GenAI RAG Solution with Nutanix and NVIDIA

NUTANIX

Table of Content

Executive Summary

Businesses are turning to AI to enhance engagement, optimize operations, and stay competitive. To provide the best business benefit, AI models require vast, high-quality data to ensure accurate insights. Outdated or insufficient data leads to a much less effective business benefit, and at worst, misinformation that can bring a negative impact to your business.

This document explains how the integrated Nutanix and NVIDIA solution, featuring Nutanix Unified Storage which is NVIDIA-Certified 'Enterprise' for Files, enables enterprises to accelerate AI workloads seamlessly through a scalable, high performance data platform. By incorporating advanced capabilities such as Retrieval-Augmented Generation (RAG) and vector databases, the solution empowers organizations to enrich Large Language Models (LLMs) with both proprietary and external data, allowing accurate, context-aware, and real-time responses.

Overview

Adopting a RAG framework instead of solely depending on standard foundational LLMs offers several key benefits, including gaining response accuracy and safety, the lessening of hallucinations, and fewer ambiguous or misleading outputs. Moreover, deploying RAG and foundational LLMs within an organization's on-prem infrastructure allows the access to sensitive data to remain internal, which is tougher to guarantee when using cloud-hosted LLM services.

Vector databases play a crucial role in RAG models by enabling efficient storage, indexing, and retrieval of supplemental data, whether proprietary or external to conventional LLM datasets. By supporting high-dimensional vector embeddings, these databases work to lift the generation process through capabilities like similarity search, document summarization, and content generation, optimizing the retrieval phase in RAG workflows.

As a key NVIDIA partner and NVIDIA-Certified 'Enterprise' Storage partner, Nutanix delivers a performant and secure infrastructure stack that simplifies deployment and accelerates generative AI adoption while helping support compliance, efficiency, and seamless data management.

Vector Databases on Nutanix Unified Storage

Vector databases differ from traditional databases by specializing in the handling of vector embeddings and numerical arrays that capture object characteristics. Designed for scalability and high-speed similarity searches, they leverage advanced indexing techniques like Locality-Sensitive Hashing (LSH) and Approximate Nearest Neighbor (ANN) algorithms.

Nutanix Unified Storage is a flexible, software defined data platform that makes it easy to manage all your data, whether its structured or unstructured, across block, file and object storage. It's designed for performance, high availability and flexible data mobility, with built-in support for GPU direct storage and multiprotocol access. Our NVIDIA-certified Nutanix Files solution when combined with our Nutanix Objects, can meet the demands of next generation AI, analytics and enterprise applications.

NUS enhances vector database performance by providing a scale-out, software-defined solution which leverages all NVMe architectures. The solution also optimizes storage capacity and compute resources to drive linear

performance as you grow, all while promoting the minimizing of resource waste, power consumption, and space utilization.

Solution Benefits

The NUS RAG solution, powered by NVIDIA, is redefining performance, scalability, and simplicity changing traditional solutions in every aspect. This advanced solution utilizes intelligent data indexing and vector database (DB) to transform the way organizations store, retrieve, and maximize the value of their data. Key benefits include:

- **Unified Data Integration**: The NUS RAG solution effortlessly integrates diverse data sources, including brochures, internal web pages, emails, PDFs, text documents, photos, and knowledge base articles. These are converted into vectors and stored in a vector DB, establishing a direct and efficient connection between user queries and the most relevant documents for fast, accurate insights
- **Advanced Query Matching**: User queries are transformed into vectors using the same embedding model as the indexed documents, enabling precise, similarity-based search matching. This approach allows the most relevant information to be retrieved, enhancing query accuracy, contextual awareness, and delivering faster, more meaningful results across the Nutanix ecosystem.
- **Augmented Query Processing**: Retrieved documents enrich user queries with additional context, designed to improve the LLM's accuracy, timeliness, and relevance. This integration aligns AI-generated insights with business-specific knowledge, enabling users to receive precise, valuable information to drive informed, context-aware decisions.
- **Enterprise-grade Performance and Scalability**: NUS RAG platform with NVIDIA enables high-speed data retrieval and indexing, powering large-scale vector databases for real-time AI-driven analytics and decision-making. Its independent scaling of performance and capacity optimizes costs while efficiently handling intensive AI workloads, ensuring seamless growth and enterprise-wide operational efficiency.
- **Seamless usability and integration**: The RAG platform supports rapid deployment and seamless integration into enterprise workflows. Its user-friendly architecture enables organizations to leverage advanced AI capabilities effortlessly, accelerating adoption.
- **Data Security**: Nutanix Unified Storage incorporates robust data security features tailored for AI/ML workloads and RAG Pipeline, promoting protection, compliance, and visibility throughout the data lifecycle. Here are the key aspects:
    - **Proactive Ransomware Protection**: Nutanix Data Lens continuously monitors data activity to detect anomalies and potential ransomware threats. It leverages both pattern recognition and behavioural analysis to detect suspicious activity, helping to protect your data by restricting client access and notifying administrators of potential anomalies
    - **Comprehensive Auditing and Governance**: Data Lens provides detailed audit trails of file access and user behaviour, helping organizations meet compliance requirements and maintain data integrity across environments.
    - **Data Privacy and Access Controls**: NUS supports fine-grained access controls and encryption to monitor access of sensitive data, aligning with enterprise-grade privacy standards.

Solution Use Cases

RAG offers a transformative approach to integrate large language models with an organization's unique knowledge and data. This integration of an LLM's deep learning capabilities with an organization's proprietary data enables RAG to effectively overcome the challenges of deploying LLMs in real-world applications. Below are detailed explanations of how RAG drives value across various use cases.

Semantic Search and Information Retrieval

RAG enhances semantic search, enabling enterprises to efficiently extract relevant insights from vast data repositories. Its precision and contextual understanding are crucial in industries where accuracy matters. For example, a healthcare organization can quickly retrieve patient case studies, medical guidelines, or clinical trial data, while a pharmaceutical company can efficiently locate specific drug research papers.

By understanding query intent and matching it with contextually relevant content, RAG streamlines information retrieval, improving both efficiency and accuracy.

Improving Customer Service

Traditional customer service chatbots are often limited by predefined responses. A Nutanix-powered RAG chatbot leverages proprietary and real-time data to provide personalized and accurate customer support. This capability benefits industries requiring high levels of accuracy, such as finance and healthcare.

Optimizing Content Creation

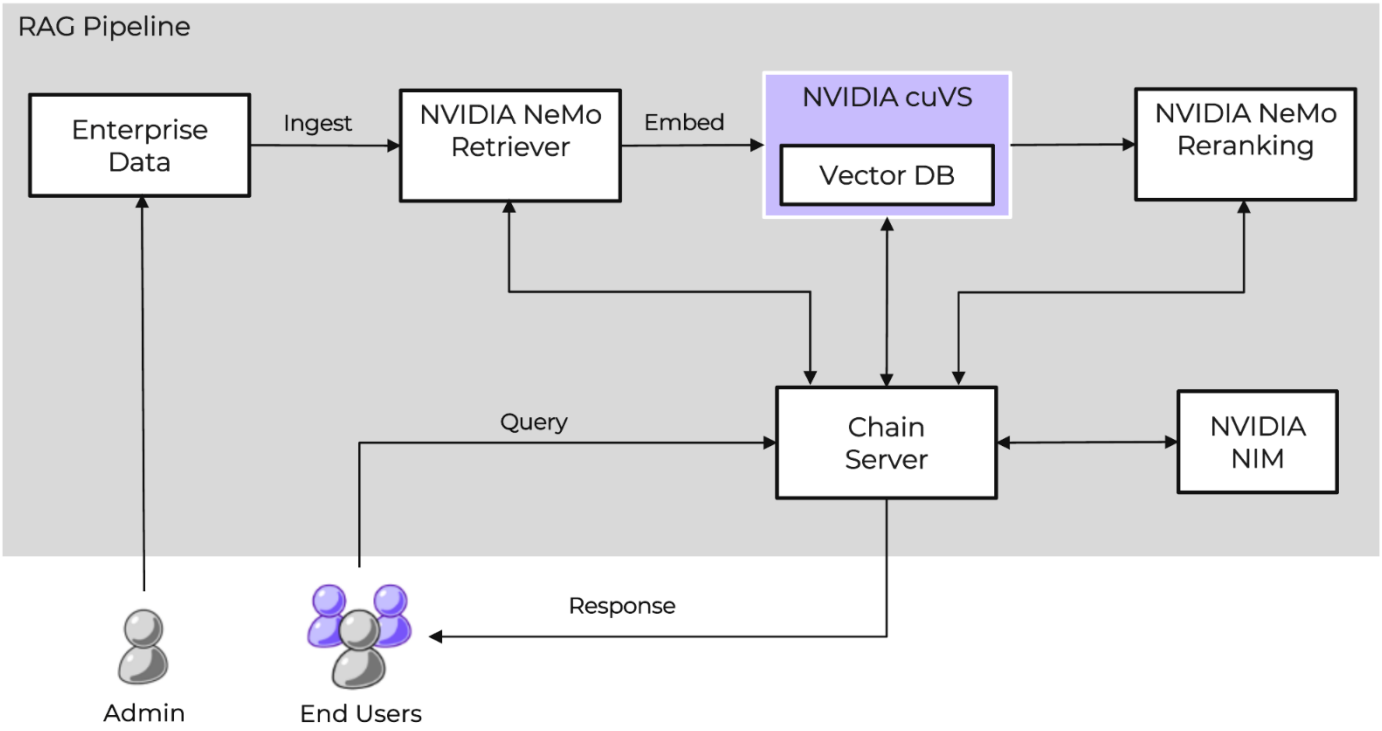
RAG is a game-changer in content creation, assisting enterprises in developing high-quality, contextually relevant materials. It enhances efficiency through various key functionalities:

- **Article Summarization**: RAG transforms lengthy documents into precise, insightful summaries. Delivering key information instantly is ideal for professionals who need quick, actionable insights without the hassle of sifting through extensive content.
- **Content Recommendations**: By analysing retrieved data, RAG suggests related content, allowing creators to seamlessly integrate complementary topics and ideas, enriching both depth and relevance in their work.
- **Adaptive Content Generation**: Leveraging retrieved data, RAG can generate contextually relevant text, including personalized marketing emails, blog posts, and technical documentation. By utilizing an organization's proprietary data, it promotes tailored content that aligns with specific audience needs.

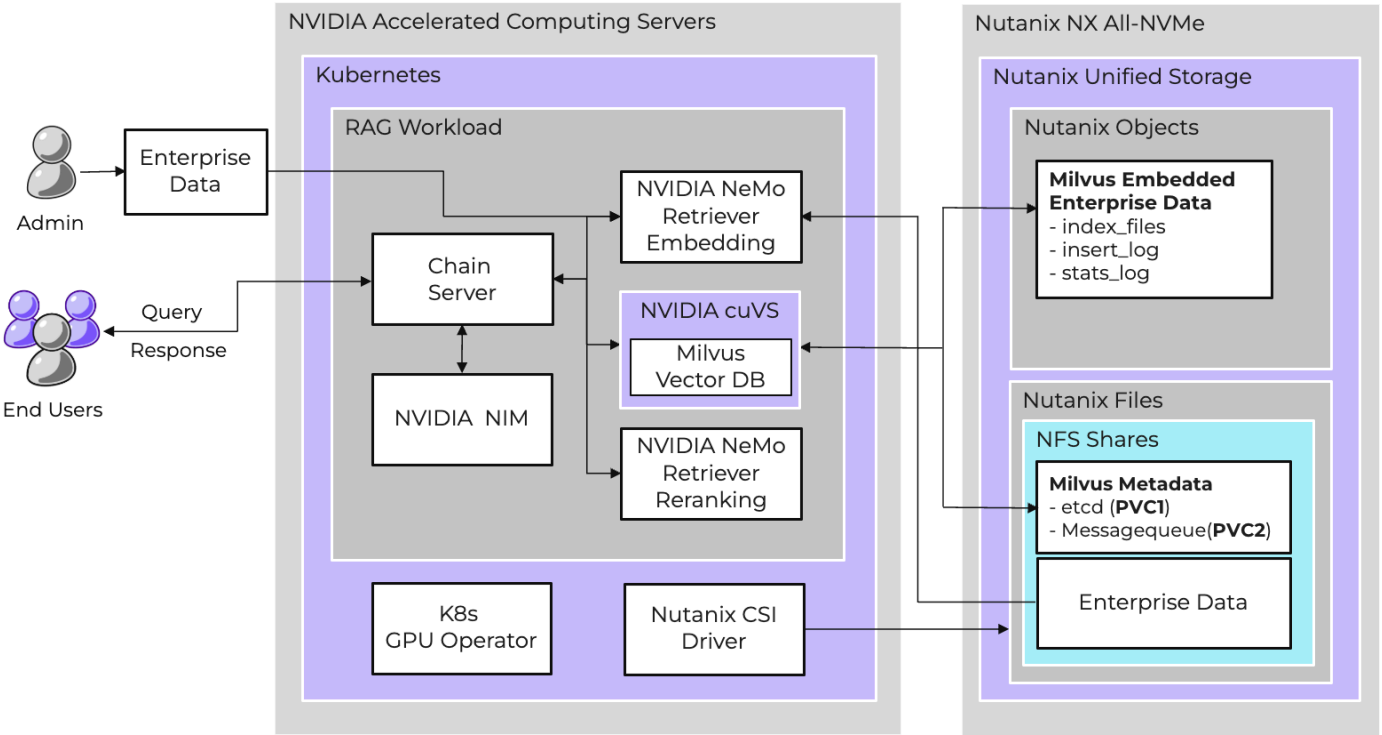
Solution Design

A well-structured RAG pipeline enhances LLM accuracy by minimizing hallucinations and integrating real-time data for reliable responses. It improves transparency by linking AI outputs to verifiable sources. Leveraging the NUS and NVIDIA ecosystem, enterprises can deploy a scalable RAG architecture for seamless data retrieval, indexing, and augmentation, allowing AI to refine accuracy and keep up to date.

RAG Pipeline



NVIDIA and Nutanix Solution Stack



- **Identify Relevant Data Sources**: Gather structured and unstructured data from various internal and external sources, with continuous updates to facilitate up-to-date responses.

- **Enterprise Data Ingestion:** Enterprise data is stored in Nutanix Files and shared via an NFS export where it is accessed through a Jupyter Lab Notebook, which acts as the interface for ingesting the data into a RAG pipeline and further embedding is carried out using NVIDIA Embedding Model which is deployed on an NVIDIA Certified Server to send the embeddings to vector DB.

- **Extract Raw Data**: Retrieve data in various formats, such as text and PDFs, to promote comprehensive knowledge representation.

- **Enterprise Data Ingestion:** Ingest enterprise data, stored in Nutanix Files and shared via an NFS export where it is accessed through a Jupyter Lab Notebook, which acts as the interface for ingesting the data into a RAG pipeline.

- **Preprocess and Chunk**: Transform raw data into smaller, manageable chunks (sentences or paragraphs) to help improve model comprehension and prevent information loss.

- **Generate Embeddings**: Convert processed data into high-dimensional vector embeddings by NVIDIA Embedding Model, deployed on an NVIDIA Certified Server and insert the embeddings to vector DB which stores data on NUS using object store and NFS filesystem, preserving the semantic meaning of the text for efficient retrieval.

- **Data Indexing**: Build vector indexes to accelerate similarity searches, allowing rapid retrieval of relevant content. Define the index by specifying the vector field name and configuring the index parameters, which specify the index type to use.

- **Retrieve Relevant Content**: Compare the query against the vector DB before querying the LLM.
    - Retrieve the most semantically similar documents along with their metadata.
    - Balance retrieval speed and result quality by optimizing embedding and latency considerations.
    - Use reranking techniques to refine results for greater relevance.

- **Response Generation**: The final step in the RAG pipeline combines the user's query with the retrieved context from the vector database, transforming it into meaningful insights. By combining the pre-trained LLM's knowledge with the most relevant retrieved data, RAG generates contextually rich and precise responses.

Solution Components

For testing and designing the integrated solution combining Nutanix Unified Storage and NVIDIA, the following software and hardware stacks were used.

Software Stack:

Component	Functionality
NVIDIA AI Enterprise	NVIDIA AI Enterprise is a software suite that provides optimized AI tools, frameworks, and support for running AI workloads efficiently on NVIDIA GPUs in datacenters, clouds, and edge environments.
NVIDIA AI Blueprint	NVIDIA AI Blueprint for RAG

NVIDIA NIM	NVIDIA NIM (NVIDIA Inference Microservices) provides ready-to-use, optimized AI model endpoints, making it easy to deploy and run AI workloads like LLMs, vision, and speech models with minimal setup. NVIDIA NIM model
NVIDIA NeMo Retriever Embedding	NVIDIA NeMo Retriever Embeddings convert text into numerical vectors, allowing AI models to quickly find and retrieve relevant information from large datasets, improving accuracy in RAG-based applications Embedding Model: NVIDIA Nemo Retriever Embedding Model
NVIDIA NeMo Retriever Reranking	NVIDIA NeMo Retriever Reranking improves search results by re-evaluating and reordering retrieved documents, ensuring the most relevant information is prioritized for AI models. Reranking Model: Nvidia Nemo Retriever Reranking Model
NVIDIA GPU Operator	Part of NVIDIA AI Enterprise, the NVIDIA GPU Operator uses the operator framework within Kubernetes to automate the management of all NVIDIA software components needed to provision GPUs.
Kubernetes	Container orchestration platform.
Nutanix CSI	Integration of Nutanix storage with Kubernetes for dynamic volume provisioning and management.
Nutanix Unified Storage	Scale-out Files and Object Storage.

Hardware Stack:

Component	Specification
Server	**2x - [NVIDIA-Certified Servers]** Below is the detailed specification per node. - CPU: 2x Intel(R) Xeon(R) Gold 6448Y - GPU: 4x NVIDIA L40S - Memory: 16 X 64GB - 2x Ethernet Controller X710 for 10GBASE-T - 2x MT43244 BlueField-3 integrated ConnectX-7 network controller - 2x ConnectX-7 (MT2910) NICs, each with 2 ports
Storage	**4x - [NX-8170-G9 - All NVMe Nodes]** Below is the detailed specification per node. - CPU: 2 x Intel(R) Xeon(R) Gold 6448H - Memory: 8 X 64GB - 2x ConnectX-7 (MT2910) NICs, each with 2 ports
Network Switch	- NVIDIA MSN4600 Ethernet Switch

Conclusion

RAG marks a major evolution in how enterprises utilize LLM by seamlessly integrating AI with proprietary organizational knowledge. This synergy enables more precise and efficient AI applications in semantic search, customer support, and content generation. By dynamically retrieving relevant information from extensive data repositories, RAG empowers businesses to make informed decisions, maximize operational efficiency, and help gain a competitive advantage.

The Nutanix Unified Storage RAG platform with NVIDIA redefines data retrieval and LLM augmentation, delivering maximum performance, scalability, and usability. By leveraging the robust infrastructure of Nutanix alongside NVIDIA's AI expertise, this platform enables organizations to streamline their AI-driven processes and maximize the value of their data assets.

Nutanix and NVIDIA are collaborating on reference architectures tailored to diverse generative AI and RAG use cases. These solutions provide enterprises with a seamless way to harness their data efficiently with scalability, high performance, and operational efficiency. This white paper marks the beginning of a series of in-depth technical discussions on scalable, production-ready RAG architectures powered by Nutanix Unified Storage.

Additional Resources

- Learn how you can accelerate adoption of AI with the Nutanix Unified Storage Platform.
- Explore AI Ready Infrastructure that simplifies enterprise AI.
- Discover Nutanix Unified Storage for all your unstructured data storage needs.