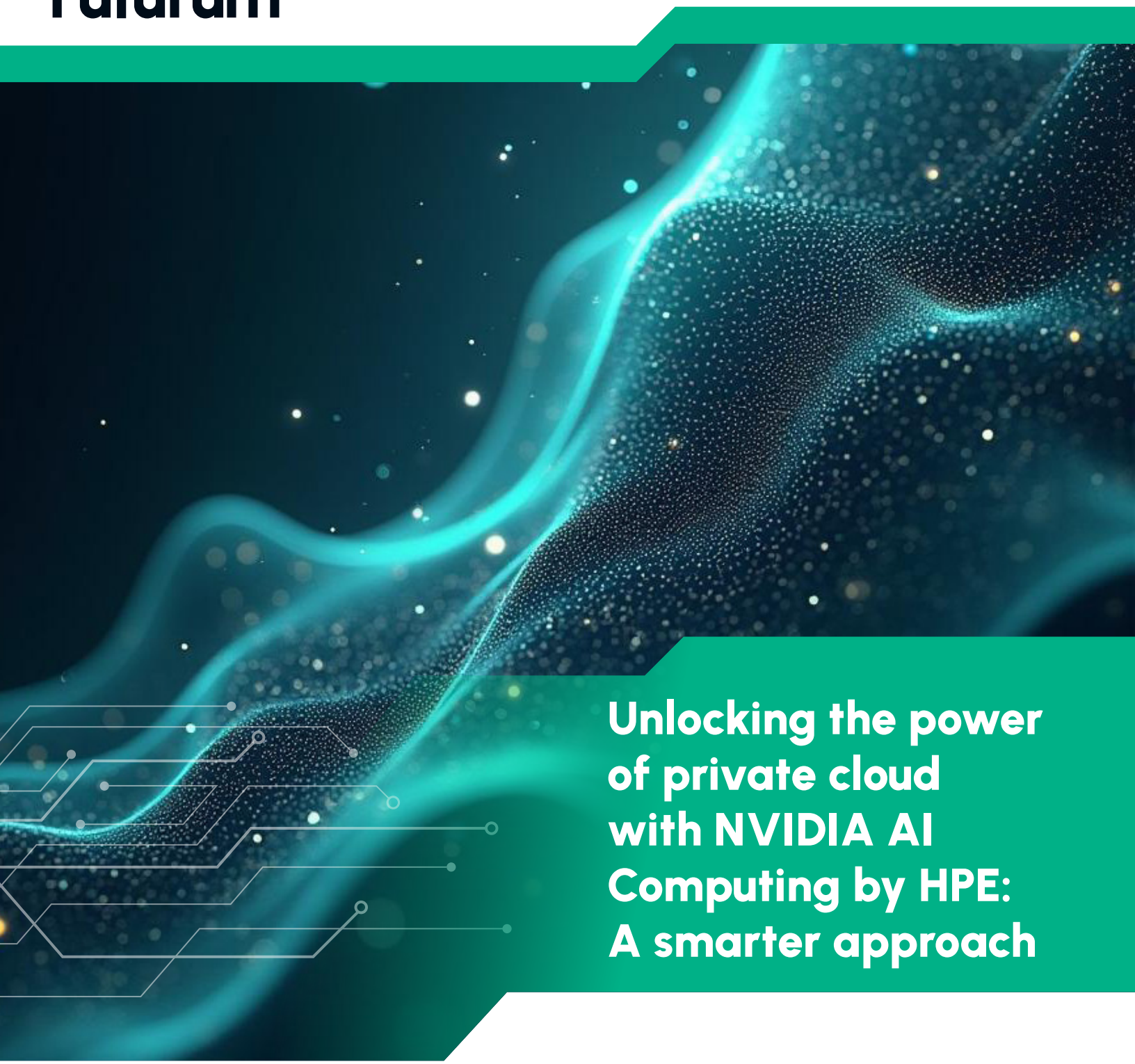
Futurum



Unlocking the power of private cloud with NVIDIA AI Computing by HPE: A smarter approach

AUTHORS

Nick Patience
VP & Practice Lead, AI | The Futurum Group

Ron Westfall
Research Director | The Futurum Group

JANUARY 2025

IN PARTNERSHIP WITH

Hewlett Packard Enterprise | NVIDIA

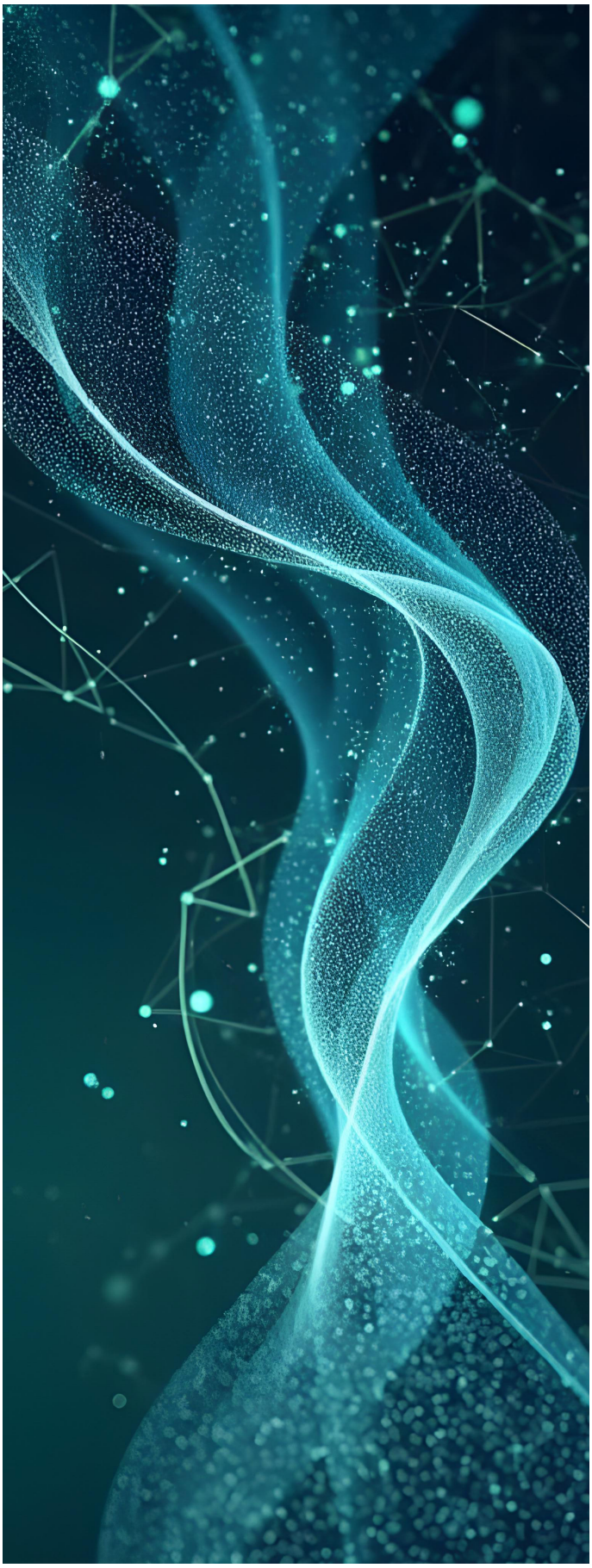
Executive Summary:

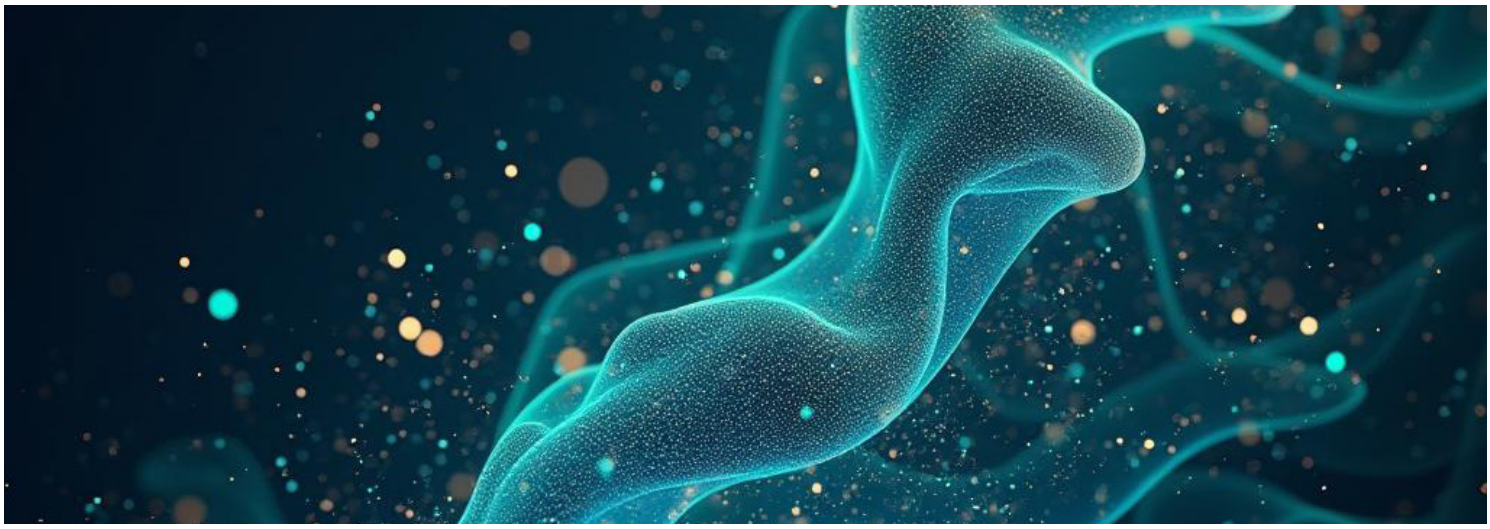
What if you could predict market trends with unparalleled accuracy? Or develop personalized customer experiences at scale? Or automate complex processes, freeing up your workforce for higher-value tasks? The answer lies in the transformative power of artificial intelligence (AI). As AI continues to mature, enterprises are increasingly recognizing its potential to transform their operations and business models.

While public cloud platforms offer a convenient and cost-effective way to deploy AI solutions, they may not always be the optimal choice for every enterprise. As AI initiatives mature and organizations handle increasingly sensitive data, the need for greater control, security, and customization becomes paramount. Private cloud emerges as a compelling solution, providing a dedicated and secure environment to power AI innovation.

As organizations embark on their AI journeys, a critical decision arises: should they build their own private cloud or leverage a commercial solution? This choice hinges on various factors, including the organization's specific needs, budget, technical expertise, and risk tolerance.

While your strong engineering team and familiarity with open-source tools might suggest a straightforward path to building a private cloud, the reality is more complex. This paper explores these complexities and helps establish why selecting a private cloud can prove the optimal approach for optimizing the AI journey across the entire organization.





Section 1: Why You Do Not Want to Build Private Cloud Yourself

The future is now for harnessing AI's full potential across the entire organization that uses a scalable, secure, and intelligent foundation. There is the impulse to view the evolving advancements in AI capabilities, the ongoing trend towards on-premises solutions, and increased IT familiarity with open-source tools to adopt a do-it-yourself (DIY) approach to AI implementation.

However, these considerations do not provide a warrant for creating your own private cloud infrastructure, as it can prove an impractical and costly decision. The countless complexities and resources required for such an endeavor often outweigh the perceived benefits, especially when considering the robust and scalable private cloud solutions already available in the market. These considerations include:

Substantial and Hidden Costs

Building and maintaining a private cloud requires substantial upfront investments in hardware, software, and infrastructure. While these costs may initially seem manageable, hidden expenses can quickly accumulate, including the ongoing costs of managing hybrid data, labeling and cleansing data, training models, and maintaining the overall solution. Moreover, once the initial infrastructure is deployed, ongoing maintenance becomes critical.

Complexity of Data Management

The realization of AI's full potential is hindered by the complexities of data management in hybrid environments. Siloed data and model bias can impede innovation and limit the scope of insights. A common approach to address these challenges is to build a DIY private cloud. While this offers greater control and customization, it introduces significant challenges in data management, security, and scalability. Additionally, the rapidly evolving AI landscape necessitates the integration of multiple open-source and proprietary solutions, which can be both time-consuming and resource-intensive to maintain.

Security Oversight and Unknowns

Security is paramount for today's digital enterprise. When deploying a DIY private cloud, security must extend beyond infrastructure to encompass data and AI models to protect sensitive data as well as prevent the introduction of corrupted or biased data into training datasets. A real-world example of the need for this type of security occurred at a major telecom company when employees inadvertently input confidential data into the company's chatbot releasing it out into the public.

To address these challenges, organizations must adopt a multifaceted approach that combines robust data protection, advanced security technologies, governance and control. This approach should be seamlessly integrated into the data science workflows without hindering productivity. Delivering this coverage in a DIY architecture requires a dedicated team of IT, security, and data science professionals. The cost of these specialized resources can be high, but the stakes are even higher as compromised AI models can lead to inaccurate predictions, biased outputs, ethical fallout, and potential regulatory penalties.

Scalibility Challenges

Based on our client interactions, a phased approach to AI and GenAI implementation is gaining traction. This approach allows organizations to start with smaller-scale projects, gradually scaling as their AI maturity and business needs evolve. By enabling controlled experimentation and focused learning, this strategy facilitates adaptation to the dynamic AI landscape while optimizing ROI.

To support this phased approach, organizations need flexible infrastructure that can scale with their evolving needs. A DIY private cloud, while offering control, requires significant upfront investment, ongoing maintenance and specialized expertise. This can hinder a phased approach to AI implementation, diverting resources from core business objectives and slowing down innovation.

Overall, the rapid evolution of software, the constant need for security patches, and the rising complexity of AI technologies demand significant operational resources. Maintaining system security, ensuring optimal performance, and staying abreast of the latest updates require continuous IT monitoring and expert technical support. These ongoing costs can easily surpass initial set-up expenses. Any downtime during maintenance or failure to apply timely updates can lead to serious consequences, such as system vulnerabilities and reduced operational efficiency. Businesses must carefully consider these hidden costs when evaluating the total cost of ownership for a DIY private cloud.

The Innovation Gap: A DIY Cloud Dilemma

Today's hybrid enterprise empowers organizations to optimize AI workload performance, security, and costs by shifting between public and private clouds. Advanced technologies like AI, machine learning, and predictive analysis are often readily available on public cloud platforms, accelerating innovation and reducing time-to-market. In contrast, building and maintaining these capabilities in-house on a DIY private cloud requires significant upfront investment and ongoing maintenance, potentially hindering innovation and competitive advantages.

Section 2: Why HPE Private Cloud AI Provides Enduring Advantages over DIY

HPE Private Cloud AI ensures that enterprises can avoid the substantial hidden costs, complexities, and security challenges associated with DIY implementations. There are numerous benefits that HPE Private Cloud AI delivers to enterprises, starting with enabling instant AI productivity and accelerating time to value for customers. Self-serve access to essential AI tools is capable of speeding developer activity by up to 90%. Plus, through NVIDIA AI Computing by HPE, enterprises can swiftly transition from AI pilot projects to full-scale production with a ready-to-use AI private cloud solution.

Unlike DIY private clouds, HPE Private Cloud AI can be up within hours, allowing customers to realize value immediately. From our view, these claims are warranted due to the following considerations:

- HPE Private Cloud AI is a comprehensive, turnkey solution that integrates server, storage, networking, and AI software components. This fully configured system can be swiftly deployed in a customer's data center and made operational within 8 hours, based on current observations. The package includes HPE GreenLake cloud, providing a complete, functional environment that can enable customers to immediately begin building and deploying AI workloads. As a result, once connected, everything can be cloud managed.

- The solution targets three distinct personas consisting of cloud administrators, AI administrators, and AI developers. Each of these roles' experiences unique interactions within the product, tailored to their specific needs and functions, augmenting the user experience for each unique persona.

- HPE Private Cloud AI is tailored to assist users with varying levels of AI expertise, from those well-versed in the technology to complete novices. This design ensures that all users of systems from HPE GreenLake manage the entire lifecycle for cloud administrators, including aspects such as management, expansion, telemetry, observability, and metrics. This solution simplifies operations by providing a centralized and automated approach to server lifecycle management.

- From a single dashboard, the cloud administrator can establish granular access rights to tools, models, data, and data sources. Also, zero-trust, end-to-end security and compliance that spans from the HPE GreenLake platform through individual apps and applications on HPE Private Cloud AI.

- For reliable software update support, HPE updates all interconnected components on a regular basis and pushes the updates out as a comprehensive software update driven by the HPE GreenLake platform.

- This means that AI professionals have access to the latest AI tools and frameworks. With DIY systems, all these updates must be performed, integrated, and tested to ensure nothing within the system breaks.

Taken together the ROI benefits are potentially substantial. The level of effort in hardware installation and setup can prove immense. For instance, the physical setup of compute resources can take up to 20 hours for a 2-server rack. Setting up the networking backplane, GPU hosts, and RDMA components can consume up to 800 IT team hours. Implementing security and role-based access to data and models alone can require up to 14 weeks in small environments with experienced staff as well as up to 9 months in large, complex environments with experienced staff.

Moreover, from our perspective, HPE Private Cloud AI strategically positions enterprises to support the topmost identified AI application use cases. At 15.7%, Conversational AI is the biggest use case across 2024, closely followed by Development Tools at 15.6% (according to Futurum Intelligence). The AI software/tools market is projected to reach a value of $238B by 2029 with a CAGR of 13.4%, fueled by advancements in generative AI technology, autonomous systems, and predictive analytics (according to Futurum Intelligence). With HPE Private Cloud AI, enterprises can take full advantage of both popular AI application use cases and the use cases that uniquely fulfill improving business outcomes.

HPE Private Cloud AI Ensures Private and Secure Data

The HPE solution provides enterprise-grade confidence, governance, and control by offering the private cloud capabilities, key to managing, securing, and governing the data, models, and infrastructure integral to AI workload optimization. Through an embedded data lakehouse, the solution enables secure access to data across the enterprise in a single global namespace. This is integral to providing a cloud experience that ensures data remains private and secure by keeping it on-premises offering multi-layered controls to protect data and models, assuring reliability and performance.

HPE Private Cloud AI Delivers Flexible and Modular Approach

Deployed on-premises yet designed for hybrid environments, HPE Private Cloud AI offers a flexible and modular approach that leverages cloud technologies while remaining economical and scalable to meet evolving business needs. The solution delivers infrastructure that is scalable and pretested, enabling organizations to scale and experiment with AI projects across a diverse ecosystem of AI models and development tools.

The platform is customizable to fulfill the unique requirements of businesses in different sectors, enabling organizations to optimize their infrastructure and AI models. This includes providing a variety of modular solutions specifically designed for different workloads, encompassing compute, storage, networking, virtualization, and specialized applications such as ML operations. As a result, HPE Private Cloud AI ensures that enterprise customers can flexibly scale their AI workloads and capabilities to lock-in access to advanced capabilities, such as real-time analytics and proactive maintenance.

NVIDIA Partnership Provides Unique Advantages

From our perspective, HPE's unique partnership with NVIDIA in co-developing this solution takes full advantage of integration with NVIDIA AI Enterprise software. This speeds up data science pipelines and streamlines development and deployment of production-grade copilots and Gen AI applications, including NVIDIA NIM inference microservices.

The partnership goes beyond just technology, providing collaborative enablement and training programs along with shared market development funds for partners. By offering a full-stack AI infrastructure, the collaboration provides a complete AI-native solution, combining HPE's compute, storage, and cloud capabilities with NVIDIA's AI computing, networking, and software.

HPE Opsramp Assures Unified Visibility Benefits

HPE Opsramp provides complete visibility, insight, and control access throughout an enterprise's hybrid IT estate, extending AI-driven, full-stack observability and intelligent automation to HPE Private Cloud AI. Opsramp AI infrastructure observability and copilot assistant features make sure IT operations are integrated with HPE GreenLake cloud to deliver observability and AIOps to all HPE products and services.

Section 3: Conclusions & Recommendations

In conclusion, while a DIY private cloud may appear at first blush to offer control and cost savings, the hidden costs, complexity, and limitations often make it a less attractive option for businesses. The substantial maintenance, security oversight, and scaling challenges can overwhelm organizations, ultimately leading to higher expenses and diminished outcomes. As a result, we believe organizations need to consider the following recommendations:

- Prioritize HPE's Demonstrated Leadership: IT decision-makers should assign top consideration to HPE Private Cloud AI since the solution demonstrates HPE's leadership in driving digital transformation across enterprises through innovations in AI, hybrid cloud, and mission-critical workloads that DIY private cloud is unable to provide comprehensively or consistently.

- Call to Action: IT decision-makers need to consider HPE Private Cloud AI across their evaluation process in adopting their organization-wide AI solutions in private cloud environments due to the enduring competitive advantages of HPE's portfolio over DIY private cloud, including achieving a secure, AI-native network, purpose-built with AI and for AI.

Important Information About this Report

CONTRIBUTORS

Nick Patience
VP & Practice Lead, AI | The Futurum Group

Ron Westfall
Research Director | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Hewlett Packard Enterprise

ABOUT HPE

HPE combines technology insights, financial expertise, and a deep rooted focus on sustainability to create smarter IT lifecycles for customers and partners of all sizes. Working across the entire tech estate, from edge to cloud to end-user, our collaborative approach delivers asset management solutions that not only free up capital and maximize capacity, but also advance sustainable practices globally and consistently. For more information, visit: hpe.com

NVIDIA

ABOUT NVIDIA

NVIDIA is the world leader in accelerated computing.

Futurum

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum

CONTACT INFORMATION

The Futurum Group LLC | futurumgroup.com | (833) 722-5337 |