The Futurum Group

Unlocking AI Potential: How HPE Private Cloud AI Accelerates AI Deployment and Innovation

AUTHOR

Steven Dickens
Vice President and Practice Lead | The Futurum Group

IN PARTNERSHIP WITH

Hewlett Packard Enterprise

NVIDIA

JUNE 2024

Executive Summary

The transformative potential of generative AI (GenAI) has sparked significant interest and growth across various industries. In 2023 alone, the number of public GenAI projects on GitHub surged by 248%, highlighting the increasing adoption of AI technologies. However, this rapid growth also brings forth several challenges. Enterprises face a complex technological landscape and often struggle with the transition from pilot to production in their AI projects. And IP protection is a top-of-mind issue for enterprises who are concerned about exposing proprietary data to public models. HPE, in collaboration with NVIDIA, addresses these challenges by offering scalable, secure, and performance-optimized private cloud AI solutions. These solutions enable enterprises to harness the power of AI while maintaining control over their data and infrastructure, thus accelerating their journey towards AI-driven innovation.

Overview of the HPE Private Cloud AI Solution

HPE Private Cloud AI, part of the NVIDIA AI Computing by HPE portfolio, is a turnkey, scalable, and AI-optimized private cloud designed to accelerate AI project deployment while ensuring data remains under enterprise control. The solution combines NVIDIA accelerated computing, networking and software with HPE's high-performance compute, storage   and the HPE GreenLake cloud, together with HPE capabilities for data pipelines, orchestration and MLOps. Thereby enabling enterprises of every size to gain a fast, flexible path for developing and deploying GenAI applications. Enterprises can leverage these capabilities to overcome barriers to AI adoption, ensuring faster time to value and seamless scalability to support future growth.

Statement of Purpose for the Research Brief

This research brief aims to provide an in-depth analysis of the challenges enterprises face in deploying AI workloads and how HPE Private Cloud AI effectively addresses these challenges. The brief will highlight key findings and insights into the solution's capabilities in terms of speed to value, repeatability, and scalability. Additionally, it will offer recommendations for organizations seeking to optimize their AI deployments, ensuring they can maximize the benefits of AI technologies while managing costs and maintaining security.

Key Findings and Insights

1. Speed to Value: HPE's and NVIDIA's solution significantly reduces the time to market for AI projects by offering pre-integrated and tested tools, models, and infrastructure. This approach enables enterprises to rapidly transition from pilot phases to full production, ensuring quick realization of AI benefits.

2. Repeatable Solutions: The turnkey standardized infrastructure offered by HPE Private Cloud AI enables consistent, performant, and scalable AI deployments. By providing pre-defined, integrated, and tested tools, enterprises can deploy AI workloads quickly and easily, often with just a single click.

3. Cloud experience: Designed to scale with business needs, HPE's solution offers flexible deployment options and pay-as-you-go financial models that support growth while effectively managing costs. The solution's CapEx model and manageability features further enhance scalability, allowing enterprises to expand their AI capabilities seamlessly.

Here's an updated overview reflecting the new focus on making the case for hybrid/private cloud over public cloud for AI, along with the specific challenges:

Overview of Business Challenges in AI Implementation

When considering AI implementations, enterprises face several specific challenges that often make hybrid/private cloud solutions more appealing than public cloud options:

1. Defining, Developing, and Operationalizing the Right Use Cases: Enterprises need to accurately identify AI use cases that will drive the most value. This requires a deep understanding of business needs and the ability to translate them into AI projects. Once identified, developing and operationalizing these use cases involves significant complexity, including integrating AI models into existing workflows and ensuring they deliver tangible business outcomes.

2. Maximizing Impact with Data while Protecting Intellectual Property (IP): Leveraging data effectively in AI models to generate meaningful insights is essential. However, this must be done without compromising sensitive IP. Public cloud solutions often raise concerns about data security and control, making it challenging to protect proprietary algorithms and sensitive data from unauthorized access or breaches.

3. Scaling AI Efforts without Exploding Costs: Scaling AI initiatives to meet growing demands can lead to prohibitive costs, especially when relying on public cloud resources. Enterprises must balance the need for computational power and storage with the financial implications, ensuring that AI deployments remain cost-effective as they scale.

4. Optimizing Technology for Specific Use Cases: Challenge: Different AI applications require different technologies and infrastructures. Ensuring that the technology stack is optimized for specific use cases is critical but complex. Public cloud solutions may not always offer the necessary customization or performance optimization needed for particular AI workloads.

5. Performing Inference and Tuning Near Data Sources: Bringing analytics closer to the data is crucial for reducing latency and enhancing performance. Public cloud solutions often involve data transfer across distances, leading to delays and inefficiencies. Ensuring that inference and tuning occur near the data sources can significantly improve the speed and accuracy of AI models.

These challenges highlight the importance of considering hybrid and private cloud solutions for AI implementations, as they can offer more control, customization, and cost management compared to public cloud alternatives.

Summary of Findings

HPE Private Cloud AI offers several key features and capabilities that address the major challenges faced by enterprises in deploying AI workloads:

1. Speed to Value: HPE accelerates AI development and deployment with pre-integrated, tested infrastructure and tools, reducing time to market.

2. Repeatability: The solution enables consistent, performant, and scalable AI deployments with standardized infrastructure stacks, ensuring reliability and efficiency.

3. Cloud experience: HPE's solution supports growth and flexibility with the HPE GreenLake platform and scalable infrastructure, accommodating evolving business needs.

4. Security: Comprehensive security measures ensure data protection and compliance, addressing the heightened vulnerabilities associated with AI adoption.

5. AI Governance and Management: The solution simplifies AI governance with a unified access point and control, enhancing manageability across hybrid environments.

6. Strategic Partnerships: HPE leverages collaborations with NVIDIA and system integrators to enhance capabilities and provide robust support for AI deployments.

Recommendations - Making the Case for Hybrid/Private Cloud over Public Cloud for AI

When considering AI implementations, enterprises can benefit significantly from hybrid/private cloud solutions. Accurately identifying AI use cases that drive the most value requires a deep understanding of business needs and the ability to translate them into AI projects. Developing and operationalizing these use cases involves integrating AI models into existing workflows to ensure they deliver tangible business outcomes.

Leveraging data effectively in AI models to generate meaningful insights without compromising sensitive intellectual property is crucial. By maintaining stringent control over data, enterprises can protect proprietary algorithms and sensitive information from unauthorized access or breaches, addressing common security concerns associated with public cloud solutions.

Scaling AI initiatives to meet growing demands while managing costs effectively is essential. Enterprises must balance the need for computational power and storage with financial implications, ensuring AI deployments remain cost-effective as they scale.

Optimizing the technology stack for specific AI applications is critical for success. Enterprises need to customize and fine-tune their technology to meet the needs of various AI workloads, ensuring the best-fit solutions for different use cases.

Conducting AI inference and model tuning close to data sources is another important factor. This approach reduces latency and enhances performance, ensuring that inference and tuning occur near the data sources for improved speed and accuracy of AI models.

Private cloud AI solutions offer several key advantages for enterprises w enterprises who require AI solutions deployed securely on-premises in colos, data centers, or edge locations. Retaining data proximity ensures faster access and processing, improving the efficiency and performance of AI models. Comprehensive governance simplifies management, ensures accountability, and facilitates secure, collaborative workflows. Cost control is enhanced through flexible financial models, avoiding the high expenses associated with public cloud solutions. Additionally, optimizing the technology stack for specific use cases ensures that AI deployments are tailored for maximum impact and efficiency.

Conclusion

The rapid evolution of generative AI presents both opportunities and challenges for enterprises. As organizations strive to harness the full potential of AI, they must navigate complex technological landscapes, ensure robust security, and manage scalability and costs effectively. NVIDIA AI Computing by HPE provides a comprehensive solution to these challenges, offering scalable, secure, and performance-optimized AI infrastructure. By leveraging HPE's turnkey private cloud for AI and strategic partnerships, enterprises can accelerate their AI journey, achieve faster time to market, and maintain control over their data and infrastructure. Looking forward, HPE remains committed to driving innovation and enabling enterprises to unlock the transformative potential of AI. Through continuous development and strategic collaborations, HPE aims to provide cutting-edge solutions that empower businesses to thrive in the AI-driven future.

Important Information About this Report

CONTRIBUTORS

Steven Dickens
Research Analyst | 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

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

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 Financial Services 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/hpefinancialservices

NVIDIA.

ABOUT NVIDIA

NVIDIA engineers the most advanced chips, systems, and software for the AI factories of the future. We build new AI services that help companies create their own AI factories. For more than 30 years, scientists, researchers, developers, and creators have been using NVIDIA technology to do amazing things. More than 4 million developers now create thousands of applications for accelerated computing. More than 40,000 companies use NVIDIA AI technologies, with 15,000 global startups in NVIDIA Inception.

.

TheFuturum
Group

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.

TheFuturum
Group

CONTACT INFORMATION

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