April 2025 | eBook

Enterprise AI on Nutanix

The simplest path is the fastest

NUTANIX

Table of Contents

Fast Track Your Enterprise AI

The emergence of generative AI (GenAI) has prompted organizations to rethink their AI plans and accelerate timetables.

However, in a recent survey of 650 IT decision-makers, DevOps and platform engineering professionals, many organizations aren't sure how to get started with GenAI. Chief among their concerns are:

- **IT infrastructure deficits:** 91% of respondents said IT infrastructure needs to be improved to support AI workloads.
- **Ensuring data security and resilience:** Data security, resilience and scalability are key challenges for enterprise AI.
- **Lack of skills:** 100% of organizations said they require additional skills to support AI-related initiatives over the next 12 months.

Before deploying AI, consider these three general approaches:

- **Rely exclusively on cloud.** Using the cloud for AI experimentation, development and production of AI apps is an option, but it can be expensive. You'll compete for cloud resources with everyone else and you may be putting your data at risk.
- **Build a custom AI stack in datacenters and at the edge.** This approach takes significant expertise and time and should only be considered if you have in-house expertise or a committed partner. Otherwise, your AI progress is likely to be slow and missteps frequent.
- **Choose a turnkey solution.** A turnkey solution can enable even inexperienced teams to get started quickly with best-in-class hardware and AI software that increases flexibility.

This eBook examines the pain points that teams face as they plan for AI and explains how the Nutanix Cloud Platform and Nutanix GPT-in-a-Box can simplify the path to production for AI – on-premises, in the cloud and at the edge – yielding better results in less time.

AI Pain Points

The majority of organizations **(85%)** plan to purchase existing AI models or leverage open-source AI models to build AI apps, with only **10%** planning to build their own models.

For many companies, there's simply no need to design and train AI models from scratch. It's easier to license and extend pretrained foundation models with your private data, using either:

- **Fine-tuning:** Use private data to fine-tune the model to fit your needs.
- **Retrieval-augmented generation (RAG):** Data from outside the model is used to deliver results tailored to your company.

With either option, as you transition from experimentation to production deployment of AI, you will face significant challenges in the following areas.

Complexity

Leveraging GenAI requires the ability to deploy containerized large language models (LLMs), fine-tune them or implement RAG, and deploy to production over and over again using MLOps.

To accomplish this, you may need infrastructure appropriate for fine-tuning and infrastructure appropriate for inferencing. Inferencing is often done at the edge – close to the consumer of the AI service – to decrease latency and increase responsiveness.

- Inferencing at the edge creates remote infrastructure management challenges.
- Inferencing and training, whether in the datacenter or at the edge, require the right mix of compute, GPU and storage resources so you can avoid over-provisioning while scaling quickly as demands change.

To bulletproof your AI operations, you will need intelligent tools that can help you determine whether the root cause of any problems that arise is a result of the model, the infrastructure or a weakness in your MLOps processes.

Compliance

AI teams are frequently organized separately from IT and often focus on developing solutions without proper consideration for mission-critical resiliency, data management and the protection of personally identifiable information (PII).

It can be a struggle to deploy and operate GenAI or other AI-based apps without compromising compliance with IT policies for security, data protection, resilience, and other Day 2 operational needs.

All training data must be scrubbed to ensure it doesn't contain any PII, such as names, addresses, phone numbers, social security numbers, financial information, and credit card numbers.

Cost

While the cloud offers some advantages for AI operations, it can be much more expensive to run an AI model in the cloud versus on-premises.

On the other hand, AI infrastructure is notoriously power-hungry. Power, cooling, physical space, and other Day 2 considerations have to be factored in when you consider running AI workloads in your datacenter or at the edge.

Governance

There are several governance challenges associated with AI.

- **Data management and data sovereignty:** Consistent, policy-based data protection and security across edge, core and cloud can be a challenge. Data generated at the edge is often needed for further training. Simple, repeatable and automated methods are essential for moving and managing data between locations.

  Sometimes data created at the edge needs to be reliably stored and protected in place to ensure data sovereignty regulations aren't violated. If training is required using that data, you may need to make provisions to move the model to the data and train it there.

  If you have multiple such locations, a federated approach may be necessary, in which your model is trained successively at each location.

- **MLOps:** AI teams must be able to accurately track and maintain complete information on each model version, including the exact data sets it was trained on and when and where the model was deployed.

  Depending on the frequency of training iterations and the number and complexity of data sets, this can be significantly more complex than tracking code revisions in traditional software development.

Security and data privacy

A number of additional security concerns come into play with AI.

- **Data leakage:** There's a strong temptation to feed company information – which could include confidential or proprietary data – into non-sovereign AI services running in the cloud.

  This could be something as seemingly harmless as feeding the text of an unpublished report to Copilot, Gemini or ChatGPT and asking it to create a summary or using OpenAI's Whisper model to generate a transcript of an internal meeting.

  But what happens to your data after the model is finished with it? Microsoft, Google, AWS, and OpenAI may be trustworthy, but an entire ecosystem of generative apps and services is emerging, which increases risk.

- **Data poisoning:** Malicious injection of unauthorized data into a LLM, even if that model is under your control, could lead to inaccurate generative content known as hallucinations and bias.

- **Prompt injection:** Direct prompts can be inserted to maliciously circumvent AI guardrails and manipulate LLM responses.

- **Security tooling:** Your current security tools may be unable to align with your MLOps, requiring either an upgrade or completely new tools.

Nutanix Cloud Platform uses innovative software and a unique hyperconverged infrastructure (HCI) to address these challenges on-premises, at the edge and in the public cloud.

Why Nutanix Cloud Platform for AI?

Nutanix is intensely focused on reducing infrastructure complexity and enabling hybrid multicloud operations. Built on our proven HCI, Nutanix Cloud Platform (NCP) provides an agile, resilient infrastructure that satisfies your AI needs from the edge to the core datacenter to the public cloud.

Simplify Your AI Operations

Operational complexity is another big challenge on the path to production deployment of AI. Given resource constraints, many organizations do a lot of their AI experimentation in the cloud.

But how do you take work that was done in the cloud and move it into the datacenter or operationalize it at the edge?

NCP removes infrastructure complications and limitations to jumpstart your AI and machine learning initiatives. With Nutanix, you can achieve seamless app, workload and data mobility across edge, datacenter and cloud environments.

Because NCP runs anywhere, your teams work efficiently across all environments.

NCP at the Edge

The edge can be a particular pain point because of the need to deploy and manage complex infrastructure remotely. With its proven HCI design – combining compute, networking and storage – NCP uniquely satisfies the needs of edge deployments with:

- A compact footprint
- Easy remote management and remote app deployment.
- Advanced data protection and security.
- Untethered operations.
- A full suite of data services.

Nutanix GPT-in-a-Box

Many enterprises are struggling with taking advantage of GenAI, especially for use cases that cannot run in the public cloud due to data sovereignty, governance and privacy concerns.

Nutanix GPT-in-a-Box delivers ready-to-use, turnkey, private AI that allows you to fine-tune and run LLMs and other AI models while maintaining full control.

Nutanix GPT-in-a-Box addresses the complexity, scaling and security challenges that you face when developing GenAI apps. Full-stack, software-defined and AI-ready, GPT-in-a-Box runs on NCP to simplify deployment and speed-up your AI initiatives.

GPT-in-a-Box delivers:

- **A full-stack AI platform.** Use your choice of hardware, CPUs or GPUs, VMs or containers, and LLMs and AI frameworks, anywhere you deploy GenAI.
- **Built-in data services for AI.** Secure and expansive data services for files, blocks and objects with unified snapshot and disaster recovery controls.
- **Scalable AI workloads.** Deliver scalable AI anywhere with less effort using a consistent set of platform services that unify cloud operations.
- **Deploy and adapt with ready-to-use LLMs:** Access validated GenAI models that can be deployed quickly from edge to cloud while accelerating time-to-value.
- **Deploy and operate secure APIs for AI apps:** Easily create secure, role-based APIs to connect apps to your AI models.

With Nutanix GPT-in-a-Box, you can maintain AI training data and models internally to help meet security, privacy and compliance requirements while optimizing IT costs.

GPT-in-a-Box includes everything you need to get started running GenAI models. All you supply is your foundation model of choice.

Flexible GPU and CPU Options

Whether you're doing inferencing or training or both, having the right GPUs and CPUs is an essential element of AI success.

GPUs

Nutanix supports a full range of NVIDIA GPUs to address diverse needs. NCP supports GPU pass-through, allowing you to utilize GPUs efficiently in virtualized or containerized environments.

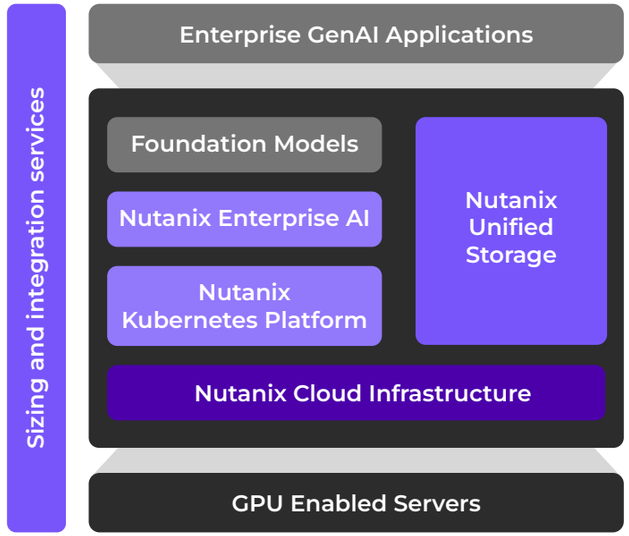
GPU pass-through gives apps running on VMs direct access to GPU resources. Nutanix provides a cluster-wide view of GPUs, allowing you to allocate any available GPU to a VM. Multiple GPUs can be allocated to a single VM. With pass-through, only one VM can use a GPU at any given time.

CPUs

While many have the impression that GPUs are essential for AI, the latest generations of CPUs include enhancements to accelerate training and inference, making them suitable to support some AI operations without GPUs.

The right CPUs can also deliver the parallelism, memory capacity and bandwidth necessary to get the most from GPUs.

Available services include Planning Workshop, Stack Design Workshop and Stack Deployment.



AI-ready bundled services help you size and configure infrastructure suitable to deploy a curated set of LLMs using the leading open-source AI and MLOps frameworks.

Industry Partnerships

Nutanix partners with industry-leading AI companies to deliver a full stack solution.

Through our partnership with NVIDIA, Nutanix AHV hypervisor has been certified to run NVIDIA AI Enterprise, a cloud-native software platform that streamlines development and deployment of production-grade AI solutions.

This includes AI agents, computer vision, speech AI, and more. Certification with NVIDIA AI Enterprise ensures that all this functionality is available and works as expected on NCP.

NCP pairs well with NVIDIA AI Enterprise, creating an environment that fosters agility, efficiency and scalability. Integration with NVIDIA GPU compute accelerators ensures that AI workloads get direct access to GPU resources to reduce latency and accelerate performance.

In addition to NVIDIA, Nutanix has close partnerships with Intel and AMD. As a result, we can closely track AI innovations from these technology leaders and add support to our platform.

AI Use Cases

For many enterprises, GenAI use cases – such as intelligent chatbots, support copilots, smart document processing– are the highest priority. However, Nutanix gives you multiple options to accommodate a full range of AI use cases.

NVIDIA AI Enterprise and NVIDIA NIM

Because Nutanix is certified to run NVIDIA AI Enterprise, we support a wide variety of models spanning diverse use cases and domains, including computer vision, speech, language understanding, and molecule generation.

For NVIDIA AI Enterprise customers, GPT-in-a-Box makes it easy to deploy NVIDIA NIM, a set of optimized cloud-native microservices for GenAI.

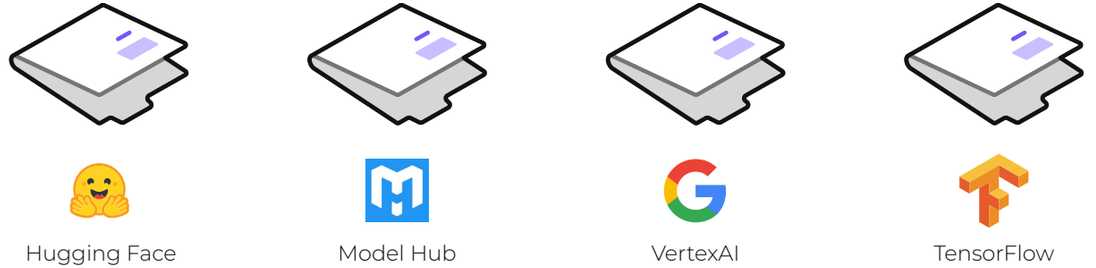
Hugging Face and other Model Hubs

Nutanix has partnered with Hugging Face to fast-track model deployment, enabling you to easily integrate LLMs from Hugging Face with GPT-in-a-Box. Deliver a seamless workflow to search, download and deploy validated AI LLMs with full support.

GPT-in-a-Box also allows you to upload and deploy non-validated and unsupported models of your choice.

Model Hubs

Repositories for opoen-source models with permissive licenses for usage



Getting Started with Nutanix

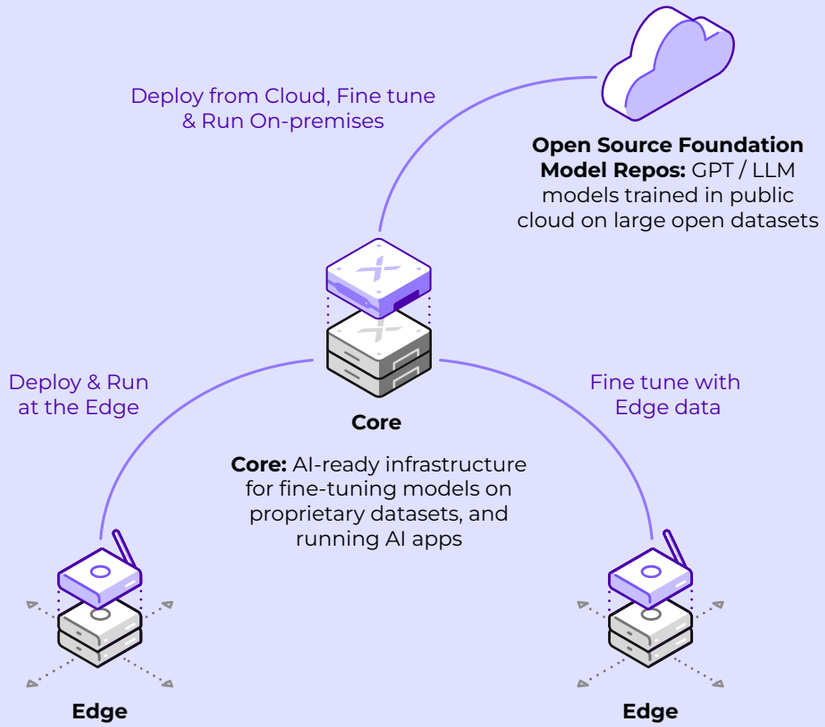
Nutanix is focused on eliminating the pain points and keeping you on a path to successfully deploy GenAI and other AI apps in production.

NCP simplifies AI operations from edge to core to cloud, simplifies data services, and delivers the data protection and security you need to move forward with confidence. Coupled with the turnkey GPT-in-a-Box solution, Nutanix makes it even easier to get started with GenAI.

To learn more about Nutanix Enterprise AI solutions, visit our AI solutions page and be sure and take our AI test drive.

Take a Test Drive

Simplify AI Operations from Core to Edge to Cloud



NUTANIX

info@nutanix.com | www.nutanix.com | @nutanix

filename here including version followed by date AI-EnterpriseAIOnNutanix-eBook-FY25Q3-v2 04302025