INFINIDAT

SOLUTION BRIEF

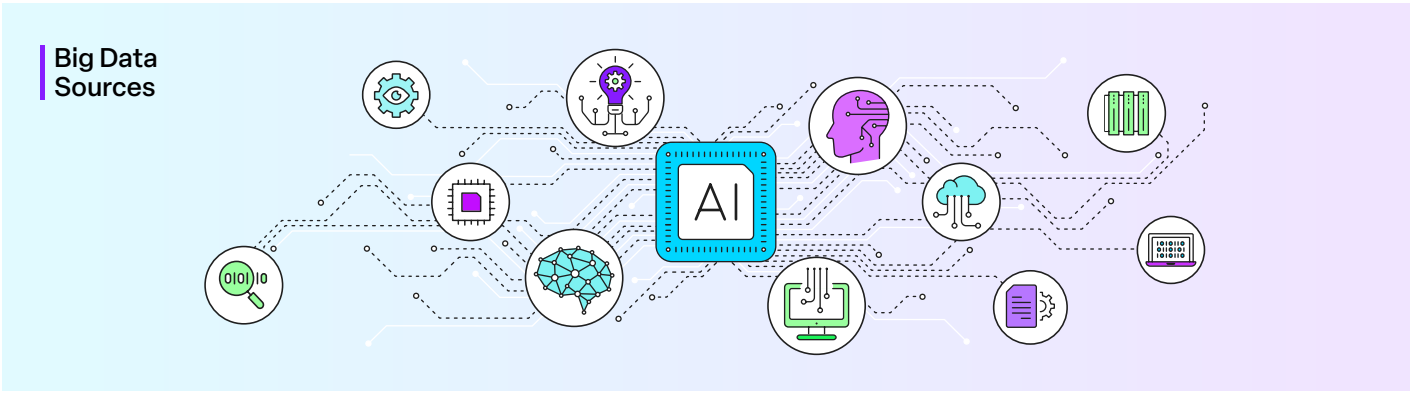
AI Workloads Leverage InfiniBox®

Retrieval-Augmented Generation (RAG) is the New "Killer App" Fueling Machine Learning (ML) and Artificial Intelligence (AI)

Enterprises of all sizes have been accumulating vast amounts of diverse datasets in huge volumes with the intent of using that data in machine learning, predictive modeling, and other advanced analytics to solve business problems and create actionable insights.

However, there's a challenge when running AI Large Language Models (LLM) because they require a substantial amount of computational resources and energy (electricity and cooling). That's why you don't see many enterprises making those investments; instead, they choose to use the AI services of cloud providers. A smaller, compact model called the Small Language Model (SLM) has also emerged and become popular. With fewer parameters and computational resources required, SLMs are more practical and easier to customize for enterprises that can't justify the high costs of running and training LLM, especially for their first AI project.

LLMs and SLMs are typically trained on publicly available data and lack exposure to specific information and to an enterprise's private data. These limitations often result in inaccurate and unreliable answers referred to as "hallucinations," a fact acknowledged by many publicly available generative AI chatbots, e.g., ChatGPT, that display warnings about potential misinformation. These hallucinations often appear to be plausible answers but are incorrect, with the possibility of damaging an enterprise's customer experience in many ways.



Using Retrieval-Augmented Generation (RAG) with InfiniBox to Solve Hallucinations

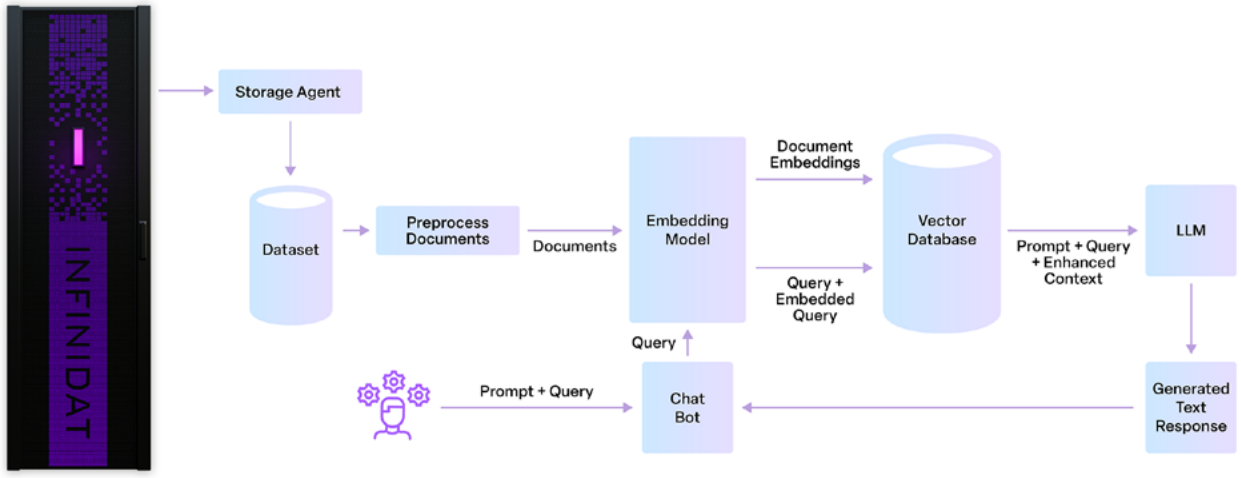
Using inaccurate data can lead to misinformed business decisions and provide customers with wrong information, damaging the enterprise's reputation and operational inefficiencies.

While powerful, Generative AI (GenAI) needs continuous augmentation with new, trusted, and accurate data to enhance and maintain its reliability. Retrieval-Augmented Generation (RAG) has emerged as a valuable tool to bridge this gap and improve accuracy. RAG combines the power of GenAI with an enterprise's private data to unlock fresh insights and streamline workflows. By integrating information from external sources into LLMs and SLMs, enterprises can achieve more accurate and specific results.

But where do you start? It doesn't need to be a daunting task. Chances are that users have data on existing systems and storage within their own infrastructure that is usable. For specific AI workloads, it's important to note that AI is just another workload for the InfiniBox, an enterprise storage solution known for its capability to deliver scalable, consistent I/O for diverse, modern workloads due to its innovative design.

A RAG workflow can be easily created from existing open-source products and data already in an enterprise's on-premises data center. Infinidat developers created an RAG workflow architecture outlining the process.

Infinidat RAG Workflow Architecture[1]



Infinidat's RAG workflow architecture runs on a Kubernetes cluster. Users who want to run RAG using data on-premises but without available GPU resources have a fast and convenient solution leveraging the cloud. Our approach uses a Kubernetes cluster as the foundation for running the RAG pipeline, enabling high availability, scalability, and resource efficiency. With AWS Terraform, we significantly simplify setting up a RAG system to just one command to run the entire automation. Meanwhile, the same core code running between InfiniBox on-premises and InfuzeOS™ Cloud Edition makes replication a breeze. Within 10 minutes, a fully functioning RAG system is ready to work with your data on InfuzeOS Cloud Edition.

The development of a RAG pipeline is an inherently iterative process that must be refreshed and maintained to keep it accurate. By keeping up with the latest advancements and continuously refining their RAG pipeline, enterprises can significantly enhance the accuracy and practicality of their AI-driven insights, maximizing the competitive advantages

The InfiniBox Advantage for RAG

For existing Infinidat customers, chances are the InfiniBox in your data center already has the "big data" repository of datasets to create the necessary documents to embed into a language model. Selecting a database to store embeddings (vectors) is crucial to your RAG system. Fortunately, current versions of commonly used database engines, such as Oracle, Postgres, MongoDB, and DataStax Enterprise, support storing and retrieving vector data, making them RAG-ready. Many of our current customers use their InfiniBox in mission-critical applications that utilize Oracle.

Our robust data services layer, provided by InfuzeOS, enables the control of all infrastructure components through software—a very powerful capability. InfuzeOS utilizes our innovative Neural Cache to optimize data placement dynamically and predictively. It places the most active data—hot data—in DRAM cache to minimize data retrieval from SSDs or HDDs. InfuzeOS ensures continuously

Fortune Business Insights projects that the big data technology market will grow from $349.40 billion in 2023 to $1,194.35 billion by 2032.

optimized performance at scale, best-in-class low-latency, and capable operational support for diverse modern workloads across on-premises and cloud environments.

What this means to a RAG workload is that during the embedding stage, where high-dimensional vectors are saved to a database, InfiniBox achieves lightning-fast performance with very low latency that will not bog down the process.

When a user poses a question (e.g. ChatGPT), their query is converted into an embedding that lives within the same space as the pre-existing embeddings in the vector database. With similarity search, vector databases quickly identify the nearest vectors to the query to respond. Again, the ultra-low latency of InfiniBox enables rapid responses for GenAI workloads.

Additionally, there are benefits to having the vector database stored on InfiniBox. Ingesting the data can be resource and time-intensive, with the results stored in the vector database. This database can be snapshotted and/or replicated, providing a vector database that is ready to plug into other RAG workflow deployments and reducing the resource requirements in other deployments.

Once the data is preprocessed and stored in an efficient, fast-retrieval data storage system (InfiniBox) the Large Language Model (LLM) comes into play. LLMs generate coherent and contextually relevant responses. After retrieving the relevant data from the database, the LLM combines the retrieved knowledge with its understanding to produce a comprehensive answer.

Lastly, InfuzeOS Cloud Edition for AWS and Azure provides cloud-based storage operations similar to those on-premises, enabling storage admins to have the same powerful InfuzeOS features of the data center in the cloud to rapidly replicate datasets to AI resources. This storage standardization reduces costs and the complexity of storage management across all environments.

The InfiniBox advantage to RAG primarily centers around our ability to provide the best-performing storage platform for this workload with innovations like Neural Cache. Infinidat's solution can encompass any number of InfiniBox platforms and enables extensibility to third-party storage solutions via file-based protocols such as NFS. But let's not forget that we also guarantee 100% availability and cyber recovery, all essential parts of an efficient AI infrastructure.

The use of GenAI and Retrieval-Augmented Generation (RAG) will substantially help make AI models more accurate and relevant. Enterprises can use their existing investment in the InfiniBox architecture, along with the above RAG workflow, to create fast and responsive AI models of their existing private datasets, unlocking fresh insights and streamlining workflows.

[1] Design and development of RAG workflow attributable to Andrew Wang

sales@infinidat.com

INFINIDAT