TECHNICAL WHITE PAPER

Pure Storage GenAI RAG Platform for Financial Services with NVIDIA NeMo Microservices and KX KDB.AI

Contents

Executive Summary

The emergence of generative AI (GenAI) has transformed business operations, and the financial services sector is no exception. This white paper delves into the potential of a GenAI retrieval-augmented generation (RAG) platform, a state-of-the-art solution that integrates NVIDIA NeMo™ microservices and the KDB.AI vector database to enhance the application of GenAI in the financial services realm. The GenAI RAG platform is a holistic solution that harnesses the latest in natural language processing (NLP), machine learning (ML), and GenAI technologies, supported by reliable NVIDIA networking, GPU compute, and the Pure Storage® platform for AI data storage. It aims to optimize various processes within the financial services industry, including customer service, risk management, fraud detection, and investment analysis using AI technologies. This paper offers deployment guidance, a detailed architecture, and benchmark results, illustrating how this platform can be successfully implemented to meet the rigorous requirements of the financial industry.

Introduction

This technical white paper is focused on providing a validated solution customized for the financial sector with details on deployment guidance and benchmarked results to implement a Pure Storage RAG platform in partnership with KX and NVIDIA. Generative AI models, including those using RAG and pre-trained large language models (LLMs), can be employed to create financial analyses that are more accurate, real-time, and cost-efficient compared to those previous labor-intensive efforts or those using LLMs alone. GenAI models have the capability to process and analyze vast amounts of information at a much higher rate than manual analysis, enabling them to generate comprehensive and up-to-date financial analyses. RAG is a method that enhances the accuracy and relevance of LLM outputs by augmenting them with a vector search lookup of cleaned and vectorized data that is custom or proprietary to organizations. This approach improves the inference capability of LLMs by providing them with relevant and contextual information from the vectorized data.

Despite the popularity of powerful LLMs like GPT-4, Command R+, and Claude, their usage in financial services poses a risk of inaccurate responses to queries involving internal enterprise data. Because these models are not updated frequently, nor trained on the proprietary data of financial institutions, queries from employees may yield unreliable answers.

Consider the scenario of creating financial projections using the last 10 years of 10-K data from a global financial firm with a general-purpose LLM like ChatGPT. Due to the lack of access to internal data from the financial institution such as resource planning and profitability metrics, the model would not be able to provide a clear and accurate response. Even if the model did have access to the data, the responses would likely contain outdated information, as the model's training dataset is typically several months old.

To address this challenge, we can leverage pre-trained domain-specific LLMs, such as FinGPT, and augment them with enterprise data that has been cleaned, modularized, and vectorized into a vector database on a Pure Storage FlashBlade® system. FlashBlade provides an optimal storage platform for the high performance needs of a RAG application. Updates to data can be made frequently so that data is fresh and accurate.

By integrating internal 10-K data into a domain-specific LLM like FinGPT and then augmenting it with RAG, we can generate accurate and up-to-date summaries of the company's financial information. This approach leverages the model's understanding of financial concepts and terminology while providing access to the most recent and relevant data from a financial institution's internal sources.

Solution Overview

RAG for financial services can be used by financial analysts and professionals working at hedge funds, banks, and quant firms. These users rely on RAG for key tasks like providing tailored financial advice that reduces the need for or enhances in-person consultations by summarizing lengthy financial documents such as reports, research papers, and regulatory filings into concise and actionable summaries. By designing ideal investment portfolios and analyzing various economic variables and investor profiles, and even creating trading signals for informed quantitative trading decisions, professionals using these services gain extensive knowledge. The benefits include significant time savings, improved accuracy, enhanced decision-making efficiency, and the ability to quickly access and act on critical financial information.

This groundbreaking RAG platform offers the performance, scalability, and ease of use that financial services teams require. This innovative solution leverages the power of vector databases and advanced indexing to revolutionize the way organizations access and utilize their valuable data. See the Solution Design section for more details of the architecture.

Benefits of Running Vector Databases like KDB.AI on FlashBlade

Vector databases differ from traditional databases in that they are designed specifically to handle vector embeddings as numerical arrays that represent the characteristics of an object. This specialized model allows for efficient vector data storage and retrieval, making the database more flexible. They are also scalable and built to handle specialized data models that allow for efficient storage and retrieval of this type of data. When performing high-speed searches, performance is critical in vector databases, as these searches leverage advanced indexing techniques like locality-sensitive hashing (LSH) and approximate nearest neighbor (ANN) algorithms to enable fast similarity searches.

Running a KDB.AI vector database on a FlashBlade system gives you:

- Native multi-protocol access for NFS, S3, and SMB and support for billions of files and objects in a single system and namespace.
- An AI storage platform that enables customers to reduce their storage-related energy, space, and administrative requirements by up to 85%, compared to competitive solid state disk (SSD) solutions and up to 95% compared to hard disk drive (HDD) solutions.[1] For many users, the data center real estate, energy requirements, and TCO of power-hungry AI workloads are a major concern when accommodating these workloads in existing or planned data center footprints and budgets.
- The ability to scale the VectorDB to tens of petabytes while independently scaling to support multiple compute GPUs. By scaling performance and capacity independently, you save money by only growing what you need when you need it.
- User experience and operational simplicity. Storage infrastructure that can upgrade non-disruptively with no performance reduction greatly enhances the user experience and productivity of our customers for RAG use cases.

Our architecture is designed to include the KDB.AI vector database—a critical component of the RAG platform for financial services—which is hosted on a FlashBlade//S500. FlashBlade//S500 provides scalability advantages by providing an easy, non-disruptive way to grow the vector database from hundreds of terabytes to multiple petabytes—something that is very challenging and resource-intensive using local SSDs or alternative all-flash products. This combination of Pure Storage and NVIDIA for RAG use cases delivers a proven and validated, production AI solution with performance, cost savings, and environmental, sustainability and governance (ESG) benefits to enterprises.

The KDB.AI Advantage

Leveraging the robust and proven KDB+ engine, independently recognized as the fastest time series vector native database, KDB.AI is revolutionizing AI technology. Providing the best price to performance and efficiency ratio, KDB.AI overcomes traditional vector database challenges such as statelessness, lack of contextual understanding, ultra-high-dimensionality problems, an absence of causal reasoning, and hurdles around explainability, fairness, and sustainability. Moreover, KDB.AI fills a significant gap in the real-time processing and streaming of vectors, a limitation profoundly felt in the AI field today. Seamlessly handling both structured and unstructured data, KDB.AI empowers real-time processing, enabling sliding window search for time-series-based vector search. Its stateful approach caters to various applications, providing native support for large language models (LLMs) and machine learning workflows.

1. **Real-time processing**: KDB.AI enables real-time processing of vast time-series and alternate datasets. Financial institutions can swiftly analyze market data, monitor trading activities, and respond quickly to market fluctuations.
2. **Actionable insights**: By leveraging KDB.AI's robust capabilities, organizations gain actionable insights from large, complex multi-source data. These insights empower decision-makers to make informed choices, optimize investment strategies, and manage risk effectively.
3. **Generative AI applications**: KDB.AI facilitates the development of scalable, enterprise-grade generative AI applications. These applications enhance user engagement, minimize uncertainty, and speed time to market. Imagine AI-driven tools that assist traders, predict market trends, and automate routine tasks.
4. **Enhanced user loyalty**: With KDB.AI, financial services providers can create personalized experiences for clients. Whether it's personalized investment recommendations or tailored financial advice, generative AI powered by KDB.AI enhances user loyalty and satisfaction.

Financial Services RAG Platform Solution Use Cases

RAG introduces a transformative approach for integrating LLMs with the specific knowledge and data of an organization. This synergy between the deep learning capabilities of LLMs and an organization's unique data sets enables RAG to address the challenges associated with effectively deploying LLMs for real-world, enterprise-specific applications. Below are detailed explanations of how RAG benefits enterprises across various use cases.

Financial Analysis Using Semantic Search and Efficient Information Retrieval

This solution aims to empower financial analysis through a question-answering interface that can take the context of a company's internal data and provide specific answers to reduce hallucinations and improve the accuracy of the results. A financial analyst could, for example, search through millions or billions of internal documents and query them in a matter of seconds. This is very different and differentiated from what GPT-4 and other open-source LLMs like LLaMa-3 can do.

Additionally, LLMs can be used to summarize lengthy financial documents while augmenting them with relevant internal data—not found in open-source LLMs—stored on a Pure Storage FlashBlade system.

RAG significantly enhances the capabilities of semantic search, making it a powerful tool for enterprises to navigate and extract pertinent information from their extensive data repositories. This is particularly beneficial in scenarios where precision and context are paramount, such as the highly regulated financial industry. By understanding the intent behind queries and matching them with contextually relevant content, RAG streamlines the process of information retrieval, making it more efficient and accurate.

Quantitative Trading

In quantitative trading, understanding context is crucial. RAG allows models to ingest relevant information from external sources, improving their contextual awareness. When analyzing a stock's performance, RAG can retrieve historical data, proprietary company reports, and news articles related to that stock. This context enhances the model's ability to generate relevant and accurate insights that will have a measurable effect on downstream applications. Quantitative traders need up-to-date information on market events, economic indicators, and company-specific news. Unlike open-source GenAI LLMs, RAG can retrieve real-time data from financial news sources, regulatory filings, and market reports, which it then uses to generate responses that reflect the latest developments. Imagine an AI trading assistant that provides instant updates on investment strategies based on earnings announcements, central bank decisions, or geopolitical events—RAG makes this competitive edge possible.Traditional algorithms can gather data such as Federal Reserve decisions, but incorporating the impact of that data across hundreds or thousands of investment strategies swiftly requires RAG (Retrieval-Augmented Generation).

Backtesting

Backtesting is essential for evaluating trading strategies, and RAG can significantly enhance this process by providing detailed risk assessments. By retrieving and analyzing historical volatility, drawdowns, and various risk metrics, RAG helps traders understand the risk exposure of their strategies. This leads to better-informed decisions, increasing the chances of success in live trading environments.

Enhancing Financial Customer Service

In the realm of customer service, RAG offers substantial improvements in both efficiency and the quality of interactions. Traditional customer service bots often provide responses based on a limited set of pre-defined answers, which may not always address the customer's specific needs accurately. Additionally, providing inaccurate information may expose a financial institution to regulatory penalties. However, with RAG a chatbot can access and understand the bank's unique policies, customer details, and proprietary information, enabling it to deliver precise and personalized responses to customer inquiries. RAG can provide support professionals with detailed customer information to streamline and enhance human-in-the-loop support. This capability is not limited to banking; it extends to any customer service scenario where personalized, accurate information delivery is crucial.

Improving Insights and Due Diligence

With a RAG system, analysts can identify, perform due diligence on, and monitor investment opportunities in public markets. LLMs using RAG can pull real-time financial data, news snippets, or even sanctions lists from constantly updated databases. Once the comprehensive analysis is generated, businesses provide analysts and advisors with current insights for informed investment decisions. The RAG platform enhances LLM outputs by incorporating a vector search of cleaned and vectorized data, providing more accurate and contextual information. This approach reduces the risks associated with the use of generalized LLMs in financial services, such as inaccurate responses to queries involving internal enterprise data.

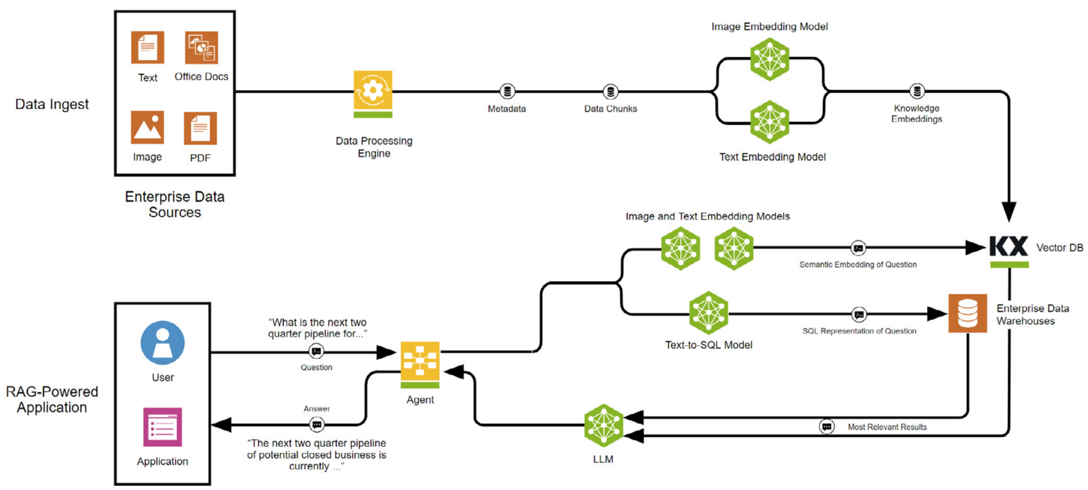
Retrieval-augmented Generation Pipelines

A RAG-powered application pipeline is an advanced approach in AI that enhances the relevance and accuracy of responses or predictions by combining retrieval mechanisms with generative models. For a financial analyst, understanding this pipeline involves recognizing its stages and the user benefits it provides. The stages are:

- **Data ingestion and preprocessing:** Data is collected from various sources such as financial reports, market data, news articles, and historical records. This data is cleaned to remove duplicates, handle missing values, and ensure consistency and quality. It is then indexed to facilitate efficient retrieval.
- **Retrieval:** This stage begins with query understanding, where the user's query is interpreted to determine intent and extract key terms. The system searches the indexed data to retrieve the most relevant documents or data points, using techniques like TF-IDF, BM25, or neural retrieval models. These documents are then ranked based on relevance using factors such as recency, source reliability, and content relevance.
- **Augmentation:** The retrieved documents are then combined with the original query to form a comprehensive input for the generative model, ensuring all necessary context is included. This stage may also involve data enrichment, where additional data, such as market trends or expert opinions, is incorporated to enhance the generative process.
- **Generation**: This stage utilizes a generative model, like GPT-4, to produce a coherent and contextually relevant response based on the integrated input from the retrieval stage. The generated response is fine-tuned to align with financial analysis standards and user expectations.
- **Post-processing and delivery:** The generated response is validated for accuracy and relevance. This may involve automated checks or human-in-the-loop validation for critical applications. The response is then formatted in a user-friendly manner, such as charts, reports, or summaries, and delivered through the chosen interface, whether it's a dashboard, report, or alert system.

The user benefits are substantial. RAG-powered pipelines enhance decision-making by providing accurate and contextually relevant responses, aiding in informed financial decisions. They offer comprehensive insights by integrating diverse data sources and perspectives, giving a holistic view of the financial landscape. Time efficiency is significantly improved as the automated processes reduce the time required to gather and analyze information compared to manual methods. Users receive real-time updates and insights, enabling swift responses to market changes.

Personalization is another key benefit, as the system can be fine-tuned to cater to individual user preferences and deliver context-aware insights. The pipeline’s scalability allows it to handle large volumes of data and queries simultaneously, making it suitable for both individual analysts and large financial institutions. Moreover, it can adapt to new data sources and evolving market conditions, ensuring continuous relevance and utility. In summary, a RAG-powered application pipeline enhances the efficiency, accuracy, and relevance of financial analysis, providing users with timely, comprehensive, and actionable insights. This advanced approach leverages the strengths of both retrieval and generative AI to support sophisticated financial decision-making processes.



Solution Benefits

This architecture is designed to enhance data processing capabilities, particularly for applications that need to rapidly access and retrieve large datasets. FlashBlade//S™ excels in providing high throughput and low latency, optimizing parallel processing and efficient data access. Key benefits include industry-leading performance density, energy efficiency, and cost-effectiveness, ensuring that financial institutions can maximize their return on investment. The architecture's scalability and flexibility allow for independent scaling of vector database capacity and accelerated compute resources, which is crucial when adapting to fluctuating data volumes and computational demands. As data volumes and computational needs grow, the system can seamlessly expand without performance degradation, ensuring continuous, efficient operation.

The enhanced user experience is another critical aspect. Non-disruptive upgrades and consistent performance mean that system maintenance does not interfere with ongoing operations, allowing financial analysts to work without interruptions. The simplified deployment and management processes reduce the complexity typically associated with advanced data infrastructure, enabling financial institutions to quickly integrate and leverage the architecture for better insights and decision-making. Overall, this architecture empowers financial analysts with faster data ingestion, superior performance, and a more streamlined operational experience, crucial for maintaining a competitive edge in the fast-paced financial industry. Additional benefits of a RAG platform with Pure Storage include:

- **Seamless data integration**: The RAG platform seamlessly integrates a wide range of media, including brochures, internal web pages, emails, videos, photos, knowledge base articles, and more. These assets are embedded into a vector database, and an index is built to establish a direct link between the documents and the user's queries.
- **Intelligent query matching**: The user's queries are also vectorized using the same embedding model, allowing for precise matching against the vector database. This process ensures that the most relevant documents are retrieved, providing users with the information they need to enhance their queries and the overall context.
- **Augmented query processing and robust guardrail filtering**: The retrieved assets are used to augment the user's query and provide additional context to the LLM. This integration of relevant information helps to improve the accuracy and relevance of the LLM's responses, ensuring that users receive the most accurate and valuable information. The retrieved assets can also be utilized to fine-tune the guardrails that filter out malicious inbound queries and/or inappropriate outbound responses. This feature helps to maintain the integrity and safety of the system, ensuring that users receive only the most appropriate and trustworthy information.
- **Unparalleled performance, scalability, and ease of use**: The KX, and Pure Storage with NVIDIA RAG platform delivers unparalleled performance and scalability, allowing organizations to efficiently manage and access their vast data repositories. This solution is designed to handle even the most demanding data requirements, ensuring that users can access the information they need quickly and seamlessly.

The RAG platform is designed with user-friendliness in mind, making it easy to integrate into existing workflows and systems. This streamlined approach allows organizations to leverage the power of this innovative solution without the need for extensive infrastructure development.

Solution Design

The importance of a RAG pipeline lies in its ability to provide LLMs with relevant and factual information, reducing the likelihood of hallucinations, which are convincing, but factually inaccurate responses when LLMs are not supplied with the proper contextual data. RAG also allows for dynamic data integration, ensuring the system's responses are based on the latest available information. Additionally, RAG can enhance the transparency and trust of AI systems by enabling users to trace back how the AI came to a response, as it fetches and presents data from specific, verifiable sources. A deeper look at the solution architecture design from Pure Storage with PX and NVIDIA will give a deeper understanding of the solution.

Supporting Components

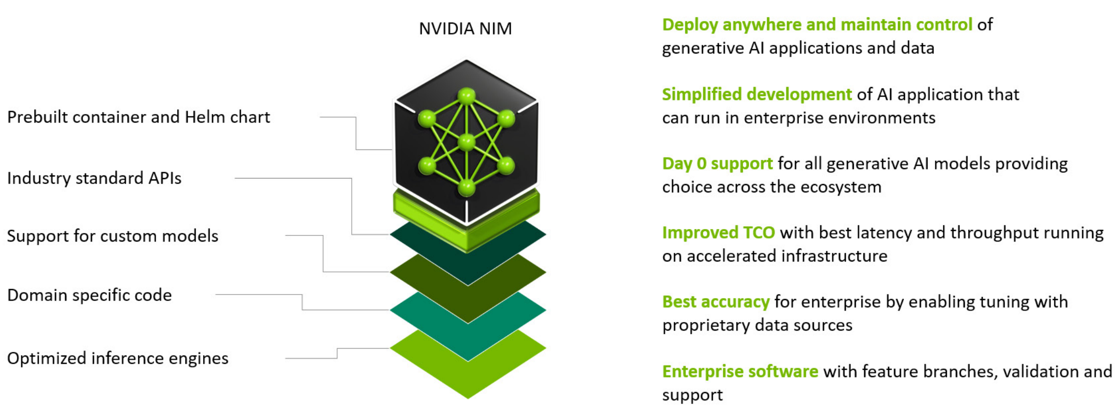
NVIDIA NeMo microservices offer analysts advanced AI models and GPU-accelerated inference capabilities, enhancing the efficiency and accuracy of financial analytics. By providing pre-trained and customizable models for tasks such as sentiment analysis, fraud detection, and risk management, NVIDIA NeMo enables rapid deployment and fine-tuning specific to financial data. Its microservices architecture ensures scalability, allowing for seamless handling of increasing workloads. Leveraging NVIDIA GPUs, NeMo accelerates data processing and model inference, facilitating real-time or near-real-time analysis essential for applications like algorithmic trading and customer service automation, ultimately empowering analysts to make faster and more informed decisions.

The KDB.AI vector database is designed to efficiently store and retrieve high-dimensional vector embeddings, which are crucial for tasks such as machine learning, recommendation systems, and natural language processing. Utilizing advanced indexing techniques like approximate nearest neighbor (ANN) search, the database ensures rapid access to relevant data points. It optimizes storage with space-efficient data structures, allowing for the handling of large-scale, high-dimensional datasets. The database also supports seamless integration with AI models, enabling real-time updates and queries, thus providing financial analysts with swift and accurate insights from complex, multidimensional data.

Pure Storage FlashBlade offers financial firms scalable and high-performance storage solutions by combining a massively parallel architecture with all-flash technology. This enables rapid data access and processing, essential for handling large volumes of financial data and performing complex analytics. The scalability of FlashBlade allows it to grow with an organization's needs, providing seamless capacity expansion without compromising performance. Additionally, its integrated data protection and management features ensure data reliability and accessibility. This results in faster decision-making and improved operational efficiency, making it an ideal choice for demanding financial applications and workloads.

NVIDIA NeMo LLM Inference Microservice (NIM)

The NVIDIA NIM plays a crucial role in bridging the gap between the complexities of AI development and the practical operational needs of enterprise environments. It is a scalable, efficient, and easy-to-deploy solution for leveraging large language models (LLMs) in enterprise applications. It provides a way to deploy, manage, and scale AI models in a production environment, ensuring that businesses can derive real value from advanced AI technologies.



NIM offers financial institutions several advantages:

Scalability and performance: NIM is designed to handle large-scale deployments, meaning it can manage multiple instances of LLMs across various nodes, which is crucial for enterprises in the finance sector, where processing large volumes of data and providing real-time responses are fundamental requirements. By leveraging NVIDIA accelerated infrastructure, NIM ensures that the inference process is highly optimized. This optimization results in faster response times and more efficient processing of extensive datasets, which is vital for financial analysts who rely on up-to-the-minute data for decision-making.

Ease of deployment: NIM utilizes container technology such as Docker, making it easy to deploy across different environments. This containerized approach ensures that LLMs are portable and can be deployed consistently in development, testing, and production environments. For financial institutions, this means reduced time to market and greater flexibility in model deployment. Furthermore, NIM's integration with Kubernetes allows for automated deployment, scaling, and management of containerized applications. This orchestration reduces operational overhead and ensures high availability and reliability, crucial for maintaining continuous financial operations.

Operational efficiency: NIM provides data for monitoring and managing the performance of deployed models using standard infrastructure tooling. Financial analysts benefit from the ability to track metrics like latency, throughput, and resource utilization, which are critical for maintaining optimal performance of financial models. Additionally, NIM can scale using Kubernetes, allowing the auto-scaling feature allows the system to automatically scale the number of model instances up or down based on demand. This ensures that resources are used efficiently and cost-effectively, enabling financial firms to handle varying workloads without over-provisioning.

Integration with existing systems: NIM provides RESTful APIs, making it straightforward to integrate with existing enterprise systems. Financial analysts and other users can easily access the capabilities of LLMs through familiar interfaces without needing deep technical knowledge of AI. This API-based access facilitates the integration of advanced AI capabilities into existing financial tools and platforms. Moreover, NIM supports various AI frameworks and can seamlessly integrate with existing data pipelines. This compatibility ensures that financial enterprises can leverage their current infrastructure and workflows without extensive modifications, promoting a smooth transition to AI-enhanced operations.

Security: NIM ensures that data is processed securely, adhering to enterprise security standards and regulatory requirements. This is crucial for financial institutions that handle sensitive data and must comply with stringent security protocols. By providing robust security guardrails, NIM helps financial institutions mitigate risks and maintain trust with their clients and stakeholders.

NVIDIA NeMo™ Retriever

NVIDIA NeMo Retriever microservices are designed to streamline and enhance the processes of document embedding, retrieval, and querying, which are critical components of a retrieval-augmented generation (RAG) pipeline.

Document embedding—the process of converting documents into numerical representations (vectors)—is simplified with NeMo Retriever pre-trained models. These models generate high-quality embeddings without requiring extensive customization, allowing you to start embedding financial documents immediately. Additionally, the scalability of NeMo Retriever ensures that even vast repositories of financial data are processed efficiently.

For document retrieval, NeMo Retriever employs advanced search algorithms that significantly speed up the process of locating relevant documents. Whether you need quarterly reports or specific financial statements, NeMo Retriever can quickly identify and retrieve these documents. The microservices also leverage sophisticated models to understand the context of your queries, providing more accurate and relevant results.

Querying your document repository is made more intuitive with NeMo support for natural language processing (NLP). This feature allows you to use natural language to query your documents, simplifying the process and eliminating the need for complex query languages. Additionally, NeMo Retriever supports interactive querying, enabling you to refine your searches based on initial results to get the precise information you need.

For a financial analyst, these capabilities translate into significant time savings, as the embedding, retrieval, and querying processes are automated and accelerated. Enhanced search algorithms and NLP ensure more relevant and accurate results, crucial for financial analysis and decision-making. NeMo scalability makes it suitable for both small and large financial institutions, handling anything from hundreds to millions of documents. The user-friendly interfaces and natural language support mean users don't need to be a technical expert to leverage these powerful tools.

For instance, if an analyst is working on a market analysis report and needs to gather relevant financial statements from the past five years, NeMo Retriever can streamline this task. By embedding your financial documents into numerical vectors once, the user can then quickly retrieve specific documents using natural language queries such as "Q4 financial statements for 2020." Searches can be refined interactively to ensure the user is getting the precise information they need.

In essence, NeMo Retriever microservices make it easier and faster to turn vast amounts of financial data into actionable insights, allowing users to focus more on analysis and less on data management.

The NVIDIA API catalog enables developers to protype their AI applications using industry standard APIs. The catalog provides APIs for every stage of a RAG pipeline, allowing developers to create RAG applications and deploy them on-premises using NIM containers.

Enterprise Stack

Table 1 outlines the components and functionality of the solution.

Component	Functionality
NVIDIA AI Enterprise	End-to-end AI platform accelerates data science pipelines and streamlines development and deployment of production-grade co-pilots and other generative AI applications.
NVIDIA NIM	Designed to bridge the gap between the complex world of AI development and the operational needs of enterprise environments, enabling 10-100X more enterprise application developers to contribute to AI transformations of their companies.
NVIDIA NeMo Retriever	A collection of generative AI microservices that enable organizations to seamlessly connect custom models to diverse business data and deliver highly accurate responses.
NVIDIA GPU Operator	Lifecycle management of software required to use GPUs with Kubernetes.
KDB.AI Vector Database (via NFS)	KDB.AI integrates RAG and mixed search—literal, semantic, and time series—enabling nuanced querying that considers context and relationships for more accurate and insightful analysis.
Docker	Container platform
NVIDIA DGX™ systems or NVIDIA OVX™ systems	Accelerated compute
FlashBlade//S500, 1-10 chassis	Storage
NVIDIA Spectrum™ Series Switches SN3700	Network

TABLE 1 Enterprise stack components

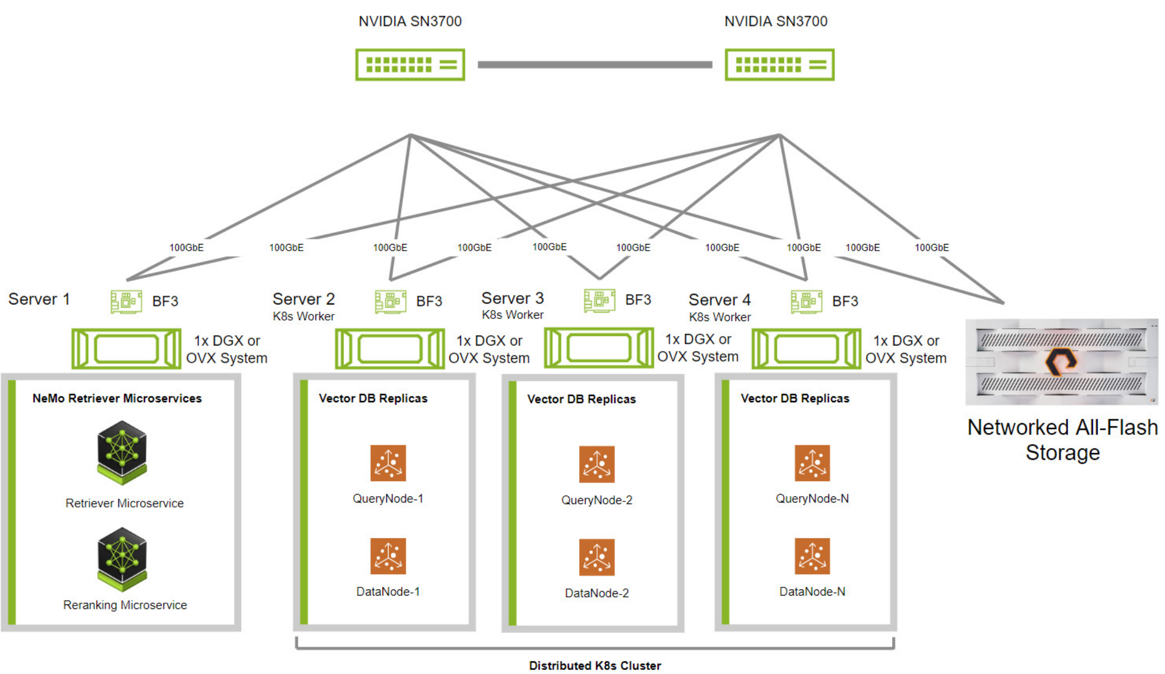
Recommended Hardware Architecture and Design

We recommend the following for architecting the solution:

- **Accelerated computing resources**: NVIDIA DGX systems and NVIDIA OVX systems deliver the necessary computational power for AI workloads, supporting analytics, training, and inference operations.
- **FlashBlade storage**: FlashBlade//S500 provides a unified storage platform for NFS, S3, and SMB protocols, supporting large-scale vector databases and facilitating high-speed data retrieval. On a FlashBlade system, objects can be natively accessed via the S3 storage protocol, and simultaneously files can be accessed natively via the NFS or SMB storage protocols.

Sizing and tuning: Proper sizing and tuning of accelerated computing and storage resources are critical for optimizing performance. Detailed guidelines on configurations ensure that the platform meets specific workload requirements. In the testing and design of the Pure Storage and NVIDIA integrated solution we used 2-nodes of NVIDIA DGX or NVIDIA OVX systems with 1-10 chassis FlashBlade//S500 systems, and an NVIDIA Spectrum-2 SN3700 switch.

- **Compute—2 NVIDIA DGX or NVIDIA OVX systems**: NVIDIA DGX systems and OVX systems are powerful and versatile systems purpose-built for all AI infrastructure and workloads, ranging from analytics to training and inference.
- **Storage—1 to 10 chassis FlashBlade//S500**: FlashBlade//S 500 excels at energy efficiency, scalability, and multi-modal performance and is designed to handle unstructured data efficiently. The modular architecture of FlashBlade//S allows you to independently scale capacity and performance. This flexibility ensures greater efficiency and minimizes waste. You can adjust it to meet your growing needs and projections.
- **Network switch—NVIDIA Spectrum-2 SN3700 Ethernet switch**: The SN3700 enables connectivity to endpoints at different speeds and carries a throughput of 6.4Tb/s, with a landmark 8.33Bpps processing capacity. As an ideal spine solution, the SN3700 allows maximum flexibility, with port speeds spanning from 10GbE to 200GbE per port.



Software Stack with NIMs

We recommend the following software stack:

- KX Vector DB Server Edition v1.1.0 running on docker. See https://code.kx.com/kdbai/gettingStarted/kdb-ai-server-setup.html for installation on NFS.
- Kdbai-client v1.1.0 for connecting to the KX Vector DB SDK in Python.
- LangChain v0.0.330 for document pre-processing and chunking.
- NVIDIA NeMo embedding-microservice to use the ***gte-large*** embedding model.
- NVIDIA NeMo inference-microservice for LLM and chat generation.
- Pytorch 2.3.0 for deployment of LLMs and GPU synchronization.
- Python 3.11.8 for running the scripts.
- NVIDIA Base Command Manager for cluster management.
- Ubuntu 22.10 for the Linux environment (included DGX OS for DGX systems)

Deployment Requirements

Below are the architecture's deployment requirements:

- **Compute**: NVIDIA DGX systems or NVIDIA OVX systems
- **Storage**: FlashBlade//S500, scalable from one to ten chassis
- **Network**: NVIDIA Spectrum SN3700 switches for high-throughput connectivity

Deployment Steps

For a successful deployment, the following steps need to be carried out:

1. **Infrastructure setup**: Deploying and configuring systems, storage, and network components
2. **Kubernetes deployment**: Setting up a Kubernetes cluster with NVIDIA GPU Operator
3. **Software installation**: Installing KX Vector DB, NeMo Microservices, and other necessary software components
4. **Configuration and tuning**: Fine-tuning settings to optimize performance for specific workloads

Benchmarking Results

The pipeline in this solution is compute-bound rather than storage-bound. To reduce the time needed to get the model production ready, one can simply add more nodes to the system. In this case, NVIDIA DGX systems with NVIDIA A100 Tensor Core GPUs were used.

For the benchmarking results we used the following:

- **Dataset:** All SEC Financial Statement and Notes Data from 2009 to 2024 was used. These datasets were processed in Python to be in readable formats for embedding and ingesting of the vectors. The datasets contained the extensive list of all 10-K filings provided to the SEC 2009-2024, creating a total of approximately 276M tokens
- **Pre-processing:** This consisted of chunking of all of the text files using the LangChain library and the ***gte-large*** embedding model from HuggingFace.
- **Embedding**: The embedding model outputs a 1024-dimensional vector of the chunked text. The embeddings were processed on 8 NVIDIA A100GPUs with 40GB of GPU memory each, in a DGX A100 system. The resulting embeddings were subsequently stored as parquet files with approximately 59M vectors and associated text. Doubling the number of GPU nodes reduced computation time by roughly half.
- **Uploading and inserting:** We used the KDB.AI vector database (KDB.AI) hosted on a docker container using all 128 available cores on a AMD EPYC 7742 64-Core Processor. The files were uploaded and ingested, but with the limitation that they could only be uploaded in 10MB batches. This noticeably increased the total time for ingestion, at a read throughput of 3GBps and a write latency of 5ms during data ingestion. We also ingested multiple files in parallel but saw no significant benefits to the total time versus the serial upload. Newer releases of the KDB.AI vector database will increase the upload size and allow for faster parallel ingestion.
- **Retrieval:** We processed 100 queries of varying sizes and took the average amount of time to retrieve the vectors over the 100 queries. This resulted in a low latency of 0.020 seconds per query with a pre-loaded index.
- **Generation:** We ran 100 queries of varying length and token output size and calculated the total number of tokens per second over all runs. Generation was consistent at around 74 tokens per second across all runs.

Overall, we have proven the full stack infrastructure required for the end user to get near real time insights on financial data with a scalable architecture across storage and compute which optimizes for efficiency and reliability for a RAG infrastructure. These insights can be used across several different use cases across financial services from sentiment analysis, risk management, quant trading and backtesting. The Pure Storage platform enables firms to run complex queries across data sets of very large scale with maximum throughput and minimum latency. Pure's ability to provide seamless scalability is crucial for the demanding requirements required by financial services institution applications.

Strategic Technology Analysis Center (STAC) recently performed the first baseline and scaling STAC-M3 Benchmarks on a stack involving Pure Storage. The stack was KX's kdb+ 4.0 DBMS using NFS version 3 to access 266TiB of total usable storage on a Pure Storage FlashBlade//S500 with Purity//FB 4.1.5 and 10 x Pure Storage FlashBlades.

Conclusion

The Pure Storage GenAI RAG Platform for financial services, with NVIDIA NIM, NeMo, KDB.AI, and NVIDIA accelerated infrastructure and networking offers a transformative solution for the financial services industry. By integrating enterprise-specific data with advanced AI models, this platform enhances the accuracy, relevance, and efficiency of financial analyses. This architecture provides a scalable, high-performance platform for deploying RAG solutions, delivering significant benefits in performance, scalability, and user experience.

RAG represents a significant advancement in the application of LLMs within enterprises. By bridging the gap between the generalized capabilities of LLMs and the specific, proprietary knowledge of an organization, RAG enables more effective and efficient use of AI in semantic search, customer service, and content creation. Its ability to understand and retrieve contextually relevant information from extensive data repositories makes it an invaluable tool for enterprises looking to leverage AI for competitive advantage.

The RAG solution from Pure Storage, NVIDIA, and KX represents a significant advancement in data retrieval and augmentation, offering organizations a powerful and versatile solution that delivers unparalleled performance, scalability, and ease of use. By combining the capabilities of these industry leaders, this platform is poised to transform the way organizations access and utilize their valuable data.

Additional Resources

- Learn more about Pure Storage solutions for financial services.
- Learn more about Pure Storage AI solutions
- Discover FlashBlade//S for all of your unstructured data storage needs.
- Explore Github for the FB py-client.

1 https://www.purestorage.com/docs.html?item=/type/pdf/subtype/doc/path/content/dam/pdf/en/misc/esg/2024-esg-pure-report-technology.pdf

PS2647-01-en 07/24