Nutanix Extends AI Platform to Public Cloud

November 12, 2024

Nutanix Enterprise AI provides an easy-to-use, unified generative AI experience on-premises, at the edge, and now in public clouds

SALT LAKE CITY, Nov. 12, 2024 (GLOBE NEWSWIRE) -- KubeCon – Nutanix (NASDAQ: NTNX), a leader in hybrid multicloud computing, today announced that it extended the company's AI infrastructure platform with a new cloud native offering, Nutanix Enterprise AI (NAI), that can be deployed on any Kubernetes platform, at the edge, in core data centers, and on public cloud services like AWS EKS, Azure AKS, and Google GKE. The NAI offering delivers a consistent hybrid multicloud operating model for accelerated AI workloads, enabling organizations to leverage their models and data in a secure location of their choice while improving return on investment (ROI). Leveraging NVIDIA NIM for optimized performance of foundation models, Nutanix Enterprise AI helps organizations securely deploy, run, and scale inference endpoints for large language models (LLMs) to support the deployment of generative AI (GenAI) applications in minutes, not days or weeks.

Generative AI is an inherently hybrid workload, with new applications often built in the public cloud, fine-tuning of models using private data occurring on-premises, and inferencing deployed closest to the business logic, which could be at the edge, on-premises or in the public cloud. This distributed hybrid GenAI workflow can present challenges for organizations concerned about complexity, data privacy, security, and cost.

Nutanix Enterprise AI provides a consistent multicloud operating model and a simple way to securely deploy, scale, and run LLMs with NVIDIA NIM optimized inference microservices as well as open foundation models from Hugging Face. This enables customers to stand up enterprise GenAI infrastructure with the resiliency, day 2 operations, and security they require for business-critical applications, on-premises or on AWS Elastic Kubernetes Service (EKS), Azure Managed Kubernetes Service (AKS), and Google Kubernetes Engine (GKE).

Additionally, Nutanix Enterprise AI delivers a transparent and predictable pricing model based on infrastructure resources, which is important for customers looking to maximize ROI from their GenAI investments. This is in contrast to hard-to-predict usage or token-based pricing.

Nutanix Enterprise AI is a component of Nutanix GPT-in-a-Box 2.0. GPT-in-a-Box also includes Nutanix Cloud Infrastructure, Nutanix Kubernetes Platform, and Nutanix Unified Storage along with services to support customer configuration and sizing needs for on-premises training and inferencing. For customers looking to deploy in public cloud, Nutanix Enterprise AI can be deployed in any Kubernetes environment but is operationally consistent with on-premises deployments.

“With Nutanix Enterprise AI, we're helping our customers simply and securely run GenAI applications on-premises or in public clouds. Nutanix Enterprise AI can run on any Kubernetes platform and allows their AI applications to run in their secure location, with a predictable cost model,” said **Thomas Cornely, SVP, Product Management, Nutanix**.

Nutanix Enterprise AI can be deployed with the NVIDIA full-stack AI platform and is validated with the NVIDIA AI Enterprise software platform, including NVIDIA NIM, a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing. Nutanix-GPT-in-a-Box is also an NVIDIA-Certified System, also ensuring reliability of performance.

"Generative AI workloads are inherently hybrid, with training, customization, and inference occurring across public clouds, on-premises systems, and edge locations," **said Justin Boitano, vice president of enterprise AI at NVIDIA**. "Integrating NVIDIA NIM into Nutanix Enterprise AI provides a consistent multicloud model with secure APIs, enabling customers to deploy AI across diverse environments with the high performance and security needed for business-critical applications."

Nutanix Enterprise AI can help customers:

- **Address AI skill shortages.** Simplicity, choice, and built-in features mean IT admins can be AI admins, accelerating AI development by data scientists and developers adapting quickly using the latest models and NVIDIA accelerated computing.
- **Remove barriers to building an AI-ready platform.** Many organizations looking to adopt GenAI struggle with building the right platform to support AI workloads, including maintaining consistency across their on-premises infrastructure and multiple public clouds. Nutanix Enterprise AI addresses this with a simple UI-driven workflow that can help customers deploy and test LLM inference endpoints in minutes, offering customer choice with support for NVIDIA NIM microservices which run anywhere, ensuring optimized model performance across cloud and on prem environments. Hugging Face and other model standards are also supported. Additionally, native integration with Nutanix Kubernetes Platform keeps alignment with the ability to leverage the entire Nutanix Cloud Platform or provide customers with the option to run on any Kubernetes runtime, including AWS EKS, Azure AKS, or Google Cloud GKE with NVIDIA accelerated computing.
- **Mitigate data privacy and security concerns.** Helping mitigate privacy and security risks is built into Nutanix Enterprise AI by enabling customers to run models and data on compute resources they control. Additionally, Nutanix Enterprise AI delivers an intuitive dashboard for troubleshooting, observability, and utilization of resources used for LLMs, as well as quick and secure role-based access controls (RBAC) to ensure LLM accessibility is controllable and understood. Organizations requiring hardened security will also be able to deploy in air-gapped or dark-site environments.
- **Bring enterprise infrastructure to GenAI workloads.** Customers running Nutanix Cloud Platform for business-critical applications can now bring the same resiliency, Day 2 operations, and security to GenAI workloads for an enterprise infrastructure experience.

Key use cases for customers leveraging Nutanix Enterprise AI include: enhancing customer experience with GenAI through analysis of customer

feedback and documents; accelerating code and content creation by leveraging co-pilots and intelligent document processing; leveraging fine-tuning models on domain-specific data to accelerate code and content generation; strengthening security, including leveraging AI models for fraud detection, threat detection, alert enrichment, and automatic policy creation; and improving analytics by leveraging fine-tuned models on private data.

Nutanix Enterprise AI, running on-premises, at the edge or in public cloud, and Nutanix GPT-in-a-Box 2.0 are currently available to customers. For more information, please visit Nutanix.com/enterprise-ai.

Supporting Quotes:

- "Thanks to the deep collaboration between the Nutanix and Hugging Face teams, customers of Nutanix Enterprise AI are able to seamlessly deploy the most popular open models in an easy to use, fully tested stack – now also on public clouds," said **Jeff Boudier, Head of Product at Hugging Face.**
- "By providing a consistent experience from the enterprise to public cloud, Nutanix Enterprise AI aims to provide a user-friendly infrastructure platform to support organizations at every step of their AI journey, from public cloud to the edge," said **Dave Pearson, Infrastructure Research VP at IDC.**

About Nutanix
Nutanix is a global leader in cloud software, offering organizations a single platform for running applications and managing data, anywhere. 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.

Media Contact pr@nutanix.com