services

Training and certification designed for IT and business professionals across all industries. Create learning paths to expand proficiency in a specific subject. Schedule training in a way that works best for your business with flexible continuous learning options.

The Defective Media Retention (DMR) service feature option applies only to Disk or eligible SSD/Flash Drives replaced by Hewlett Packard Enterprise due to malfunction. Comprehensive Defective Material Retention (CDMR) allows you to keep all data retentive components.

HPE GreenLake

HPE GreenLake edge-to-cloud platform is HPE's market-leading as-a-Service offering that brings the cloud experience to apps and data everywhere – data centers, multi-clouds, and edges – with one unified operating model, on premises, fully managed in a pay per use model.

If you are looking for more services, like **IT financing solutions**, please explore them here.

Explore **HPE GreenLake**

Make the right purchase decision. Contact our presales specialists.

Chat now (sales)

Share now

Get updates

PSN1014828256WWEN, October, 2024.

Hewlett Packard Enterprise