DATASTAX

WHITEPAPER

The Hyper-Converged Database

The Modern Architecture to Make Your Data AI-ready and Cloud-ready

Self-managed and hybrid data architectures are facing a range of new demands. Here's what's required for enterprises to modernize and keep up with the requirements of hyperconverged infrastructure, generative AI, and more.

Introduction

Many enterprises face a huge challenge with the infrastructure they use to store, retrieve, and manage data: the technology is outdated. The issue is particularly glaring with the rapidly shifting paradigm of how traditional, structured data is stored and used, and how new applications are being developed. Flexibility, security, and the ability to appropriately manage what data moves to the cloud and what remains within the corporate server room are all considerations that can complicate matters for organizations that are looking to modernize.

And then there's the huge push to infuse GenAI in the enterprise. This poses a particular set of challenges for the many companies that manage the majority of their data in their own data centers or in self-managed, private/public cloud environments: much of the advances in data storage, vector search, embedding and indexing, and other areas critical to GenAI has taken place in the public cloud or in hyperscaler environments.

When it comes to accelerating the development of applications that can leverage AI and retrieval-augmented generation (RAG), data infrastructure is a foundational area that needs to be addressed.

Here, we'll explore the technological shifts that are influencing enterprise data architecture, the new demands that GenAI places on hyperconverged infrastructure, and the requirements for self-managed data infrastructure in the age of GenAI.

Converging technological shifts

Technology shifts tend to be the major catalyst for modernization of enterprise infrastructure. The "if it ain't broke, don't fix" mentality is generally a tried-and-tested approach to operational success, but when technology advances to a point where investing in modernization comes with big enough benefits, an upgrade is required.

This is happening right now with much of what is being developed and deployed around GenAI both at the hardware and the software infrastructure layers. There are four major technology shifts happening now that play a significant role in how enterprise data architecture is changing and being modernized.

Hyperconverged infrastructure

Hyperconverged infrastructure (HCI) represents a major advance in enterprise architecture. HCI is generally described as a software-defined, integrated IT infrastructure that virtualizes the traditional enterprise elements of networking, compute, and storage onto on-premises servers.

Hyperconverged concepts are already built into a cloud-native architecture: consider the separation of storage and compute, which enables a componentized approach to rapidly scale up and scale down resources and features on demand. But with the technology advances being made by hardware providers like Nvidia, Dell, and Intel, the hyperconverged approach is extending to on-premises applications to take advantage of virtualized compute, software-defined storage, and virtualized networking—no matter where enterprise data resides..

Hybrid cloud

The original promise of cloud computing was to eliminate the need for self-managed hardware. However, the architectural reality of many enterprises requires a hybrid architecture that supports both self-managed on-premises, self-managed private/public cloud, and serverless managed public cloud.

In the telco industry, for example there are many use cases that require expansion into different regions where cloud services aren't available; a hybrid approach of both self-managed and cloud infrastructure is required in cases like this. Then there's the need for data security and privacy in the financial services or health care industries, where some data needs to be self-managed on-premises and another portion can be serviced in the cloud.

This hybrid cloud approach can present challenges if all the data is required to be in one location or stored in a single cloud environment. A truly scalable enterprise requires a hybrid cloud approach that combines the best of on-premises, hyper-scaler clouds, specialized AI clouds, and edge compute clouds and brings them together seamlessly.

GenAI

In 2023, many organizations started the journey to explore what GenAI could do to provide a better customer experience. Natural-language processing (NLP) chatbots were the talk of the town, and bringing enterprise-specific data into the conversation was key. Architectural patterns like RAG started to become the de facto standard on bringing enterprise data to GenAI applications.

NLP chatbots are just the beginning; now we are seeing specialized AI cloud environments designed to deploy enterprise-grade RAG and vector search capabilities directly alongside sensitive customer data—where that data resides. Take for example NVIDIA DGX Cloud, which runs on "AI super pods" and provides an AI platform to help enterprise developers build GenAI applications.

This environment is designed specifically to enable the deployment of highly scalable HCI on the edge.

Many enterprises are trying to figure out how to run enterprise RAG on sensitive customer data; this has to be accomplished where the data lives and requires a modern componentized approach that can scale across on-premises and specialized AI cloud environments.

5G edge computing

Another major area that will increase the demands for a modern approach to enterprise data is 5G edge computing. Over the past half-century, consumers' expectations regarding responsiveness have evolved dramatically. Fifty years ago, it was accepted that bank loans required a significant amount of paperwork and time (weeks on average) before any response was provided. Today, a financial services provider that doesn't offer the ability to apply for a loan on a mobile device with rapid feedback on eligibility is probably losing customers.

The demand for "instant gratification" has had a major impact on how we service data in a mobile, distributed world. With the advent of 5G, there's been a significant shift in the compute power of edge infrastructure and, more importantly, the architectural demands to push a large portion of the workloads out to the edge—to bring them as close to the customer as possible. Relying on legacy approaches and large monolithic application infrastructure won't support these kinds of workloads. It demands a lightweight, low-latency infrastructure that can service data at scale and be deployed across a hybrid environment, including environments like 5G base stations that can segment storage and compute and leverage data as close to the end use as feasible.

Hybrid GenAI workloads

Build an architecture designed to meet the most demanding challenges and typically it will get used for a broad range of different workloads. This holds true for the concepts and design principles around a hyperconverged architecture. Much HCI today is very well-suited for traditional workloads; there are clear benefits (like scalability, portability, and simplified deployment) that justify transitioning these workloads to HCI. Whether they'd planned for it or not, companies that have adopted HCI have built a strong foundation for the most demanding workload: GenAI.

As stated earlier, most of the experimentation in GenAI today has been focused on NLP chatbots. Most were built on public data; take OpenAI's ChatGPT, for example. There is some sensitive data being built into the chat bots, but most of the data being used is information that can be found on the internet or publicly available in an enterprise's documentation or on their website.

Consequently, NLP chatbots don’t really have to perform particularly quickly or efficiently. If a chatbot can provide me with an answer in five seconds versus two minutes of manually clicking through documentation to find my answer, that’s a marked improvement.

But GenAI is already moving from a primary focus on text into other modalities, like video, audio, and images. OpenAI’s Sora is a real-time text-to-video generator that “can create realistic and imaginative scenes from text-based instructions.” With Sora, I can enter a simple prompt:

“Animate a scene that features a close-up of a short fluffy monster kneeling beside a melting red candle. The art style is 3D and realistic, with a focus on lighting and texture. The mood of the scene is one of wonder and curiosity, as the monster gazes at the flame with wide eyes and open mouth. Its pose and expression convey a sense of innocence and playfulness, as if it is exploring the world around it for the first time. The use of warm colors and dramatic lighting further enhances the cozy atmosphere and the image.”

A still of a GIF created with OpenAi.com/index/sora

The above image is from a GIF created with Sora based on the prompt. As this technology starts to become more mainstream and these kinds of real-time GenAI use cases become more prominent, it will become critical to move the execution of these applications as close as possible to the customer.

This presents a significant challenge. As these AI workloads move closer to the edge, they’ll increasingly require a software infrastructure and stack designed for hyperconverged architecture.

A key benefit of HCI is the fact that it's designed to provide high compute on the edge, which provides faster processing for GenAI applications. (These capabilities are available now with DataStax HCD and NVidia inference microservices on AWS Wavelength- AWS's and Verizon's 5G Edge computing cloud.)

The other area that benefits from a hyperconverged hybrid approach: GenAI workloads that require high levels of security around sensitive data. When processing moves near the edge, a significant amount of that data will be considered sensitive information (photos and engagement data, for example). Enterprises will need to run inference in their private data centers with RAG as well as in public and private cloud deployments. Taking an open model like Meta's LLAMA 2 and deploying it onto an enterprise's infrastructure to conduct the inference might be acceptable, but deploying proprietary/personal data in a public cloud environment isn't, in most cases.

New solutions are being developed to address these concerns. Dell's AI Factory hardware, for example, consists of purpose-built on-premises clouds leveraging NVIDIA processors, fast storage, and is an ideal hardware platform to deploy a database for RAG.

Lastly, LLMs aren't getting smaller. Especially with the advent of large multi-modal models such as Google Gemini, trained on all the videos on YouTube, enterprises will want their own LLMs that have been customized to their specific data or fine tuned to their specific business, such as health care or financial services. This is where specialized AI clouds like NVIDIA DGX Cloud are attractive, because enterprises can get access to the latest AI superpods.

Self-managed database requirements

It should come as no surprise that HCI modernization efforts—particularly those aimed at supporting demanding GenAI use cases—require modern database technologies that can accommodate advanced workloads and are specifically designed for enterprises that are modernizing their datacenters with HCI.

Self-managed databases for the hyperconverged enterprise must meet a set of important requirements. They include:

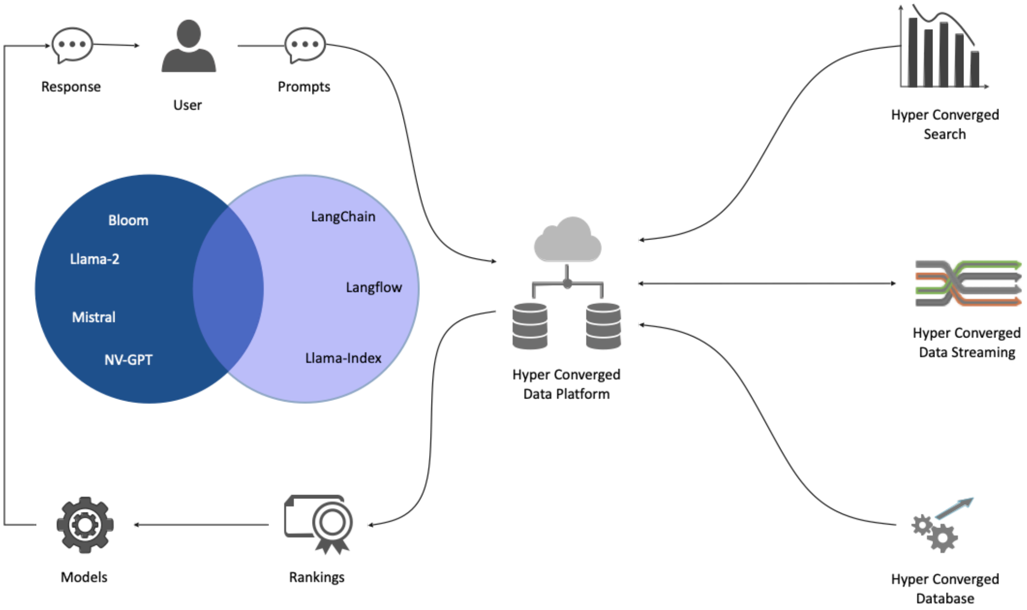
- **Separation of storage and compute –** Decoupling data storage from compute is becoming a hallmark of the modern database, and it's critical to enabling virtualized compute, software-defined storage, and virtualized networking. It also supports reduction of total cost of ownership (TCO), which is a key goal of HCI.
- **Developer productivity –** A flexible, scalable Database-as-a-Service that offers rapid provisioning and powerful, flexible data modeling, and easy-to-use data APIs makes it easy for developers to focus on building production-grade applications.

- **Support for AI workloads –** Production GenAI applications place new demands on databases, including enterprise-grade RAG and vector search capabilities. Security is critical when using sensitive proprietary data to build the smartest AI.
- **A hybrid design –** A database designed with an eye toward hybrid deployment is built on Kubernetes and powered with dynamic scalability that can be deployed on-premises and on any cloud with multi-environment replication capabilities.

The DataStax Hyper-Converged Database

GenAI has challenged many enterprises whose data is stored and managed in inflexible systems. How do you continue to service large-scale legacy applications that require traditional NoSQL functionality, for example, but adapt to the requirements and demands of GenAI?

DataStax recently introduced DataStax Hyper-Converged Database (HCD) to address these issues. DataStax HCD is data infrastructure for AI clouds, built for enterprises that have invested in data center modernization and HCI to support new AI and other advanced workloads, enabling enterprises to get the most out of data center modernization and total cost of ownership (TCO).



A hyper-converged architecture for GenAI

HCD offers a cloud-native architecture with elastic scalability. It's built on Apache Cassandra, which is renowned for its reliability and scalability in handling AI workloads. HCD includes DataStax Mission Control, which is designed to support modern DevOps practices and enables seamless deployment and management of data infrastructure while providing comprehensive observability capabilities.

Developer productivity is a key HCD benefit, as it enables rapid provisioning and intuitive data APIs, which help to streamline the development process. Additionally, HCD provides a comprehensive GenAI stack, facilitating RAG for enterprise applications, and it offers a vector search add-on, powered by the open-source JVector vector search engine, bringing SAI-powered vector search alongside OpenSearch-based full-text search capabilities.

While HCD provides a significant upgrade to managing Cassandra in HCI, it's one piece of what we're calling the "Hyper-Converged Data Platform" (HCDP). In addition to HCD, HCDP offers the ability to leverage data from microservices and events using Hyper-Converged Streaming (HCS), along with the added ability to explore, enrich, and visualize data for machine learning and rich data processing with Hybrid search with JVector and OpenSearch . (Learn more about DataStax HCDP here.)

The future is here

In this paper we explored how rapid shifts in technology require enterprises to look at modernization of the infrastructure to take advantage of HCI and the massive amount of compute that is being virtualized and pushed to the edge, specifically to drive and power generative AI applications.

We're at a crossroads with this technology, and what seemed like fantasy is rapidly becoming a reality. Take the 2013 movie HER, where Joaquin Phoenix's character Theodore Twombly falls in love with a multi-modal, real-time, GenAI assistant. The assistant has access to Joaquin's most sensitive data and delivers a multi-modal, real-time experience via Twonbly's smartphone. Just a decade later, GenAI, 5G, AI clouds, and hybrid cloud have all converged to enable comparable enterprise digital experiences with customer service and sales agents. Just look at the recent announcement made by OpenAI showcasing an AI assistant that provides language translation in real-time.

The future seems to be approaching faster than ever before, and enterprises that are modernizing with hyperconverged infrastructure will do well to prepare for that future by ensuring their data infrastructure keeps up.

Download the DataStax HCD 1.0 preview.

About DataStax

DataStax is the company that powers generative AI applications with real-time, scalable data with production-ready vector data tools that generative AI applications need, and seamless integration with developers' stacks of choice. The Astra DB vector database provides developers with elegant APIs, powerful real-time data pipelines, and complete ecosystem integrations to quickly build and deploy production-level AI applications. With DataStax, any enterprise can mobilize real-time data to quickly build smart, high-growth AI applications at unlimited scale, on any cloud. Hundreds of the world's leading enterprises, including Audi, Bud Financial, Capital One, SkyPoint Cloud, VerSe Innovation, and many more rely on DataStax to deliver real-time AI. Learn more at DataStax.com.