Enterprise AI - Nutanix Enterprise AI

[PDF generated June 03 2025. For all recent updates please see the Nutanix Bible releases notes located at https://nutanixbible.com/release_notes.html. Disclaimer: Downloaded PDFs may not always contain the latest information.]

Key Use Cases of Enterprise AI

Key use cases include:

- Creating Better Security
  - Build on traditional AI capabilities like fraud detection and threat monitoring by using generative AI to enable more adaptive threat simulations, automate incident response, and generate realistic data for training and testing security systems.

- Accelerating Code & Content Creation
  - Enable code co-pilots, intelligent document processing, and fine-tuned models trained on domain-specific data to significantly accelerate the development of software and the generation of high-quality content.

- Supercharging the Customer Experience
  - Leverage advanced analytics to understand customer feedback, deploy personalized chatbots, and deliver tailored interactions that drive deeper engagement and satisfaction.

Challenges with Enterprise AI

Typical challenges implementing Enterprise AI include:

- Not knowing where to begin
  - The necessary skillsets are still evolving, making it challenging to find and hire qualified talent.

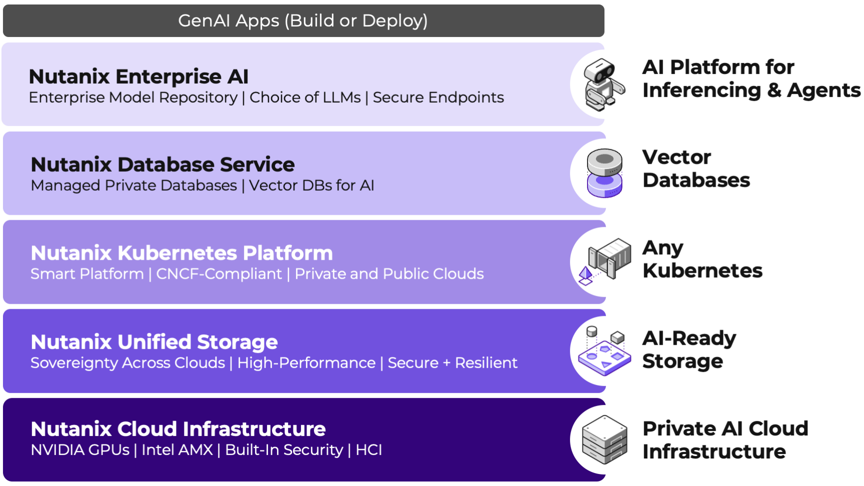
- Ensuring that intellectual property and data stay secure and private
  - While public cloud solutions are easy to deploy, they may not always meet requirements for data sovereignty and security.

- Needing help to get a working solution
  - Application developers and IT teams often have distinct expertise and requirements, so a platform that supports both groups is essential.

Building an Enterprise AI platform

Building an Enterprise AI platform involves integrating multiple components and technologies to ensure scalability, reliability, and performance.

- Private AI Cloud Infrastructure
- AI-Ready Storage
- Kubernetes
- Vector Databases
- AI Platform for Inferencing & Agents

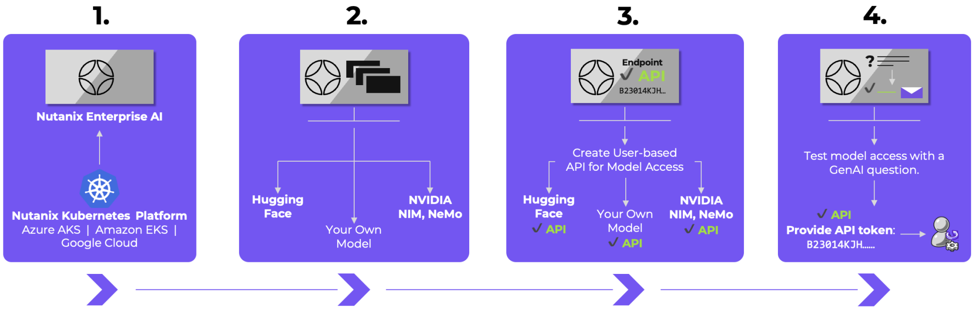

GenAI Apps (Build or Deploy)
Nutanix Enterprise AI
Enterprise Model Repository | Choice of LLMs | Secure Endpoints
AI Platform for Inferencing & Agents
Nutanix Database Service
Managed Private Databases | Vector DBs for AI
Vector Databases
Nutanix Kubernetes Platform
Smart Platform | CNCF-Compliant | Private and Public Clouds
Any Kubernetes
Nutanix Unified Storage
Sovereignty Across Clouds | High-Performance | Secure + Resilient
AI-Ready Storage
Nutanix Cloud Infrastructure
NVIDIA GPUs | Intel AMX | Built-In Security | HCI
Private AI Cloud Infrastructure

Nutanix Enterprise AI

Nutanix Enterprise AI is a Kubernetes application that enables the AI platform component of the Enterprise AI stack and enables IT organizations to manage and deploy LLMs and inference endpoints. Nutanix Enterprise AI can be deployed on:

- Nutanix Kubernetes Platform
- Amazon EKS
- Azure AKS
- Google Cloud GKE

How Nutanix Enterprise AI works



1. Deploy and run Nutanix Enterprise AI on Kubernetes.
2. Login to the interface and deploy your choice of LLM from Hugging Face, NVIDIA, or import your own custom model.
3. Create a secure endpoint and API key.
4. Test the model directly from the UI before sending the token to the application developer or data scientist.

Then, monitor and manage the endpoint usage, infrastructure, events, and other metrics for understanding how the organization is using AI and troubleshoot any issues.

Nutanix Enterprise AI Key Features

- Elegant user interface
- Choice of AI Models (LLMs) from Hugging Face or NVIDIA NIM, including Text Generation, Safety, Embedding, Reranker, and Vision models
- Upload Your Own AI Models (LLMs)
- API Token Creation and Management
- Partner API Token Management for Hugging Face and NVIDIA NIM
- API Code Samples
- Role-Based Access Controls (RBAC)
- AI Model Preflight Testing

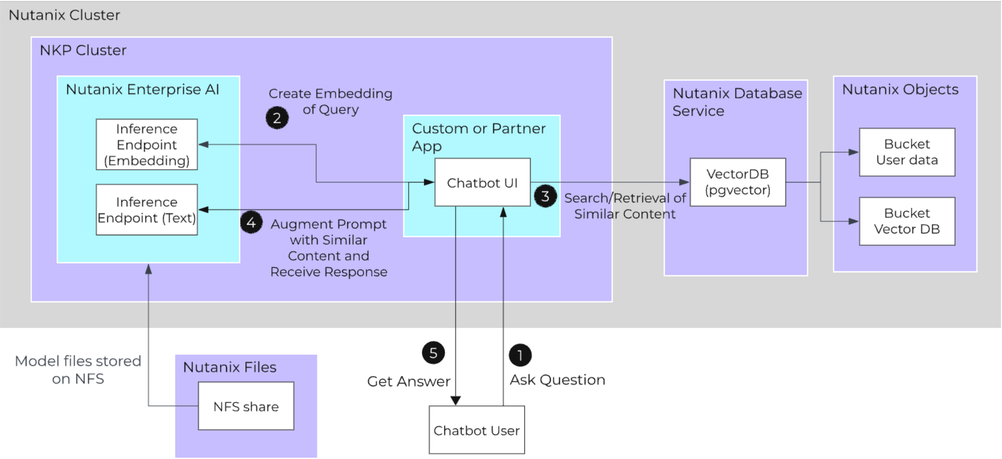
- AI Model and API Monitoring
- Kubernetes Resource Monitoring
- GPU Usage Monitoring
- Event Auditing
- Integrated Nutanix Pulse Reporting
- Export Metrics with OpenTelemetry
- Hibernate and Resume Endpoints

Example Use Case - Retrieval Augmented Generation

Open source LLMs, such as Meta's Llama, are pre-trained on vast amounts of data from the internet, but may not know anything about your own organization. For example, if you asked about your next company holiday, it might know about national holidays, but not holidays specific to your organization. That's where Retrieval Augmented Generation (RAG) comes in.

A crucial part of RAG is the document store and vector database. The typical workflow involves ingesting documents from file or object storage, processing them with a function that splits and embeds the content, and then storing these embeddings in a vector database. Both open source and commercial tools are available to streamline this process.

Once your documents have been embedded, the end-user workflow of a RAG-enabled chatbot looks similar to the below diagram.



1. **Ask Question**
   - User asks a question to the chatbot.
2. **Create Embedding of Query**
   - Instead of going directly to the inference API, the application will first create an embedding of the query using an embedding model hosted on Nutanix Enterprise AI.
3. **Search/Retrieval of Similar Content**
   - With that embedding, the application will search for similar embeddings in the vector database that has been populated with the embeddings of source documents.
4. **Send Prompt to Inference API**
   - The application augments the user's prompt with the found context and sends this to a text generation model hosted on Nutanix Enterprise AI.
5. **Get Answer**
   - The chatbot returns an answer to the user.

For more information for designing and implementing a Retrieval Augmented Generation workflow, check out the Nutanix Validated Design.

Other Resources

To learn more about Nutanix Enterprise AI and to see it in action, check out the following resources:

- Try Nutanix Enterprise AI by taking a Test Drive
- AI on Nutanix YouTube playlist