Solution overview

HPE Private Cloud AI: empowering ITOps and data scientists

Businesses everywhere recognize AI's transformative power and are deep in a strategic race to implement it for a competitive advantage. But for many, the path to effective AI deployment is fraught with challenges. They're battling significant headwinds: the dizzying speed at which AI technology evolves, its inherent complexity, and perhaps most critically, the growing chasm between the demand for highly specialized AI talent and its scarce availability.

The private cloud imperative

Addressing these challenges brings us to the pivotal question of AI deployment: where will your intelligence reside? The choice is stark: risk exposing vital data and corporate IP in the public cloud or secure it within a private environment designed for control and compliance. The shift towards private cloud for AI is not merely a trend, but a rapid acceleration driven by customers acutely aware of the data sovereignty concerns and the often unpredictable, escalating costs tied to public cloud. Yet, the nature of your private cloud — its architecture and capabilities — is what truly defines its success.

One of the keys to a successful private cloud strategy lies squarely in the underlying infrastructure. It's the critical foundation for any successful AI initiative, providing the indispensable components necessary to develop, train, deploy, and scale AI effectively. Consider the cost of spending a year or more building your AI infrastructure from scratch or reference architecture. Such an undertaking becomes a significant bottleneck, directly delaying the integration of the software layer required for tangible AI outcomes. In essence, any delay in establishing your AI infrastructure translates to a slow, or completely stalled, path to impactful AI.

Accelerate AI deployment, decrease complexity

Want a simpler and faster AI deployment? HPE Private Cloud AI, co-developed with NVIDIA, is your answer. This end-to-end, integrated platform is designed to get your AI private cloud up and running in just one day. It's a factory-integrated solution of hardware, software, and services that provides everything you need for popular AI use cases. Post-installation, HPE refreshes the solution with the latest GPU, CPU, and hardware innovations. One-click upgrades ensures customers have access to the latest AI software, tools, and models.

AI solutions for every stage and scale

Every business is at a different stage of AI adoption. That's why HPE Private Cloud AI isn't a one-size-fits-all solution. The solution is deployed on one of four scalable starting points based on your business needs. From development to enterprise-wide production use cases across inference, retrieval augmented generation (RAG), and fine-tuning models. With HPE Private Cloud AI, you can confidently start your AI journey, knowing that HPE provides investment protection as AI technology roadmaps evolve.

Empowering ITOps

For ITOps and cloud admins, managing HPE Private Cloud AI transforms your role, letting you move from AI zero to AI hero. Through a unified interface, you gain complete control: easily define clusters, add users and their roles, set access controls, establish robust security, monitor resource allocation, and seamlessly expand, observe, and manage multiple HPE Private Cloud AI clusters.

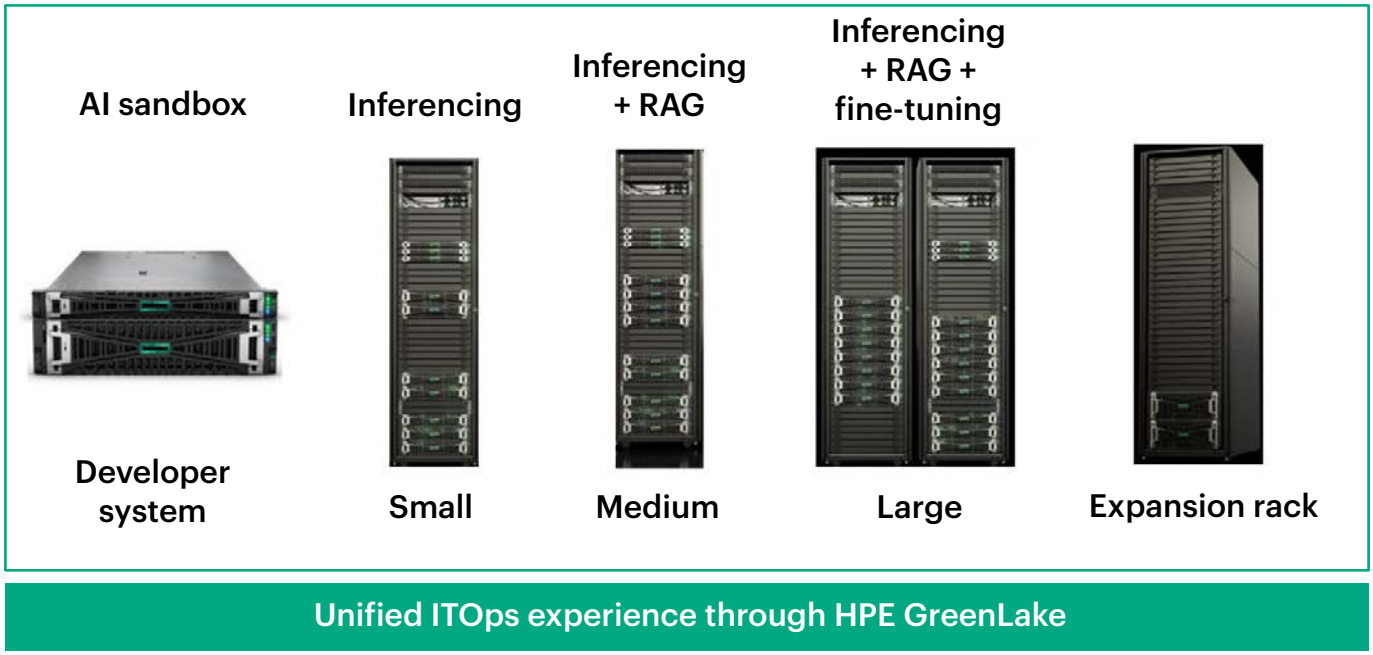

Figure 1. HPE Private Cloud AI: the right size for every stage of your AI journey

The power behind this ease lies in automated back-end processes. These intelligent systems set up configurations to your exact specifications, even at the complex Kubernetes layer, without you needing to be an expert. Once in the HPE GreenLake cloud, simply define your policies, which will be instantly propagated to the built-in software layer, HPE AI Essentials Software. This process establishes platform-wide configurations, reducing the tedious, application-by-application setup for each user and enabling consistency across your entire AI environment.

Air-gapped AI for ultimate security

For customers operating in dark or highly secure environments, HPE Private Cloud AI is available in an air-gapped version. Your sensitive data remains fully protected, allowing for model customization and inference in an environment completely isolated from external networks.

Like its connected counterpart, the air-gapped version includes robust security and compliance features. This makes it an ideal solution for organizations facing the strictest regulatory and sovereignty requirements, such as those in the military, government, healthcare, and financial sectors.

Heart of AI innovation

At the heart of HPE Private Cloud AI lies HPE AI Essentials Software, an integrated powerhouse designed to accelerate your AI journey. This software provides cutting-edge, ready-to-use AI and open-source tools, empowering AI practitioners to construct AI workloads at unprecedented speed. With HPE AI Essentials Software, diverse types of AI practitioners get seamless access to built-in data, AI tools and models, and low-code wizards. This enables them to effortlessly create, deploy, and govern ML pipelines, and manage models. Best of all, HPE handles maintenance and updates automatically, freeing your teams to focus on innovation and delivering tangible AI outcomes instead of wrangling with complex updates.

The built-in model catalog provides persona-based access to your preferred tools, including NVIDIA® NIM™ as part of NVIDIA AI Enterprise. Can't find a specific tool you prefer or have a trusted third-party application? Our intuitive wizard helps you import Helm-based tools, apps, or NVIDIA Blueprints in just a few steps. Once configured into the platform, these imported applications are covered under the platform's robust security and governance framework.

Bring AI to your data

AI is a data-hungry application, and data is the key to delivering reliable outcomes. It's the fuel that powers AI, enabling models to learn, adapt, and make intelligent decisions. The quality and relevance of this data are paramount, as flawed or insufficient data inevitably leads to flawed AI results.

HPE AI Essentials Software includes a built-in data lakehouse gateway that provides secure and transparent access for authorized users to your corporate data, regardless of format or location. A global namespace connects operational databases across S3 and NFS endpoints directly to the data lakehouse in near real-time. This capability removes the need to move data before processing, effectively providing users with a consistent, unified view of data across diverse sources.

Fine-grained access controls govern this unified access, providing data security while enabling seamless collaboration. It also allows customers to utilize their preferred analytics engines — whether that's Microsoft SQL, Apache Spark, BI tools, or AI/ML tools — to work together on data stored across various architectures.

Enterprise multi-tenancy

HPE AI Essentials Software enhances its collaborative capabilities with the introduction of robust enterprise multi-tenancy. This advancement allows multiple users to collaborate seamlessly on projects within a secure, shared environment. Administrators gain granular control over roles and permissions, enabling them to define

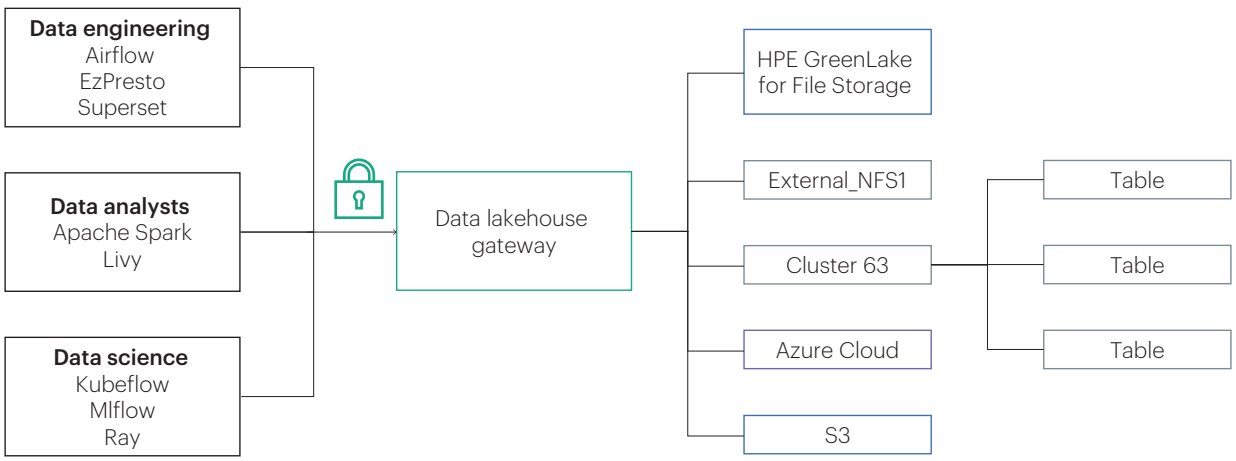

Figure 2. Data lakehouse gateway enables unified, governed access to data from diverse sources

who can access what, and precisely specify data source access based on individual projects. Furthermore, with GPU quota setting capabilities, resources can be efficiently allocated and managed across different teams and projects, enabling optimal performance and fair usage for users. This multi-tenancy model empowers teams to accelerate their AI initiatives with greater efficiency, security, and collaborative capabilities.

Deliver AI value quickly

Beyond seamless data access and multi-tenancy, data engineers and data scientists are empowered with a pre-integrated catalog of AI and data tools. This reduces infrastructure complexities, allowing them to innovate and deploy solutions rapidly. Beyond the pre-integrated offerings, data practitioners gain full flexibility. They can seamlessly import and integrate models and tools from external repositories, such as Hugging Face, as well as NVIDIA Blueprints, directly into the catalog.

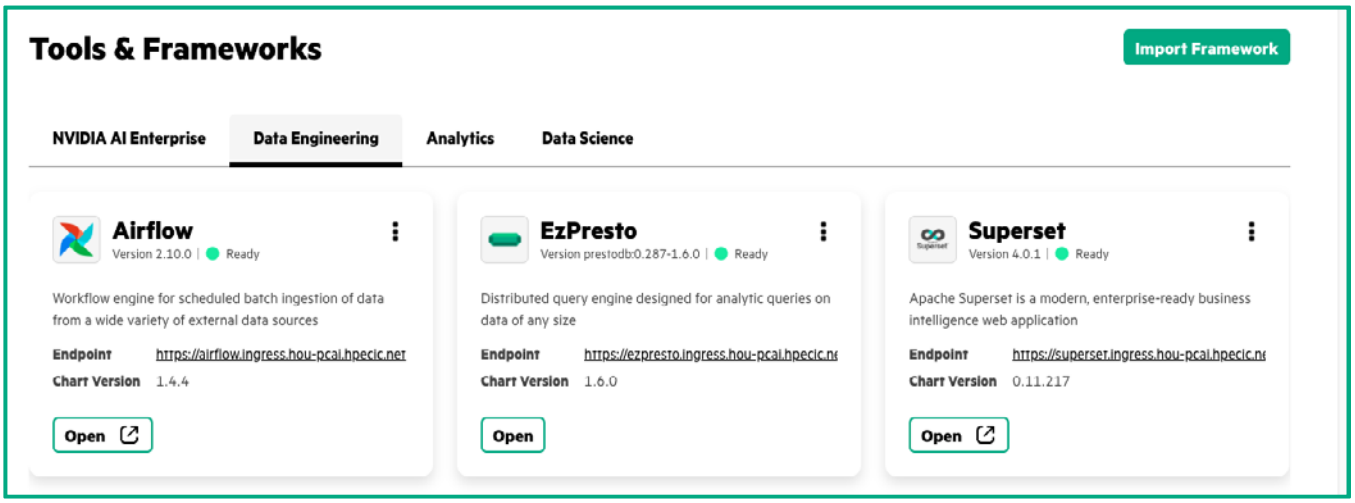

Figure 3. Data engineers benefit from instant access to Airflow for orchestration, EzPresto for querying, and Superset for visualization of data pipelines

Say goodbye to complex data transformation woes. The model catalog simplifies everything with pre-validated hardware and software stacks that enable seamless interoperability, making integration challenges a thing of the past. Data engineers can directly access data in the data lakehouse, bypassing tedious pre-ingestion or data movement. This newfound freedom lets them dedicate their energy to crafting robust pipelines, unburdened by infrastructure complexities and data silos.

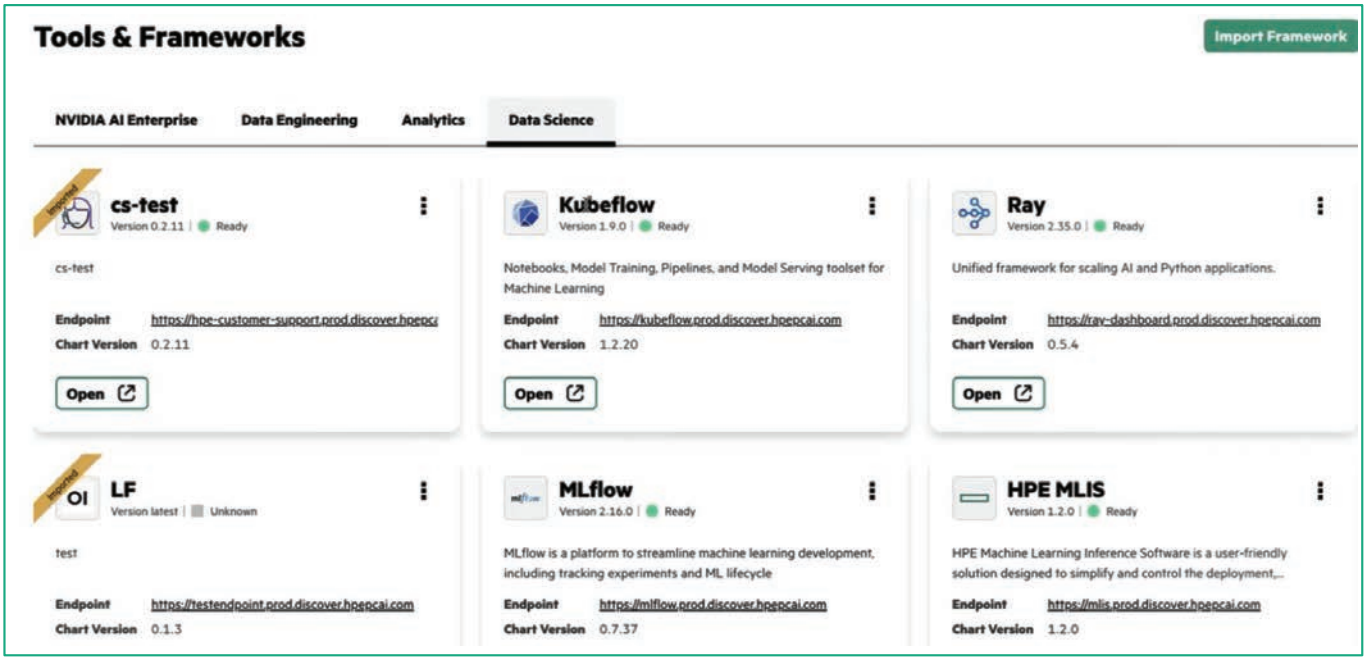

Figure 4. Out-of-the-box model catalog for data scientists. Items with a yellow ribbon have been imported into HPE AI Essentials Software

A simplified infrastructure also sets data scientists free to experiment. Pre-integrated tools let them focus squarely on advanced model development, insightful experimentation, and delivering tangible AI value. Customizable Jupyter notebooks and JupyterLab environments provide the exact control needed to create, develop, run, and deploy applications precisely to their specifications.

Rich ISV ecosystem

HPE AI Essentials Software delivers exceptional value through its rich ecosystem of validated ISV partners. This extensive network complements the platform by providing a diverse array of pre-validated solutions and applications. Customers can rapidly import and deploy dozens of pre-validated use cases, accelerating their time to value and achieving faster ROI with significantly less risk. This rich ecosystem, combined with features like robust multi-tenancy, enables organizations to access the specialized tools and collaborative environment needed to tackle complex AI challenges and unlock new possibilities confidently.

True enterprise-ready AI platform

Can HPE Private Cloud AI accelerate your AI deployments? Here are some key questions to explore.

1. What specific steps in the AI deployment process lead to the most significant delays or resource drain for your team?

2. Have you been able to quantify the time and resources spent on integrating and maintaining your AI infrastructure and software stack?

3. Could you describe any challenges or delays your team has experienced in the past when trying to launch AI applications due to underlying infrastructure limitations?

4. How does the ability to view and control your AI's underlying components impact your team's ability to debug, optimize, or scale your models?

Accelerate your AI deployment and increase ROI

Tired of the complexity and delays of piecing together hardware, software, and orchestration for AI? HPE Private Cloud AI cuts through that complexity. Industry analysts validate that our co-engineered solution with NVIDIA can save you seven to twelve months in deployment time compared to build-it-yourself initiatives. This isn't just about speed; it also significantly lowers your total cost of ownership by an estimated $1.72 million to $3.57 million, depending on the size of your system.[1]

If efficiency and rapid time to value are crucial for your private cloud, you'll appreciate HPE Private Cloud AI. Its turnkey approach was ranked as the simplest and fastest to value against both reference architectures and self-built alternatives.[2]

It's ready to run, so you can stop integrating and start innovating.

[1] "The economic benefits of HPE Private Cloud AI with NVIDIA AI Computing by HPE", Enterprise Strategy Group, March 2025.

[2] "On-premises AI approaches: the advantages of a turnkey solution, HPE Private Cloud AI", Principled Technologies, April 2025.

a00141386ENW

HEWLETT PACKARD ENTERPRISE

3299471