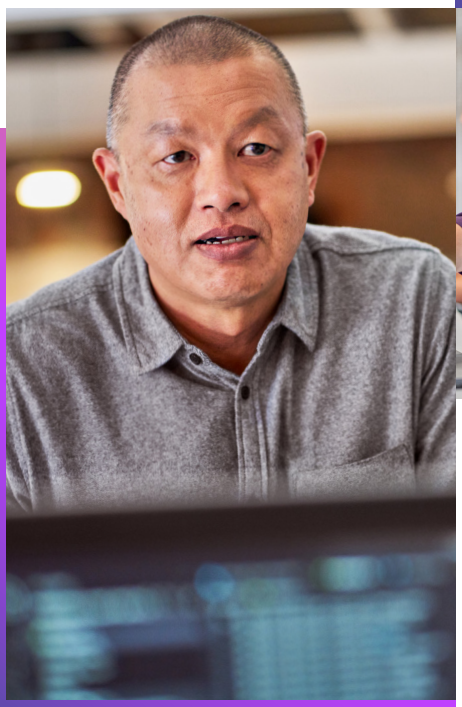
Hewlett Packard Enterprise

The secret to shortcutting your AI journey

Implementing AI at scale can be daunting. NVIDIA AI Computing by HPE can reduce complexity and speed enterprise adoption



Artificial intelligence (AI) is now a key destination on virtually every enterprise technology road map. Today, 72% of organizations have deployed some form of AI, with GenAI by far the most common.[1] However, where organizations are on their AI journey varies wildly.

Most companies are still at the getting-to-know-you stage with GenAI, typically by encouraging their employees to use public-facing chatbots. A smaller group of enterprises are running proofs of concept using limited datasets, to identify the best use cases for the technology. Less than a third of most enterprise GenAI projects have moved from experimentation into production.[2]

A lot of this early experimentation is happening in the public cloud, says Peter Moser, senior distinguished technologist at HPE. The reason? It's easy to get started, develop a simple model, and test it in a sandboxed environment. But that can change quickly.

"Once you start scaling and dealing with larger amounts of data, AI quickly gets complicated and expensive," says Moser. "A lot of customers discover they don't want their data in the public cloud, out of their control. They want it on-premises." That shift is already being well documented: Today, more than 60% of enterprises say they plan to train and deploy their AI apps in an on-premises private cloud.[3]

[1] "The state of AI in early 2024: GenAI adoption spikes and starts to generate value," McKinsey & Co., May 30, 2024

[2] "Now decides next: Moving from potential to performance," Deloitte, August 2023

[3] "Essential Elements for Private Cloud Strategies," IDC, August 2024

A key driver for private cloud is data sovereignty and security, as no organization wants to risk exposing its crown jewels in the public cloud or running afoul of data governance regulations, which continue to emerge. The private cloud also offers greater control over app performance, latency, scalability, and cost, adds Moser.

But going from limited experiments in the public cloud to production-scale AI applications on-premises is an enormous leap, which is why 9 out of 10 pilot projects don't make it to the other side.[4] "The vast majority of enterprises lack the necessary technology infrastructure and expertise to make AI work at scale," says Moser. In other words, they need help.

So, you've decided to adopt a private AI cloud. Now what?

Only the world's most advanced organizations, many of which have been working with machine learning for years, have the internal tools and expertise necessary to deploy GenAI applications without aid today, Moser says. The rest are facing a daunting learning curve, as well as sizable investments in technology and time.

"High performance compute workloads like AI use parallel file systems," Moser explains. "They require GPUs and complex tooling that most IT organizations do not have. This is all new to them."

To train these AI models, organizations will need to figure out how to consolidate huge volumes of unstructured data stored across multiple formats and locations. They'll have to recruit data specialists, upskill their IT teams on how to work with AI, deploy appropriate data governance guardrails to minimize risk, and integrate all these new systems into their existing technology infrastructure.

Meanwhile, the clock is ticking. Market pressure to rapidly adopt and deploy GenAI is only going to increase.

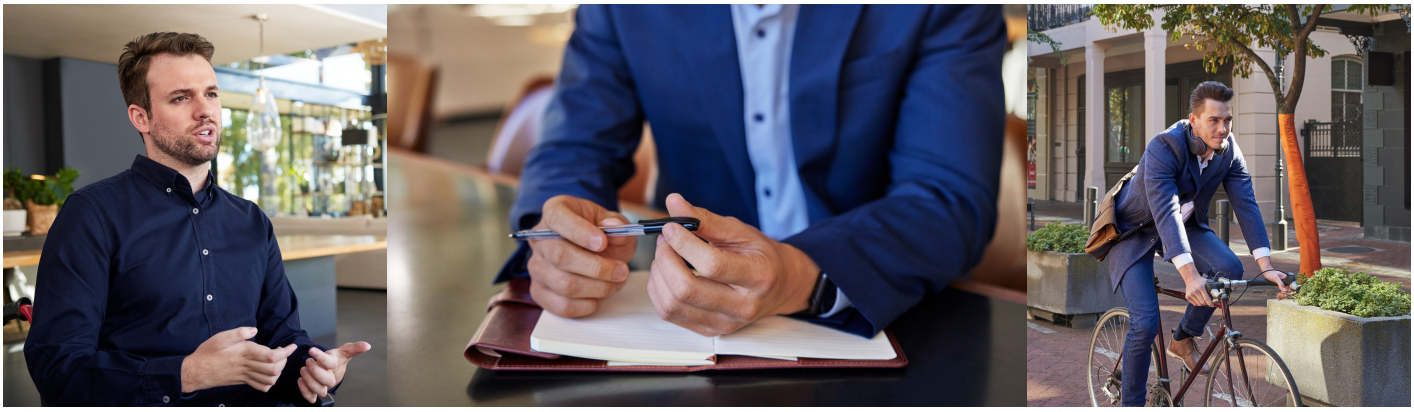
"Trying to do all of that yourself takes a tremendous amount of time and involves a great deal of risk," Moser says. "Most enterprises simply don't know enough about AI to do this in a timely manner."

HPE Private Cloud AI: Add data and stir

All the factors described earlier are what drove Hewlett Packard Enterprise to partner with NVIDIA on a unique, turnkey solution for enterprises: HPE Private Cloud AI.[5] This is more than just infrastructure. The scalable, pre-configured, full stack private cloud option is part of the NVIDIA AI Computing by HPE portfolio. This completely integrated AI solution encompasses compute and storage, design and planning, integration and optimization, education and training, consulting, and more.

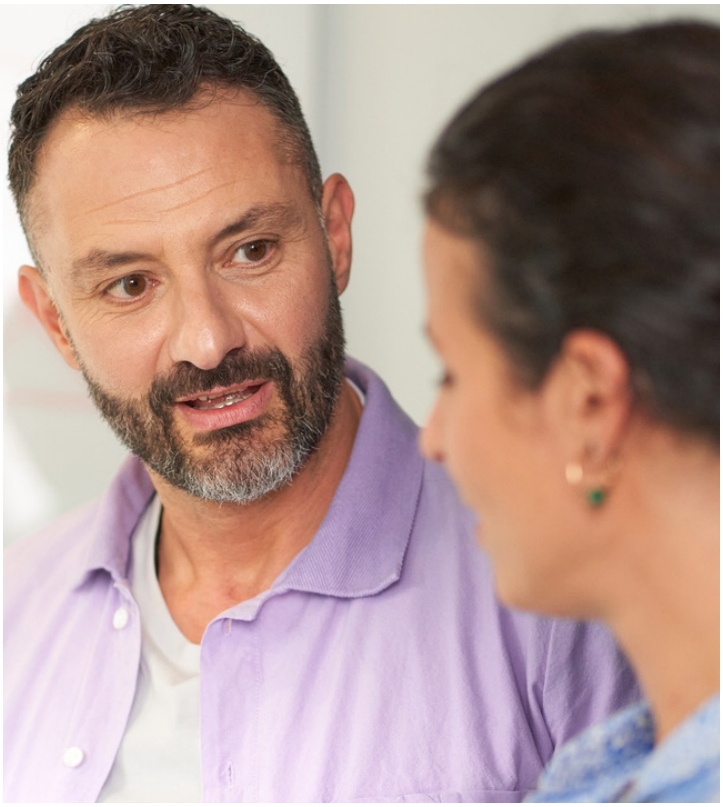
The process starts with an in-depth conversation about the kinds of opportunities AI presents for your enterprise, an assessment of how AI ready your organization is, a discussion of best practices, and the development of business use cases tuned to your needs. HPE offers workshops for enterprises that want to get started on their AI journeys but don't know where to begin or even what questions to ask, says Moser.

"One of the big mistakes a lot of enterprises make is they don't know everything AI can do," he says. "They limit themselves right out of the gate on the art of the possible.



[4] "Reasons Why Generative AI Pilots Fail To Move Into Production," Forbes, Jan 8, 2024

[5] "Hewlett Packard Enterprise introduces one-click-deploy AI applications in HPE Private Cloud AI," HPE, Sep 5, 2024



These workshops help them focus on the priorities that are most valuable to the business. They change the whole conversation."

The next step is technical design and planning. Not all enterprises will have the same needs or levels of ambition for their GenAI projects. That's why HPE Private Cloud AI comes in four preconfigured server offerings, depending on the compute and storage needs of each use case and whether they'll be deployed for training, fine-tuning, or inference.

"You're getting something that's already been designed and assembled by the companies that make all these tools and certify it's all going to work together," says Moser. "On the next day, we'll come in and install it. Within 48 hours, it's up and running."

Each HPE ProLiant server HPE configures is certified to work with 70 large language models (LLMs) available for download, along with common business use cases. HPE Ezmeral Data Fabric allows enterprises to knit their distributed datastores into a cohesive whole, making it easier for LLMs to ingest petabytes of unstructured training data. Enterprises simply need to choose the right LLM for their business needs, feed it their AI-ready data, and begin uncovering insights.

Supporting the entire AI lifecycle

As systems enter production, enterprises can choose to manage the systems themselves or call upon HPE and its partners for guidance on how to integrate turnkey solutions from HPE and NVIDIA with their existing security and data governance frameworks. Expertise is available throughout the entire AI lifecycle, from implementation through optimization.

HPE can also help enterprises walk through thorny issues around how to use data in an ethical and transparent manner, avoid privacy violations and copyright infringement, and comply with an increasing number of AI regulations.

"Getting AI right is incredibly complex," Moser adds. "People don't know what they don't know, because they've never done it before. And those who've tried to do it all themselves are finding out it's a lot more complicated than they thought."

Learn more at
HPE.com/AI

Visit **HPE.com**

Chat now

Hewlett Packard Enterprise

a50011771ENW, Rev. 1