On-premises AI approaches: The advantages of a turnkey solution, HPE Private Cloud AI

Artificial Intelligence (AI) theories and approaches have been around for decades, but in 2022, AI became more accessible and understandable to users everywhere with the release of ChatGPT. Since then, the pace of AI improvement has been fast and furious, with a whole ecosystem emerging of model development, AI-specific hardware product lines, and services that tie all the technology together. The rapid pace of AI ecosystem change can be overwhelming for organizations looking to harness the power of this tool. Understanding, assembling, and deploying the hardware and software elements necessary for these complex systems might require greater technical expertise than these organizations have.

Alongside the complexities of building AI solutions are the ever-present concerns of compliance and security. How do organizations begin to capture value from AI without surrendering their private data and intellectual property to public-cloud-based AI systems, which may have varying ways of capturing and using that data? Decision-makers who wish to strike this balance are looking at private cloud AI solutions, which can offer a turnkey way to reap the potential benefits of AI while maintaining control and governance of their systems.

Gain the advantages of a turnkey fully managed on-premise AI solution with HPE Private Cloud AI

Streamline procurement and deployment with pre-engineered, full-stack configurations

Accelerate data science productivity with pre-installed data engineering applications

Get a RAG-based LLM application up and running in just a few steps

Summary of findings

Turnkey approach – HPE Private Cloud AI – Simplest and fastest time to value

- Deployment involves selecting a pre-determined size and after receiving, can be up and running in days after delivery
- Updates handled entirely by HPE
- Built-in tools and private cloud integration allow data management from a single pane of glass, saving engineering and IT management cycles

Reference architecture approach – Dell AI Factory with NVIDIA – Moderate complexity and time to value

- Deployment involves use case definition with sales, pre-scoping, and services
- Updates provided by Dell and NVIDIA
- No private cloud integration out of the box

Build-your-own approach – Greatest complexity and time to value

- Deployment involves manual selection, configuration, and procurement of components, can be a months-long undertaking
- Updates procured and installed by internal IT staff or a third party

Introduction

In this paper, we discuss the considerations around public cloud versus on-premises for AI, dive deeper into an on-premises discussion, and finally contrast three on-premises approaches available to buyers today: build-your-own, reference architecture + services, and turnkey. As part of the discussion with each approach, we contrast two examples: a turnkey approach with HPE Private Cloud AI and a competing solution, a reference architecture + services approach with Dell™ AI Factory with NVIDIA.

We have attempted to evaluate details such as procurement processes and hardware and software availability, and to infer time to value from these details. That said, we did not purchase any of these products or conduct any hands-on evaluation, and so we cannot precisely quantify pricing, procurement time, or deployment time. We base our comparative conclusions on our own in-house knowledge of AI complexities for the build-your-own approach, and on publicly available information for the reference architecture and turnkey approaches.

Based on publicly available data as of April 2025, we believe HPE Private Cloud AI is a strong solution using a turnkey approach for customers looking to quickly onboard and host a full-featured AI technology stack in a secure on-premises environment. The HPE solution operates within a private cloud operating model (HPE GreenLake Cloud), offers detailed predetermined solution sizes, includes pre-installed data engineering and data science applications, and uses a wizard-based deployment approach that guides the deployment team through the required screens.[1] Customers requiring a turnkey approach and little ongoing maintenance, such as the one HPE provides, would likely have an easier time with procurement, deployment, and initial use than those using a build-your-own or a reference architecture approach, which would both likely require more pre-sales research, configuration work, and ongoing maintenance. HPE Private Cloud AI customers may also avoid integration failures, security gaps, and delayed rollouts that often derail AI projects before they begin.

The merits of on-premises AI compared to public cloud

Implementing AI applications requires determining the best platform for your specific workloads, and there is no single correct approach. AI is not a monolith, but a myriad of use cases, models, workloads, and needs. Early in your AI deployment decision-making, you must make several core architectural choices, including whether to host your data and application on the public cloud or on premises. Compared to public cloud options, on-premises AI solutions can offer advantages in data control, compliance, security, and cost predictability. Below we discuss some of these advantages in more depth before reviewing several different approaches to on-prem AI deployments.

Public cloud AI considerations

For newcomers to AI, the public cloud is a practical starting point, offering resources and management tools for quickly building, training, and deploying AI models. However, there can be considerable downsides to using public cloud solutions, including:

- Limited control over the underlying environment
- Risk of security breaches and cloud outages
- Data sovereignty and compliance issues, especially in regulated industries
- Lack of data and compute adjacency, with the attendant possibility of latency issues
- Lack of cost predictability
- Locked-in to a vendor-specific set of cloud-centric tools

Security is a key reason that many companies decide to deploy AI on-premises. If you deploy in the public cloud, having underlying virtualization, container, or storage software out of your control means you must trust your public cloud provider not to collect more data than advertised—intentionally or not.[3] Even with the best security practices in place, public cloud providers may provide privileged access to engineers investigating issues during outages or troubleshooting periods, which can expose sensitive user or company data.[4]

Perhaps more troublingly, public cloud security breaches, while rare, do happen. Many companies cannot take this risk. According to a recent Apple-sponsored study, "In the first three quarters of 2023, the number of ransomware attacks increased by 70% compared to the first three quarters of 2022."[5] The report also states that "98% of organizations have a relationship with a vendor that experienced a data breach in the last two years" and "Over 80% of data breaches involved data stored in the cloud."[6] For organizations whose employees may use AI chat-based applications—which ingest data in the form of prompts—in the public cloud, these statistics should be a serious reason to consider transitioning to a on-premises architecture. Attacks can happen anywhere, but by keeping all data and infrastructure on premises, you ensure that you remain in full control of your deployment and its security.

Data location can also be a compliance issue that some companies must consider when planning for and deploying AI applications. Many industries, and some regulatory bodies, have strict requirements for where and how organizations may store user data. Bodies of regulations, such as GDPR, require businesses to protect data they have collected from users in the EU.[7]

Data security and compliance are just two reasons that a company would want to control the location of its data, particularly sensitive proprietary and user data. Another potential benefit of location control is improved performance. AI applications have strict latency requirements for optimal performance, and locating your data closer to the compute-intensive AI application can decrease latency and increase performance.

On-premises AI challenges

While on-premises AI deployments can address many of the concerns companies have about hosting their AI applications on public clouds, they also present their own challenges. It is important to select a deployment approach that mitigates these challenges, which we discuss more in depth below. Researching, procuring, building, managing, and hosting AI applications on site means that your teams are responsible for understanding those applications' needs, right sizing the environment, and performing many other activities. The myriad choices across every aspect of the environment—from data storage to model version to hardware—can slow the pace of business. In addition, your infrastructure, security, and data science/AI teams may lack the requisite knowledge, training, or time to adequately handle all these decisions.

These solutions also require piecing together a variety of components up front, including server hardware, networking, software licenses, and professional or third-party services, all of which must function together seamlessly. Even once you have assembled the components and your solution is operational, your teams must manage it, which may require additional staff. To make an on-premises solution as efficient as possible, it would be based on private cloud technologies and function in as turnkey a manner as possible, mitigating these planning, deployment, and management challenges or removing them altogether.

In the remainder of this paper, we look at three on-premises deployment approaches—build-your-own, reference architecture + services, and turnkey with HPE Private Cloud AI—and review example offerings to see what they provide and how well they address these challenges.

On-premises AI deployment approaches

As executives push their companies to take advantage of AI for cost reduction or growth, IT organizations everywhere are evaluating what they must do to create a successful and scalable AI technology stack deployment. The choices can be overwhelming from multiple perspectives: financial, training and staffing, deployment time, hardware and software compatibility, user access and security, and ongoing maintenance. The right approach for each organization depends on exactly what that organization needs and what resources and staffing it already has in place.

For this paper, we focus primarily on larger enterprise solutions that combine decoupled storage and compute, NVIDIA GPUs, and software. We do not include hyperconverged solutions. The three approaches that we studied for this paper fall on a continuum, from more complex to less.

We define the three approaches as follows:

- A **build-your-own approach** involves the IT organization selecting and purchasing the hardware and software for its AI implementation, installing it all themselves, and maintaining it all themselves. We discuss this option without a specific example, as each company's piecemeal approach would be different.
- A **reference architecture + services** approach involves a vendor pre-validating hardware and software, but does not include such additional tools as built-in portals or a built-in private cloud framework. The solution we review for this approach is the Dell AI Factory with NVIDIA.
- A **turnkey approach** includes the pre-validated hardware and AI software that the prior approach used, but is truly turnkey: it offers quick delivery and easy installation, and is ready for use with an assortment of higher-level, AI-related software that enriches the base set of tools. The solution that we review for this approach is HPE Private Cloud AI.

For all approaches, software is required to accelerate application development with the GPUs in each solution. Based on documentation, the latter two approaches above include access to NVIDIA AI Enterprise for this purpose. For simplicity and comparison purposes, we also assume that the build-your-own approach would use NVIDIA AI Enterprise. We discuss this in more detail later in this paper.

In Figure 1, we show what each deployment approach includes.

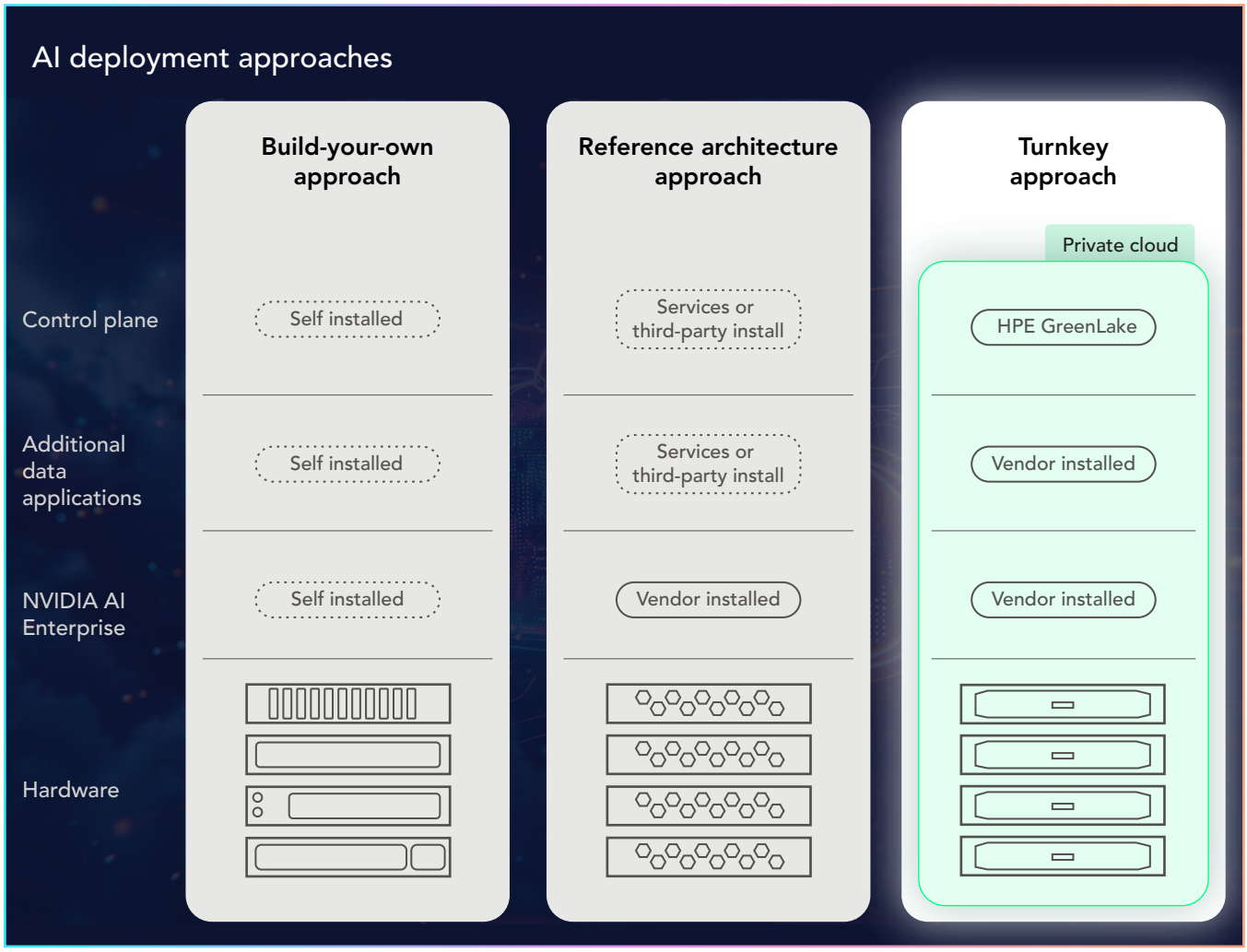

Figure 1: Hardware and software included in the three deployment approaches we discuss.

We present greater detail on the different approaches in Table 1, with categories of features or components and our assessment of what each includes.

Table 1: Features and components of the three approaches to on-premises AI implementation. Source: Principled Technologies.

	Build-your-own approach	Reference architecture approach – Dell AI Factory with NVIDIA	Turnkey approach – HPE Private Cloud AI
Pre-validated hardware and AI software	No	Yes	Yes
Customization required	Maximum	Moderate	None
Complexity	Greatest	Moderate	Least
Procurement approach	Manual selection, configuration, and procurement	Use case definition with sales, pre-scoping, and services	Order pre-determined size and receive
Component and software version interoperability	User must determine	Pre-validated	Pre-validated
Deployment time after procurement (time to value)	Months[8]	Unknown	Hours[9]
NVIDIA AI Enterprise	Included with GPUs, IT or third party must install manually	Installed by professional services, accessed via NVIDIA toolsets	Integrated with HPE AI Essentials and data tools from the factory. Accessed within private cloud management console and IDE
Libraries of models, frameworks, and connectivity options (NVIDIA AI Enterprise)	IT or third party must install manually via NVIDIA AI Enterprise toolset	Vendor installs via NVIDIA AI Enterprise toolset	Vendor installs via NVIDIA AI Enterprise toolset, plus others
Integrated development environment	IT or third party must install manually	Not clear from our research	Integrated with no installation required
Data engineering and data science toolsets (Apache® Spark, Apache Airflow, similar applications)	IT or third party must install manually	IT or third party must install manually	Integrated with no installation required
Deployed in a private cloud ecosystem	No	No	Yes
Role-based control access (RBAC) active out of the box	No	No	Yes
Primary interface to toolsets	A mix of command-line and UI	A mix of command-line and UI	Private cloud portal for infrastructure ad-min, AI admin, and AI user
Ongoing maintenance	IT or third party must maintain	Hardware provider and NVIDIA provide updates, IT or third party must install	Vendor provides

Build-your-own

> The build-your-own approach requires the largest number of decisions, the most ongoing responsibility, and the greatest staffing needs.

In a build-your-own approach, a company constructs its own solution from scratch. While maximizing flexibility, this approach also requires the largest number of decisions, the most ongoing responsibility, and the greatest staffing needs. Organizations must research and specify compatible hardware (servers, GPUs, networking, storage) and compatible software elements (OS, orchestration, containerization, GPU libraries, model interfaces such as vLLM, and data engineering and data science software). Then, they must order, install, configure, license, and maintain all those components. Beyond procurement, deployment, and ongoing IT maintenance, organizations must develop an approach—perhaps an application or an assortment of applications—for consuming the services or offering the technology to users. This might involve scoping additional installations, such as Kubernetes, Apache Spark, Apache Airflow, MLFlow, Grafana, and so on.

The build-your-own approach may provide the most options for customization, but it is certainly the most complex of the three scenarios we describe. To staff such a solution, you would need server hardware experts, networking experts, and ongoing IT admin staff, in addition to the AI developers, data engineers, and data scientists that would eventually use the solution. Keeping up with the daily changes of AI-related codebases, open-source branches, and hardware compatibilities could consume multiple full-time employees. For role-based security, you would need to define those roles and associated policies, then integrate the AI technology stack into your security practices. You'd also need to coordinate ongoing maintenance schedules with hardware, firmware, and software patches.

It is difficult to precisely quantify the time required for build-your-own planning and implementation phases, as each environment, toolset, and staff skillset is unique. Most organizations would experience many, if not all of these phases:

- Multi-team planning and coordination (infrastructure, security, finance, development)
- Hardware and software compatibility research
- Proposal development
- Procurement
- Delivery and preliminary setup
- Testing and validation
- Production deployment

As we note above, the idea-to-production lifecycle can vary greatly, but recent sources suggest timelines approaching or exceeding eight months. For example, according to Gartner in 2024, the "prototype to production lifecycle takes around eight months on average."[10] When including the extra time necessary to plan for the prototype, the total amount of time would probably be longer than eight months.

Smaller organizations, more-experienced staff, and simpler deployments could shorten these timelines, but the combination of larger organizations, less-experienced staff, or more-complex deployments would likely increase the time required.

While the build-your-own approach may provide the ultimate in flexibility, we believe it would also require the biggest price tag in labor cost, take the longest to implement, and greatly increase complexity.

Validated reference architecture + services

An approach that mitigates many of the issues with build-your-own is a reference architecture + services approach. In this scenario, a company may know that they have certain AI use cases—say, an application requiring a RAG-based LLM or specific data science initiatives—but not know how to configure or deploy a solution. In such a situation, a vendor could provide value by pre-validating hardware and software compatibilities; completing the installation with professional services; and providing additional professional services, third-party engagements, and/or the components necessary for subsequent development.

However, using this approach, the sales and configuration processes would likely require heavy touch and be driven by the organization's specific use cases, requiring moderate to heavy input from the organization to estimate sizing and software needs. Furthermore, the eventual solution could require additional application installations or specialized staff to access and consume the technology. As we define it, this scenario does not include a ready-to-go, easy-to-use portal for data engineers or cloud administrators to log in and begin their work. The IT organization would need to do additional work to set up RBAC, SSO, data security, QoS, monitoring, and so on.

> The reference architecture approach would likely require moderate to heavy input from the organization to estimate sizing and software needs.

NVIDIA AI Enterprise

NVIDIA AI Enterprise is a bundle of software components NVIDIA offers and supports to enable AI and machine learning. It includes libraries, containers, models, and repositories for streamlining development of generative AI applications. For more information, see: https://docs.nvidia.com/ai-enterprise/index.html.

Example: Dell AI Factory with NVIDIA

The Dell AI Factory with NVIDIA solution employs the reference architecture + services approach as we have defined it, combining a defined hardware solution with software and professional services.[11,12] The solution is multi-node, based on NVIDIA GPUs and networking, Dell servers, and separate (i.e., not hyperconverged) storage.[13] To explore this offering, we researched publicly available information.

Dell AI Factory is an umbrella term for all Dell AI services, hardware, offerings, etc., while Dell AI Factory with NVIDIA is a specific offering within that umbrella.[14] Dell AI Factory with NVIDIA consists of Dell compute and storage, NVIDIA networking hardware, Dell services, and NVIDIA AI Enterprise software.[15]

The Dell AI Factory with NVIDIA solution can include the following:

- A rack-scale solution with Dell and NVIDIA hardware and software for several different AI applications,[16] including GenAI: [17,18,19,20,21]
- Dell professional services[22]
- NVIDIA components including NVIDIA AI Enterprise software, NVIDIA Tensor Core GPUs, NVIDIA Spectrum-X Ethernet, and NVIDIA Bluefield DPUs[23]
- Several Dell PowerEdge™ server models (PowerEdge XE7740, XE7745, XE8712,[24] and XE96808L) that can support NVIDIA Blackwell GPUs and direct liquid cooling (DLC[25]
- NVIDIA Metropolis, NVIDIA Riva, and NVIDIA NIM offering[26]

For Dell AI Factory with NVIDIA, preconfigured solution sizes are difficult to find online, which could frustrate customers who want to move quickly or require technical visibility. The Dell solution is driven by use case scoping,[27] meaning that procurement would likely require time and involvement from sales, professional services, the organizations' teams involved in use case sizing, and more.

Dell AI Factory includes NVIDIA AI Enterprise.[28] This bundle of software and tools is strong, but is merely a bundle of related toolsets. It includes only NVIDIA components and does not tie all services together in a unifying portal or include other data engineering applications outside of the NVIDIA ecosystem. The Dell solution is not natively part of a private cloud ecosystem, and to install additional data engineering applications or configure connectivity to your organization's security and monitoring, you'd need to put in additional time (or engage professional services).

Some key considerations of a reference architecture approach include:

- Precise architecture details (exact hardware, software, sizing, and applications) may not be immediately available, and the sizing activity is part of use case scoping.
- Customers might need to work with additional third-party providers or engage additional professional services for AI use cases that go beyond a hardware + NVIDIA AI Enterprise installation.[29]
- A reference architecture approach such as the one Dell offers provides valuable hardware and NVIDIA AI toolsets, but does not by itself qualify as a turnkey solution as we define above (additional software, governance, private cloud, etc.). From our research, it also lacks a unified private-cloud experience and pre-installed data pipeline tools.

Turnkey approach

The last approach on our continuum is the turnkey solution. This approach builds on the advantages of the reference architecture solution in that it provides pre-validated hardware and AI software, but includes multiple additional layers of value. In this scenario, the deployment is wizard-driven and quick; upstream data engineering applications are pre-installed; and frameworks compatible with normal enterprise operations—such as RBAC, single sign-on (SSO), quality of server (QoS), and other administrative tools—are included out of the box. In addition, the vendor handles software version maintenance and hardware updates, which should increase the stability and reduce the complexity of the solution. HPE Private Cloud AI falls into this category.

> The turnkey approach provides pre-validated hardware and AI software, but includes multiple additional layers of value.

Example: HPE Private Cloud AI

HPE Private Cloud AI, co-engineered by HPE and NVIDIA, directly addresses many of the challenges that exist with on-premises solutions. We base our discussion of HPE Private Cloud AI on publicly available documentation and a private demonstration with HPE.

HPE Private Cloud AI bases its cloud model on HPE GreenLake and offers a pre-built, fully engineered solution that includes both the underlying hardware and software available at different sizes. This approach helps reduce the hardware and software research required to make decisions. Like the Dell solution, it integrates with the NVIDIA AI Enterprise software stack, but unlike the Dell solution, it pre-installs additional data science and data engineering applications, such as Apache Spark, Apache Airflow, Kubeflow, MLFlow, EzPresto, Ray, and Apache Superset.[30] On deployment day, according to HPE, the "installation and integration of the solution is done by HPE and takes less than 8 hours."[31] Finally, for day 2 operations, the solution includes IT lifecycle management, with HPE handling ongoing maintenance and updates.

HPE Private Cloud AI is a well-defined offering with messaging and documentation that informs customers about what they are purchasing, adding clarity to the process. In our research, publicly available information on the HPE solution provided ample information regarding the turnkey nature of the product and a high-level list of pre-installed software.[32]

Time to value

According to HPE, "The HPE Private Cloud AI solution offers a turnkey process for customers who want to simplify the overall setup experience at their site."[33] According to the HPE site, there is a 5- to 7-day response time for a quote from HPE for HPE Private Cloud AI.

HPE Private Cloud AI also includes pre-installed software, and it can reduce or eliminate some of the costs of the build-your-own approach:

- Less multi-team planning and research time is necessary
- HPE handles hardware and software compatibility research
- HPE handles testing and validation
- HPE handles build and deployment time of hardware
- HPE handles patch cycle and update management dependencies
- HPE includes AI framework tools and data pipeline tools on the solution at install time

Architecture

Built on the HPE GreenLake cloud, HPE Private Cloud AI provides an edge-to-cloud unified private cloud experience comprising both hardware and software. HPE Private Cloud AI is not a siloed AI appliance—it is embedded into a broader hybrid cloud strategy via GreenLake, enabling consistent governance, metering, and expansion across the enterprise. According to its documentation, once HPE deploys the solution in your data center or colocation, you can get a RAG-based LLM up and running with only three steps.[34]

> Built on the HPE GreenLake cloud, HPE Private Cloud AI provides an edge-to-cloud unified private cloud experience via GreenLake, enabling consistent governance, metering, and expansion across the enterprise.

The hardware footprint ranges from the HPE Private Cloud AI for developers featuring two 96GB NVIDIA H100NVL GPUs and 32 TB of storage,[35] to the Large configuration, which includes four HPE ProLiant DL380a Gen11 servers with 16 GPUs total.[36] The system comes pre-installed with software to aid in data pipeline management and AI/ML workflows and includes a single global namespace via the embedded data lakehouse.[37] Below, we discuss additional details of the pre-deployment, deployment, hardware, and software. For greater flexibility, HPE can deploy the solution in your data center or in a colocation facility through its partnership with Equinix to provide an easy, affordable deployment option for companies with data center constraints.[38]

Hardware

"HPE Private Cloud AI includes AI-optimized hardware that is delivered as a single rack in small or medium configurations. A developer configuration is available for a predictable starting point, small configurations are for basic LLM inference, and medium configurations can support RAG for LLMs. Additionally, large multi-rack configurations are available, capable of fine-tuning complex models."[39] The solution is built on a foundation of HPE ProLiant DL325 Gen 11 control nodes, HPE ProLiant DL380a Gen 11 worker nodes, and HPE GreenLake for File with Object Storage enabled.[40] Table 2 presents details on the pre-configured size configurations.

Table 2: Description and details of the available HPE Private Cloud AI sizes. Source: Principled Technologies, based on information in the HPE Private Cloud AI User Guide.

	Developer edition	Small – AI inference	Expanded Small – AI inference	Medium – AI inference and RAG	Expanded Medium – AI inference and RAG	Large – AI inference, RAG, and model fine tuning
Number of DL380a Gen11 AI nodes	1		2		4	
GPU type	NVIDIA H100NVL	NVIDIA L40				NVIDIA H100NVL
Number of GPUs	2	4	8		16	
Number of DL325 Gen11 control nodes	1	3				
HPE GreenLake for file storage capacity)	32TB	109TB		217TB		670TB
Switch type	N/A	SN4600cM		SN4700M		
Bandwidth	200GbE	100GbE		200GbE		400GbE
Number of 42U racks with PDUs	N/A	1				2
HPE AI Essentials with NVIDIA AI Enterprise	Included					

Software

As we note above, at its top layer of management, HPE Private Cloud AI is integrated within the HPE GreenLake console.[41] Administrators who are used to the GreenLake console will find user management and screen navigation familiar. The software presents three roles for cloud management and consumption: Cloud administrator, AI administrator, and AI developer. The unified presentation of screens and toolsets for each role combine to provide a single approach to managing users, tools, data sources, and applications.

> Administrators who are used to the GreenLake console will find user management and screen navigation in HPE Private Cloud AI familiar.

At the top management level, the GreenLake administrator can access features related to observability, resource management, system updates, and user management.

At its next layer of access, HPE Private Cloud AI includes existing role frameworks and portal access for two roles specific to AI: AI administrator and AI developer. The AI administrator can access and configure additional AI features via HPE AI Essentials, including configuring new data sources, importing AI tools and frameworks, and observing cluster resources.[42]

HPE AI Essentials software "delivers a ready to run set of curated AI and data foundation tools with a unified control plane."[43] It also includes open-source tools, such as Hugging Face. In the demonstration we experienced, we observed that using HPE AI Essentials offered several practical benefits:

- Additional screens the AI administrator can use to manage resources
- Views of GPU usage and GPU-per-application allocation and QoS
- Data source configuration screens for structured data sources
- Object store data sources and data volumes
- The ability to manage additional frameworks and user permissions

> HPE Private Cloud AI includes tight integrations with NVIDIA AI Enterprise inside HPE AI Essentials.

Delving further into the software stack, the AI developer role has immediate access to Jupyter® notebooks, data pipeline tools such as Apache Airflow, query editor tools to immediately query data, and the ability to import AI frameworks. HPE Private Cloud AI also incorporates other third-party tools into its HPE AI Essentials software, including EzPresto, Airflow, and Superset for data engineering; Apache Spark for data analytics; and Kubeflow, Feast, MLflow, and Ray for data science tasks. By incorporating familiar tools and keeping them up to date,[44] HPE Private Cloud AI provides a helpful environment for AI engineers and ongoing maintenance.[45]

HPE Private Cloud AI also includes tight integrations with NVIDIA AI Enterprise inside HPE AI Essentials. The integration includes an assortment of pre-built containers and microservices, to help optimize retrieval and fine-tuning, among other AI tasks. According to HPE, this integration "simplifies AI model fine-tuning and deployment for fast and efficient development of AI applications."[46]

Additionally, HPE Private Cloud AI includes pre-trained models, such as Llama 3,[47] to quickly deploy AI applications such as a virtual assistant from highly customizable templates in incorporated Jupyter notebooks.[48] Users can also import other publicly available models or custom Kubernetes applications and frameworks using the AI Essentials UI or API.[48] This allows engineers to import their preferred K8s-containerized tools and applications and bring them into the HPE Private Cloud AI environment.

Together, the HPE Private Cloud AI management tools have the potential to greatly reduce the time IT staff spends on routine maintenance, freeing them to devote a greater portion of their time to innovation.

To learn more about getting started with HPE Private Cloud AI, email pcai.accountdev@hpe.com.

Conclusion

In this paper, we have examined the considerations around public cloud versus on-premises for AI, and then contrasted three on-premises approaches available to buyers today: build-your-own, reference architecture + services, and turnkey.

On-premises solutions are a strong option worth considering for deploying your AI workloads, especially if you want complete control over sensitive data. Compared to the public cloud, an on-premises solution grants greater control over data protection and location, with the potential to increase the security and the performance of your AI applications.

Many organizations, however, also want to give their data services teams a more complete, cloud-like experience while maintaining this security and control. While a validated reference architecture + services approach such as Dell AI Factory simplifies deployment by offering validated components, it leaves management, integration, and orchestration to the customer, that company's services group, or a third party. In contrast, the turnkey HPE Private Cloud AI provides a pre-validated stack with hardware and AI toolsets, integrated data services software, and a cloud delivery model for administrators, data engineers, and data scientists—all of which lets organizations hit the ground running. It combines ease of procurement, setup, and management with a rich selection of out-of-box AI software and data tools, while providing the control and security of an on-premises deployment. For organizations seeking an on-premise AI implementation that comes with a built-in private-cloud experience, HPE Private Cloud AI is a stronger choice than build-your-own or reference architecture + services solutions.

---

1. Hewlett Packard Enterprise, Deploy Infrastructure in Minutes with HPE Private Cloud AI, YouTube video, accessed April 15, 2025, https://www.youtube.com/watch?v=VESIizULnkU&t=28s.
2. Apple Security Research, "Private Cloud Compute: A new frontier for AI privacy in the cloud," accessed April 15, 2025, https://security.apple.com/blog/private-cloud-compute/.
3. Apple Security Research, "Private Cloud Compute: A new frontier for AI privacy in the cloud."

4. Stuart E. Madnick, "The Continued Threat to Personal Data Key Factors Behind the 2023 Increase," accessed April 15, 2025, https://www.apple.com/newsroom/pdfs/The-Continued-Threat-to-Personal-Data-Key-Factors-Behind-the-2023-Increase.pdf.
5. Stuart E. Madnick, "The Continued Threat to Personal Data Key Factors Behind the 2023 Increase."
6. GDPR.EU, "GDPR compliance checklist for US companies," accessed April 15, 2025, https://gdpr.eu/compliance-checklist-us-companies/.
7. McKinsey, "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," accessed March 10, 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.
8. Hewlett Packard Enterprise, Deploy Infrastructure in Minutes with HPE Private Cloud AI, YouTube video, accessed April 15, 2025, https://www.youtube.com/watch?v=VESIizULnkU&t=28s.
9. CIOdive, "Enterprises struggle to show the value of AI projects," accessed April 1, 2025, https://www.ciodive.com/news/generative-ai-adoption-barrier-project-value/716504/.
10. Dell Technologies, "How Dell Makes the AI Factory Real," Direct2Dell Blog, accessed March 10, 2025, https://www.dell.com/en-us/blog/how-dell-makes-the-ai-factory-real/.
11. CIO Staff, "Dell Shares Its Vision of the AI Factory, Powered by Nvidia," CIO, accessed March 10, 2025, https://www.cio.com/article/3610094/dell-shares-its-vision-of-the-ai-factory-powered-by-nvidia.html.
12. Dell Technologies, "Your Way to AI," accessed March 10, 2025, https://www.delltechnologies.com/asset/en-us/solutions/business-solutions/briefs-summaries/dell-ai-factory-with-nvidia-ebook.pdf.
13. Dell Technologies, "Dell Offers Complete NVIDIA-Powered AI Factory Solutions to Help Global Enterprises Accelerate AI Adoption," accessed April 15, 2025, https://investors.delltechnologies.com/news-releases/news-release-details/dell-offers-complete-nvidia-powered-ai-factory-solutions-help.
14. Dell Technologies, "Dell Offers Complete NVIDIA-Powered AI Factory Solutions to Help Global Enterprises Accelerate AI Adoption."
15. Dell Technologies, "Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption," accessed April 15, 2025, https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2024~05~20240520-dell-technologies-expands-dell-ai-factory-with-nvidia-to-turbocharge-ai-adoption.htm.
16. Dell Technologies, "Dell Offers Complete NVIDIA-Powered AI Factory Solutions to Help Global Enterprises Accelerate AI Adoption," accessed April 15, 2025, https://investors.delltechnologies.com/news-releases/news-release-details/dell-offers-complete-nvidia-powered-ai-factory-solutions-help.
17. Dell Technologies, "Dell Offers Complete NVIDIA-Powered AI Factory Solutions to Help Global Enterprises Accelerate AI Adoption."
18. Dell Technologies, "Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption," accessed April 15, 2025, https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2024~05~20240520-dell-technologies-expands-dell-ai-factory-with-nvidia-to-turbocharge-ai-adoption.htm.
19. Dell Technologies, "Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption."
20. Dell Technologies, "Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption."
21. Dell Technologies, "How Dell Makes the AI Factory Real," Direct2Dell Blog, accessed March 10, 2025, https://www.dell.com/en-us/blog/how-dell-makes-the-ai-factory-real/.
22. Veronica Thums, "Dell AI Factory with NVIDIA at SIGGRAPH," accessed March 11, 2025, https://www.dell.com/en-us/blog/dell-ai-factory-with-nvidia-at-siggraph/.
23. Dell Technologies, "Dell Technologies Accelerates Enterprise AI Innovation from PC to Data Center with NVIDIA," accessed April 1, 2025, https://investors.delltechnologies.com/news-releases/news-release-details/dell-technologies-accelerates-enterprise-ai-innovation-pc-data.
24. Dell Technologies, "Your way to AI," accessed April 1, 2025, https://www.delltechnologies.com/asset/en-us/solutions/business-solutions/briefs-summaries/dell-ai-factory-with-nvidia-ebook.pdf.
25. Dell Technologies, "Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption," accessed April 15, 2025, https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2024~05~20240520-dell-technologies-expands-dell-ai-factory-with-nvidia-to-turbocharge-ai-adoption.htm.

26. Dell Technologies, "Accelerate your innovation with the Dell AI Factory," accessed March 11, 2025, https://www.delltechnologies.com/asset/en-us/solutions/business-solutions/briefs-summaries/dell-ai-factory-infographic.pdf.
27. Dell Technologies, "Dell collaborates with NVIDIA to Supercharge AI efficiency," accessed March 11, 2025, https://www.dell.com/en-us/blog/dell-collaborates-with-nvidia-to-supercharge-ai-efficiency/.
28. Dell Technologies, "Dell AI Factory Advancements Ease Enterprise AI Adoption," accessed March 11, 2025, https://www.dell.com/en-us/blog/dell-ai-factory-advancements-ease-enterprise-ai-adoption/.
29. Hewlett Packard Enterprise, Unlocking Private AI Power: Insurance Fraud Detection and Beyond, YouTube video, accessed April 15, 2025, https://www.youtube.com/watch?v=BYzF2Twg8UY.
30. Hewlett Packard Enterprise, Deploy Infrastructure in Minutes with HPE Private Cloud AI, YouTube video, accessed April 15, 2025, https://youtu.be/VESIizULnkU?t=28.
31. Hewlett Packard Enterprise, "HPE Private Cloud AI," accessed March 11, 2025, https://support.hpe.com/connect/s/product?language=en_US&kmpmoid=1014847366&tab=manuals.
32. Hewlett Packard Enterprise, "HPE Private Cloud AI Accelerate your AI path with a turnkey AI private cloud," accessed March 11, 2025, https://www.hpe.com/psnow/doc/a00143345enw?section=Product%20Documentation.
33. Hewlett Packard Enterprise, "HPE Private Cloud AI," accessed April 15, 2025, https://www.hpe.com/psnow/doc/PSN-1014847366WWEN?section=Product%20Documentation&softrollSection=1.
34. Hewlett Packard Enterprise, "Hewlett Packard Enterprise introduces new enterprise AI solutions with NVIDIA to accelerate time to value for generative, agentic and physical AI," accessed April 1, 2025, https://www.hpe.com/us/en/newsroom/press-release/2025/03/hewlett-packard-enterprise-introduces-new-enterprise-ai-solutions-with-nvidia-to-accelerate-time-to-value-for-generative-agentic-and-physical-ai-models.html.
35. Hewlett Packard Enterprise, "HPE Private Cloud AI Data sheet," accessed April 1, 2025, https://www.hpe.com/psnow/doc/PSN1014847366WWEN.pdf.
36. Hewlett Packard Enterprise, "Accelerate enterprise AI," accessed March 11, 2025, https://www.hpe.com/psnow/doc/a00141386enw.
37. HPE, "Hewlett Packard Enterprise extends private cloud portfolio at Equinix data centers to give customers more choice and fast access to hybrid cloud," accessed April 15, 2025, https://www.hpe.com/us/en/newsroom/press-release/2023/06/hewlett-packard-enterprise-extends-private-cloud-portfolio-at-equinix-data-centers-to-give-customers-more-choice-and-fast-access-to-hybrid-cloud.html.
38. Hewlett Packard Enterprise, "Accelerate enterprise AI," accessed March 11, 2025, https://www.hpe.com/psnow/doc/a00141386enw.
39. Hewlett Packard Enterprise, "HPE Private Cloud AI QuickSpecs," accessed March 11, 2025, https://support.hpe.com/hpesc/public/docDisplay?docId=a50009216enw&docLocale=en_US.
40. Hewlett Packard Enterprise, "HPE Private Cloud AI – Introduction and Demo," accessed March 11, 2025, https://www.hpe.com/h22228/video-gallery/us/en/v100005056/hpe-private-cloud-ai---introduction-and-demo/video/?-jumpId=in_ResourceLibrary&lang=en.
41. Hewlett Packard Enterprise, "HPE Private Cloud AI User Guide," accessed March 11, 2025, https://support.hpe.com/hpesc/public/docDisplay?docLocale=en_US&docId=sd00005025en_us.
42. Hewlett Packard Enterprise, "Hewlett Packard Enterprise and NVIDIA announce 'NVIDIA AI Computing by HPE' to accelerate generative AI industrial revolution," accessed April 15, 2025, https://www.hpe.com/us/en/newsroom/press-release/2024/06/hewlett-packard-enterprise-and-nvidia-announce-nvidia-ai-computing-by-hpe-to-accelerate-generative-ai-industrial-revolution.html.
43. Hewlett Packard Enterprise, "HPE Private Cloud AI User Guide, accessed March 11, 2025, https://support.hpe.com/hpesc/public/docDisplay?docId=sd00005025en_us&page=GUID-07D8A00E-99B7-422E-AE7B-0650FF355FE2.html.
44. Hewlett Packard Enterprise, "Simplify data analytics and AI across hybrid environments," accessed April 15, 2025, https://paths.ext.hpe.com/c/hpe-ezmeral-unified-analytics-software?x=Kep7F4&cc=us&lang=en&lb-mode=overlay&lb-height=100&lb-width=1 00&pf_route=a50007855.
45. Hewlett Packard Enterprise, "HPE AI Essentials Software 1.5.2 Documentation" accessed April 15, 2025, https://support.hpe.com/hpesc/public/docDisplay?docId=a00aie15hen_us&page=pcai/Catalogs/catalogs.html.

46. Hewlett Packard Enterprise, "HPE AI Essentials Software 1.6.x Documentation, accessed March 11, 2025, https://support.hpe.com/hpesc/public/docDisplay?docId=a00aie16hen_us&page=pcai/Catalogs/catalogs.html.

47. Hewlett Packard Enterprise, "HPE GreenLake Interactive Demo Experience," accessed April 15, 2025, https://www.hpe.com/psnow/ebook/e9add60f-3877-491c-8fd0-eefec3e98472?hf=hidden.

48. Hewlett Packard Enterprise, "Importing Frameworks and Managing the Application Lifecycle," accessed April 15, 2025, https://support.hpe.com/hpesc/public/docDisplay?docId=a00aie15hen_us&page=ManageClusters/managing-application-lifecycle.html.

This project was commissioned by HPE.

Facts matter.®