Google Cloud and Hugging Face Announce Strategic Partnership to Accelerate Generative AI and ML Development

Developers will be able to train, tune, and serve open models quickly and cost-effectively on Google Cloud

SUNNYVALE, Calif., Jan. 25, 2024 /PRNewswire/ -- Google Cloud and Hugging Face today announced a new strategic partnership that will allow developers to utilize Google Cloud's infrastructure for all Hugging Face services, and will enable training and serving of Hugging Face models on Google Cloud.

The partnership advances Hugging Face's mission to democratize AI and furthers Google Cloud's support for open source AI ecosystem development. With this partnership, Google Cloud becomes a strategic cloud partner for Hugging Face, and a preferred destination for Hugging Face training and inference workloads. Developers will be able to easily utilize Google Cloud's AI-optimized infrastructure including compute, tensor processing units (TPUs), and graphics processing units (GPUs) to train and serve open models and build new generative AI applications.

Google Cloud and Hugging Face will partner closely to help developers train and serve large AI models more quickly and cost-effectively on Google Cloud, including:

- **Giving developers a way to train, tune, and serve Hugging Face modelswith Vertex AI** in just a few clicks from the Hugging Face platform, so they can easily utilize Google Cloud's purpose-built, end-to-end MLOps services to build new gen AI applications.
- **Supporting Google Kubernetes Engine (GKE) deployments,**so developers on Hugging Face can also train, tune, and serve their workloads with "do it yourself" infrastructure and scale models using Hugging Face-specific Deep Learning Containers on GKE.
- **Providing more open source developers with access to Cloud TPU v5e,**which offers up to 2.5x more performance per dollar and up to 1.7x lower latency for inference compared to previous versions.
- **Adding future support for A3 VMs,**powered by NVIDIA's H100 Tensor Core GPUs, which offer 3x faster training and 10x greater networking bandwidth compared to the prior generation.
- **Utilizing Google Cloud Marketplace** to provide simple management and billing for the Hugging Face managed platform, including Inference, Endpoints, Spaces, AutoTrain, and others.

"Google Cloud and Hugging Face share a vision for making generative AI more accessible and impactful for developers," said Thomas Kurian, CEO at Google Cloud. "This partnership ensures that developers on Hugging Face will have access to Google Cloud's purpose-built AI platform, Vertex AI, along with our secure infrastructure, which can accelerate the next generation of AI services and applications."

"From the original Transformers paper to T5 and the Vision Transformer, Google has been at the forefront of AI progress and the open science movement," said Clement Delangue, CEO of Hugging Face. "With this new partnership, we will make it easy for Hugging Face users and Google Cloud customers to leverage the latest open models together with leading optimized AI infrastructure and tools from Google Cloud including Vertex AI and TPUs to meaningfully advance developers ability to build their own AI models."

Vertex AI and GKE will be available in the first half of 2024 as deployment options on the Hugging Face platform.

About Google Cloud
Google Cloud accelerates every organization's ability to digitally transform its business and industry. We deliver enterprise-grade solutions that leverage Google's cutting-edge technology, and tools that help developers build more sustainably. Customers in more than 200 countries and territories turn to Google Cloud as their trusted partner to enable growth and solve their most critical business problems.

SOURCE Google Cloud

For further information: press@google.com

Additional assets available online: Photos

Accelerate-Generative-AI-and-ML-Development