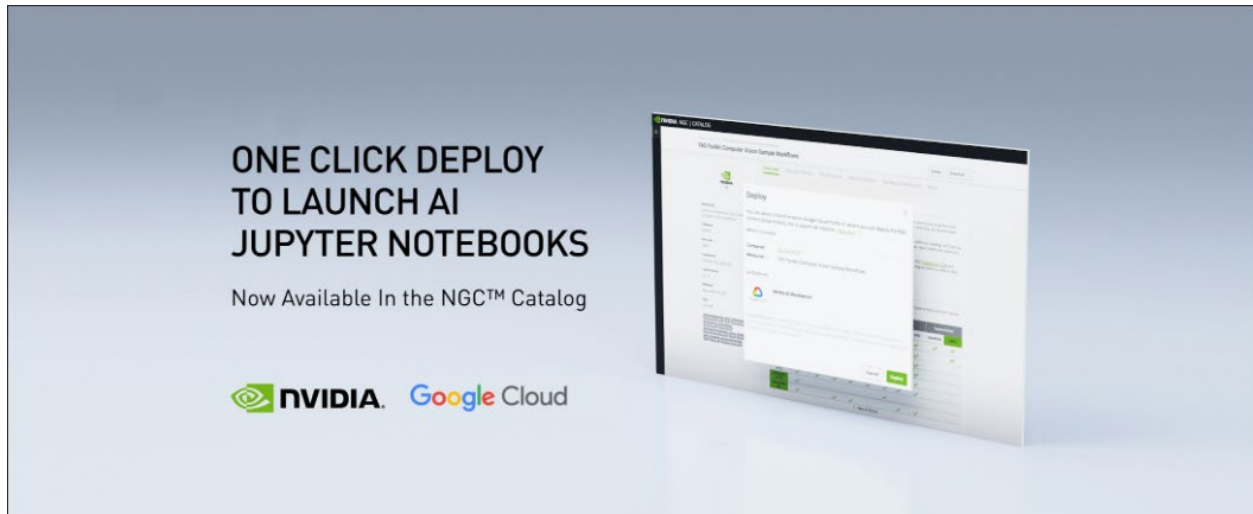AI & MACHINE LEARNING

Building AI in the cloud: An easier way with Google Cloud and NVIDIA

Chintan Patel
Sr. Manager, Product Marketing, NVIDIA

NVIDIA GPU-powered instances on Google Cloud provide an optimal platform for organizations to develop their AI applications on the latest hardware and software stack, then seamlessly deploy those applications at scale in production.
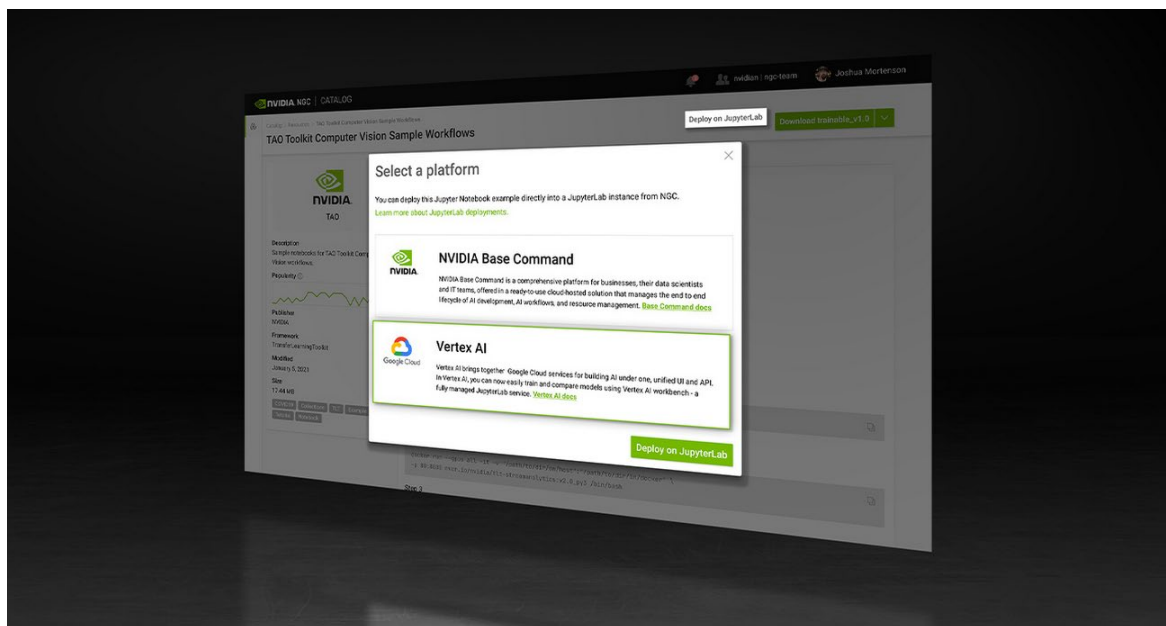
Simplifying Workflows to Speedup AI Developments

NVIDIA recently announced the One Click Deploy feature on the NVIDIA NGC catalog, the hub for GPU-optimized AI software. Developed in collaboration with Google Cloud, this feature simplifies the deployment of AI software, to a single click from the NGC catalog.
This allows data scientists to deploy frameworks, software development kits and Jupyter Notebooks directly to Google Cloud's Vertex AI Workbench, a new managed Jupyter Notebook service on top of Vertex AI, Google's service for machine learning operations.

Under the hood, this feature launches the JupyterLab instance on Google Cloud Vertex AI Workbench with optimal instance configuration, preloads the software dependencies, and downloads the NGC notebook in one go.

NGC Catalog - One Stop for AI Software

NVIDIA is expanding the rich trove of NVIDIA AI software in the catalog to ensure AI practitioners have everything they need to get started — from frameworks to models.

All of the AI models in the catalog come with credentials. They're like resumes that show the model's skills, the dataset that trained it, how to use the model and how it's expected to perform.

These model credentials provide transparency, which gives developers the confidence in picking the right model for their use case.
The NGC catalog also hosts Jupyter Notebooks tailored for the most popular AI/ML applications. Examples include:

- Computer Vision – A collection of models for detecting human actions, gestures and more.
- Automatic Speech Recognition – An end-to-end workflow for text-to-speech training.
- Recommendation – A collection of example notebooks to help build end-to-end recommendation services.

Serve Robotics uses Vertex AI and the simple easy-to-use interface that the NGC One Click Deploy delivers.

"NGC catalog allows our ML research engineers to launch environments for experiments on Vertex AI with a single click. This saves us the efforts on ML infra setup and lets researchers focus on the ML problem in the computer vision and robotics space more efficiently."—Kaiwen Yuan, Director of ML/Head of Perceptions & Predictions at Serve Robotics

Accelerate ML Deployments

Explore hundreds of Jupyter notebook examples for speech, computer vision and recommenders and, if you're just getting started with AI, browse NVIDIA's collection of Jupyter notebook examples and run it using the One Click Deploy feature on Google Cloud Vertex AI.