It Takes a Village: 100+ NVIDIA MLOps and AI Platform Partners Help Enterprises Move AI Into Production

Software providers team with NVIDIA to help customers build more effective AI pipelines.

Author: Scott McClellan

Building AI applications is hard. Putting them to use across a business can be even harder.

Less than one-third of enterprises that have begun adopting AI actually have it in production, according to a recent IDC survey .

Businesses often realize the full complexity of operationalizing AI just prior to launching an application. Problems discovered so late can seem insurmountable, so the deployment effort is often stalled and forgotten.

To help enterprises get AI deployments across the finish line, more than 100 machine learning operations (MLOps) software providers are working with NVIDIA. These MLOps pioneers provide a broad array of solutions to support businesses in optimizing their AI workflows for both existing operational pipelines and ones built from scratch.

Many NVIDIA MLOps and AI platform ecosystem partners as well as DGX-Ready Software partners, including Canonical, ClearML, Dataiku, Domino Data Lab, Run:ai and Weights & Biases, are building solutions that integrate with NVIDIA-accelerated infrastructure and software to meet the needs of enterprises operationalizing AI.

NVIDIA cloud service provider partners Amazon Web Services, Google Cloud, Azure, Oracle Cloud as well as other partners around the globe, such as Alibaba Cloud, also provide MLOps solutions to streamline AI deployments.

NVIDIA's leading MLOps software partners are verified and certified for use with the NVIDIA AI Enterprise software suite , which provides an end-to-end platform for creating and accelerating production AI. Paired with NVIDIA AI Enterprise, the tools from NVIDIA's MLOps partners help businesses develop and deploy AI successfully.

Enterprises can get AI up and running with help from these and other NVIDIA MLOps and AI platform partners:

Canonical: Aims to accelerate at-scale AI deployments while making open source accessible for AI development. Canonical announced that Charmed Kubeflow is now certified as part of the DGX-Ready Software program , both on single-node and multi-node deployments of NVIDIA DGX systems . Designed to automate machine learning workflows, Charmed Kubeflow creates a reliable application layer where models can be moved to production.

ClearML : Delivers a unified, open-source platform for continuous machine learning — from experiment management and orchestration to increased performance and ML production — trusted by teams at 1,300 enterprises worldwide. With ClearML, enterprises can orchestrate and schedule jobs on personalized compute fabric. Whether on premises or in the cloud, businesses can enjoy enhanced visibility over infrastructure usage while reducing compute, hardware and resource spend to optimize cost and performance. Now certified to run NVIDIA AI Enterprise , ClearML's MLOps platform is more efficient across workflows, enabling greater optimization for GPU power.

Dataiku : As the platform for Everyday AI, Dataiku enables data and domain experts to work together to build AI into their daily operations. Dataiku is now certified as part of the NVIDIA DGX-Ready Software

program, which allows enterprises to confidently use Dataiku's MLOps capabilities along with NVIDIA DGX AI supercomputers.

Domino Data Lab : Offers a single pane of glass that enables the world's most sophisticated companies to run data science and machine learning workloads in any compute cluster — in any cloud or on premises in all regions. Domino Cloud , a new fully managed MLOps platform-as-a-service, is now available for fast and easy data science at scale. Certified to run on NVIDIA AI Enterprise last year , Domino Data Lab's platform mitigates deployment risks and ensures reliable, high-performance integration with NVIDIA AI.

Iguazio : Provides a platform that automates, accelerates and scales MLOps, transforming AI projects into real-world business outcomes. Iguazio's global clients use its "ML factory," which allows continuous deployment of new AI services in a repeatable, scalable and reproducible way while automating development, deployment and management of AI applications in live business conditions. The platform works across use cases and in on-premises, multi-cloud and hybrid cloud environments. Iguazio is a DGX-Ready certified partner and has recently been validated as an NVIDIA AI Accelerated partner .

Run:ai : Functions as a foundational layer within enterprises' MLOps and AI Infrastructure stacks through its AI computing platform, Atlas. The platform's automated resource management capabilities allow organizations to properly align resources across different MLOps platforms and tools running on top of Run:ai Atlas. Certified to offer NVIDIA AI Enterprise , Run:ai is also fully integrating NVIDIA Triton Inference Server , maximizing the utilization and value of GPUs in AI-powered environments.

Weights & Biases (W&B;) : Helps machine learning teams build better models, faster. With just a few lines of code, practitioners can instantly debug, compare and reproduce their models — all while collaborating with their teammates. W&B; is trusted by more than 500,000 machine learning practitioners from leading companies and research organizations around the world. Now validated to offer NVIDIA AI Enterprise , W&B; looks to accelerate deep learning workloads across computer vision, natural language processing and generative AI .

NVIDIA cloud service provider partners have integrated MLOps into their platforms that provide NVIDIA accelerated computing and software for data processing, wrangling, training and inference:

Amazon Web Services : Amazon SageMaker for MLOps helps developers automate and standardize processes throughout the machine learning lifecycle, using NVIDIA accelerated computing. This increases productivity by training, testing, troubleshooting, deploying and governing ML models.

Google Cloud : Vertex AI is a fully managed ML platform that helps fast-track ML deployments by bringing together a broad set of purpose-built capabilities. Vertex AI's end-to-end MLOps capabilities make it easier to train, orchestrate, deploy and manage ML at scale, using NVIDIA GPUs optimized for a wide variety of AI workloads. Vertex AI also supports leading-edge solutions such as the NVIDIA Merlin framework, which maximizes performance and simplifies model deployment at scale. Google Cloud and NVIDIA collaborated to add Triton Inference Server as a backend on Vertex AI Prediction, Google Cloud's fully managed model-serving platform.

Azure : The Azure Machine Learning cloud platform is accelerated by NVIDIA and unifies ML model development and operations (DevOps). It applies DevOps principles and practices — like continuous integration, delivery and deployment — to the machine learning process, with the goal of speeding experimentation, development and deployment of Azure machine learning models into production. It provides quality assurance through built-in responsible AI tools to help ML professionals develop fair, explainable and responsible models.

Oracle Cloud: Oracle Cloud Infrastructure (OCI) AI Services is a collection of services with prebuilt machine learning models that make it easier for developers to apply NVIDIA-accelerated AI to applications and business operations. Teams within an organization can reuse the models, datasets and data labels across services. OCI AI Services makes it possible for developers to easily add machine learning to apps without slowing down application development.

Alibaba Cloud: Alibaba Cloud Machine Learning Platform for AI provides an all-in-one machine learning service featuring low user technical skills requirements, but with high performance results. Accelerated by NVIDIA, the Alibaba Cloud platform enables enterprises to quickly establish and deploy machine learning experiments to achieve business objectives.

Learn more about NVIDIA MLOps partners and their work at NVIDIA GTC , a global conference for the era of AI and the metaverse, running online through Thursday, March 23.

Watch NVIDIA founder and CEO Jensen Huang's GTC keynote in replay:

Original URL:
https://blogs.nvidia.com/blog/2023/03/22/mlops-ai-platform-partners-move-ai-into-production/