Nutanix Accelerates Enterprise Adoption of Generative AI

May 21, 2024

Company delivers an Enterprise AI foundation in collaboration with NVIDIA, Hugging Face, and ecosystem of partners to speed time to value for on-premises use cases

BARCELONA, Spain--(BUSINESS WIRE)--May 21, 2024-- **.NEXT Conference** – Nutanix (NASDAQ: NTNX), a leader in hybrid multicloud computing, today announced new functionality for Nutanix GPT-in-a-Box, including integrations with NVIDIA NIM inference microservices and Hugging Face Large Language Models (LLMs) library. Additionally, the company announced the Nutanix AI Partner Program, aimed at bringing together leading AI solutions and services partners to support customers looking to run, manage, and secure generative AI (GenAI) applications on top of Nutanix Cloud Platform and GPT-in-a-Box. Nutanix GPT-in-a-Box is a full-stack solution purpose-built to simplify Enterprise AI adoption with tight integration with Nutanix Objects and Nutanix Files for model and data storage.

“We saw a great response to our original launch of Nutanix GPT-in-a-Box, validating the needs of Enterprise customers for on-premises software solutions that simplify the deployment and management of AI models and inference endpoints,” said Thomas Cornely, SVP of Product Management at Nutanix. “Enterprise is the new frontier for GenAI and we're excited to work with our fast growing ecosystem of partners to make it as simple as possible to run GenAI applications on premises at scale while maintaining control on privacy and cost.”

Nutanix GPT-in-a-Box 2.0

The company announced GPT-in-a-Box 2.0, which will deliver expanded NVIDIA accelerated computing and LLM support, along with simplified foundational model management and integrations with NVIDIA NIMs microservices and the Hugging Face LLM library. GPT-in-a-Box 2.0 will include a unified user interface for foundation model management, API endpoint creation, end-user access key management, and will integrate Nutanix Files and Objects, plus NVIDIA Tensor Core GPUs.

GPT-in-a-Box 2.0 will bring Nutanix simplicity to the user experience with a built-in graphical user interface, role-based access control, auditability, and dark site support, among other benefits. It will also provide a point-and-click-user interface to deploy and configure NVIDIA NIM, part of the NVIDIA AI Enterprise software platform for the development and deployment of production-grade AI, to easily deploy and run GenAI workloads in the Enterprise and at the Edge.

“We've partnered with Nutanix as one of our key technology partners to enable our AI ambitions while empowering a future where technology serves humanity,” said Khalid Al Kaf, COO at Yahsat. “We are not just keeping pace with the future; we're actively shaping it, leveraging Nutanix GPT-in-a-Box, which provides us with simple, end-to-end management capabilities and ability to maintain control over our data.”

Partnership with Hugging Face to Deliver Integrated Access to LLM Library

Nutanix also announced a partnership with Hugging Face to help accelerate customers’ AI journey by providing integrated access to the Hugging Face library and execution of LLMs for Nutanix customers. Joint customers will be able to leverage Nutanix GPT-in-a-Box 2.0 to easily consume validated LLMs from Hugging Face and execute them efficiently.

Through this partnership, Nutanix and Hugging Face will develop a custom integration with Text Generation Inference, the popular Hugging Face open-source library for production deployment of Large Language Models, and enable text-generation models available on the Hugging Face Hub within Nutanix GPT-in-a-Box 2.0. It will deliver a seamless workflow to deploy validated AI LLMs from Hugging Face with full support from Nutanix, significantly expanding the number of supported LLMs. This will create a jointly validated and supported workflow for Hugging Face libraries and LLMs, ensuring customers have a single point of management for consistent model inference.

Strengthened Unstructured Data Platform for AI/ML

Nutanix also enhanced its unstructured data platform for AI/ML and other applications with increased performance, density, and TCO. Nutanix Unified Storage (NUS) now supports a new 550+ Terabytes dense low-cost all-NVMe platform and up to 10 Gigabyte/second sequential read throughput from a single node (close to line speed for a 100 Gigabit ethernet port), enabling faster data reads and more efficient use of GPU resources. Nutanix will also add support for NVIDIA GPUDirect Storage to further accelerate AI/ML applications. Additionally, to protect the extremely valuable and often confidential data such as data sets used to train and process AI/ML workloads, Nutanix Data Lens extends cyber resilience value to Objects in addition to Files data. A new Data Lens Frankfurt-based point of presence enables broader adoption in EU customers, meeting their own compliance needs.

Nutanix has collaborated with major server OEMs to provide customers breadth and choice with a wide range of AI-optimized GPUs and density-optimized GPU systems. These AI-optimized GPUs, including the NVIDIA L40S, H100, L40, and L4 GPUs, are now supported on Nutanix GPT-in-a-Box. Nutanix also now supports density-optimized GPU systems from Dell, HPE, and Lenovo to help lower the total cost of ownership by allowing customers to deploy a smaller number of systems to meet their workload demands. The company also announced planned support for NX-9151 which is based on the NVIDIA MGX reference architecture.

Nutanix AI Partner Program

To further support customers’ AI strategies, Nutanix announced the new AI Partner Program providing customers with simplified access to an expanded ecosystem of AI partners to deliver real-world GenAI solutions. Partners will help organizations build, run, manage, and secure third-party and homegrown GenAI applications on top of Nutanix Cloud Platform and the Nutanix GPT-in-a-Box solution, targeted at prominent AI use cases.

This broad ecosystem of partners will help address diverse use cases including operations, cybersecurity, fraud detection, and customer support, across verticals such as healthcare, financial services, and legal and professional services. Initial partners include: Codeium, DataRobot, DKube, Instabase, Lamini, Neural Magic, Robust Intelligence, RunAI, and UbiOps.

Program benefits to partners include:

- **Nutanix AI Ready validation:** All partners will receive Nutanix AI Ready validation to demonstrate interoperability with Nutanix Cloud Infrastructure, Nutanix AHV hypervisor, and Nutanix Kubernetes Platform.
- **Full-stack solutions:** Partners in the program will benefit from individual solution briefs highlighting the benefits of the Nutanix and partner solution. Additionally, select partners will receive a tech note, best practice guide, or reference architecture to simplify implementation with joint customers.
- **Promotional and go-to-market alignment:** Partners will benefit from go-to-market and demand generation activities centered around the AI use case for enterprises. They will be provided with access to promotional and go-to-market opportunities including content creation, blog posts, webinars, podcasts, and other events, to help drive awareness and demand for key use cases.

Nutanix GPT-in-a-Box 2.0 is expected to be available in the second half of 2024. Support for NVIDIA GPU Direct and NX-9151 are currently under development. Additional features announced in NUS as well as Data Lens are currently available to customers. More information can be found here.

Supporting Quotes:

- "Our partnership with Nutanix will help extend Hugging Face to more enterprises looking for private and secure control of their LLMs," says Clem Delangue, CEO and co-founder, Hugging Face. "Our mission is to enable organizations of all sizes to build their own AI with open source. Our work with Nutanix will make it easy for enterprises to build AI applications on premises leveraging Text Generation Inference."
- "Customers looking to leverage GenAI have a responsibility to maintain control over sensitive, private data," said Dave Pearson, IDC's Research Vice President for Infrastructure. "Through these new features and partnerships, Nutanix looks to enable organizations to build a private and secure AI-ready platform and accelerate GenAI adoption in the Enterprise."

Additional Resources:

- Blog: GPT-in-a-Box 2.0 is Here With Four Ways to Level Up Your GenAI
- Blog: Accelerating the Delivery of Complete AI Solutions for Organizations: Announcing the Nutanix AI Partner Program
- Join the Alliance Partner Program
- Nutanix AI
- Nutanix Validated Design for GPT-in-a-Box
- NVIDIA MGX

About Nutanix

Nutanix is a global leader in cloud software, offering organizations a single platform for running apps and data across clouds. With Nutanix, companies can reduce complexity and simplify operations, freeing them to focus on their business outcomes. Building on its legacy as the pioneer of hyperconverged infrastructure, Nutanix is trusted by companies worldwide to power hybrid multicloud environments consistently, simply, and cost-effectively. Learn more at www.nutanix.com or follow us on social media @nutanix.

View source version on businesswire.com: https://www.businesswire.com/news/home/20240521060096/en/

Gabrielle Moynan
pr@nutanix.com

Source: Nutanix