Making AI Models More Accurate with Retrieval Augmented Generation

Bill Basinas

Bill Basinas
Senior Director, Product Marketing
Infinidat

Biography

Bill Basinas is Senior Director, Product Marketing at Infinidat (https://www.infinidat.com) and has been focused in the storage industry since 1994 when he joined Legato Systems as the first field systems engineer.

He was also an early employee at Avamar and spent time at enterprise companies such as EMC and HPE Storage in Global Marketing and Engineering roles.

Bill blogs at https://www.infinidat.com/en/blog

Keywords Enterprise storage, Retrieval augmented generation (RAG), Enterprise data, Generative AI learning (GenAI)
Paper type Opinion

Abstract

Generative AI learning (GenAI) models are powerful, modern tools for automating interactive delivery of knowledge and information, but they often struggle to understand proprietary enterprise data. This leads large language models (LLMs) being prone to so-called "hallucinations" – fabricating answers instead of acknowledging that they don't know the answer. Using Retrieval augmented generation (RAG) helps LLMs address this issue by providing all the information needed to answer questions even if they are on topics on which it has not been trained. In this article, the author looks at why RAG is important and clarifies the important role that storage infrastructure now plays in RAG.

Introduction

Many analysts and industry experts are saying the same thing "enterprises deploying RAG will be able to do so on their existing storage infrastructure regardless of whether the storage is all-flash SSA or Hybrid". In the case of Infinidat, we believe that is especially true because of the performance capabilities that InfuzeOS™ enables in our InfiniBox® platforms and our ability to deliver outstanding performance for many diverse workloads. With over 17PB of enterprise -class capacity in a single rack, enterprises deploying RAG don't need to run out and spend big dollars for new storage infrastructure.

So, what do we do in our InfiniBox family that is different to support an AI RAG solution? Nothing! Ironically, that might be a strange answer. But the truth is that we don't have to do anything new because of the powerful capabilities we already deliver, enabled by our proven platform technology — the power of InfuzeOS, our patented Neural Cache technology, and our industry-leading low latency performance. We already deliver performance and low latency at scale. These benefits extend to vector databases, a critical component in GenAI. We are already recognized for how well we perform with mission-critical databases, like Oracle, which already supports embedding vector databases. Current versions of commonly used database engines support storing and retrieving vector data, for example, Oracle, Postgres, MongoDB, and Datastax Enterprise, making them RAG-ready. But that is a story for a later time.

The importance of RAG

Back to why RAG is important. Without iterative updating and fine-tuning of these models, which are static or only tend to leverage publicly available data, the return of a query will often deliver incorrect or misleading results referred to as "AI hallucinations." AI hallucinations appear as factually inaccurate content, false attributions, or citation of non-existent information. RAG workflow has emerged as a key tool to bridge this gap and provide continued refinement of data queries. RAG combines the power of generative AI models with enterprises' active private data to produce continuously updated, correctly informed responses to live queries, like ChatGPT.

In today's data-driven world, organizations accumulate vast repositories of data. Despite this, extracting actionable insights is a persistent challenge. The technology of Large Language Models (LLMs) has improved tremendously in recent years but is very resource intensive and requires substantial amounts of computational resources and energy (electricity and cooling to run the computational resources). To combat this, more compact versions of these models, called Small Language Models (SLMs), have emerged and are increasingly popular. Most enterprises do not have the budget or resources to deploy AI systems at the scale required to create Large Language Models. Regardless of the model, training workloads should not be confused with RAG; they are different, and training is well known to be resource intensive. Hyper-scalers have bulked up tremendously in this area to provide those types of training services, but once trained and operating within an enterprise, RAG is hugely important to produce accurate results.

Let's highlight a RAG workflow and how it can be a game-changer for your organization.

Figure 1: Simplistic view of a RAG architecture

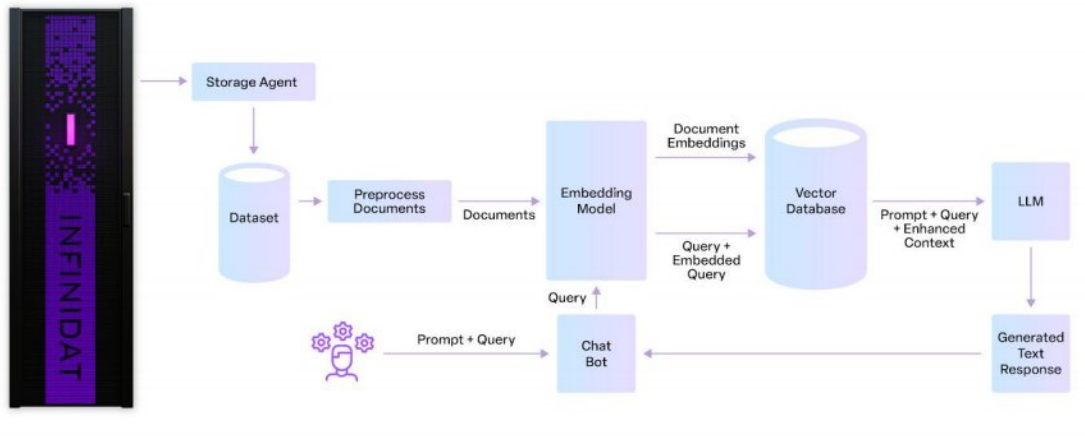
Basic components

The graphic (see *figure 1*) is a simplistic view of a RAG architecture. *Figure 2* shows that the development of a RAG pipeline is an inherently iterative process. By continuously refining a RAG pipeline with new data, organizations can significantly enhance the accuracy and practicality of AI-model driven insights, maximizing the benefits and promise of Generative AI technology.

AI RAG workflow diagram

I am not going to dive into all those areas in this article because we have a solution brief that goes into the details on many of these components that get deployed to support a GenAI architecture that leverages RAG to enhance and produce reliable results.

Figure 2: AI RAG workflow diagram



Infinidat's RAG workflow deployment architecture

We used a Kubernetes cluster as the foundation for running a RAG pipeline, which results in making it portable, scalable, resource efficient, and highly available. We used Terraform to significantly simplify the process of setting up a RAG system enabling just one command to run the entire automation. Within ten minutes, a fully functioning RAG system, hosted in the cloud, was ready to work with the data replicated from on-premises to InfuzeOS Cloud Edition (AWS and Azure).

Conclusion

Enterprises should leverage RAG using existing storage that is in-place data. In most cases, they do not need to invest in a lot of specialized resources like GPUs and storage. Many will tell you that you do, and they are expensive. We all know the answer is somewhere in between. To host and run the Large Language Model, yes, you likely may need specialized resources if hyper-scalers don't fit. But for RAG, it is a very different story, and we are well aligned to support an enterprise and can provide flexibility that can leverage the cloud as a fast and convenient deployment solution as well. 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. Our RAG solution makes it easy to leverage your most important data to produce the best possible and accurate results from your GenAI.