Hewlett Packard Enterprise | NVIDIA

Lösungsübersicht

Beschleunigte KI für Unternehmen

Eine sofort nutzbare KI-Private Cloud im Rahmen des NVIDIA AI Computing by HPE Portfolios

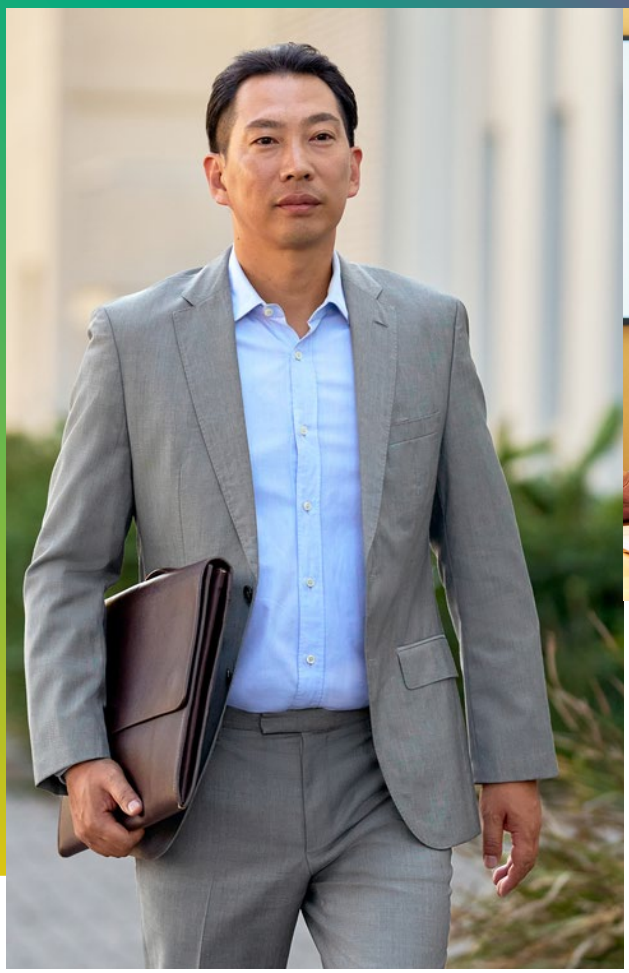


Abbildung 1. Illustration von HPE Private Cloud AI als Bestandteil von NVIDIA® AI Computing by HPE

Ihre Vorteile durch die jüngsten KI-Errungenschaften

Künstliche Intelligenz (KI) bietet ein nahezu grenzenloses Potenzial bei der Gewinnung von Einblicken, Kenntnissen und Erfahrungen, mit denen Unternehmen und unsere Gesellschaft komplett umgestaltet werden können.

Immer mehr Unternehmen führen KI ein, um bessere Unternehmensergebnisse und schnellere Innovationen zu erzielen als ihre Mitbewerber.

Der Schwerpunkt der meisten KI-Strategien in Unternehmen liegt auf zwei konkreten Bereichen: Inferenz-Workloads, die von Anwendungen mit Large Language Models (LLMs) bis hin zu branchenspezifischen Anwendungsfällen reichen, und dem Einsatz der Unternehmensdaten, um mit Techniken wie Feinabstimmung und Retrieval Augmented Generation (RAG) einen Kontext bereitzustellen. Diese Fortschritte eröffnen ein breiteres Spektrum an Möglichkeiten, mithilfe der generativen KI (GenAI), alles, was wir tun, zu beschleunigen. Viele Unternehmen experimentieren zwar mit mehreren KI-Projekten, allerdings ist die Rentabilität der einzelnen Projekte nicht immer eindeutig erkennbar. Daher benötigen Unternehmen eine einfache und risikoarme Möglichkeit, mit transformativen Technologien zu experimentieren.

Public Cloud-Lösungen bieten zwar einen großen Funktionsumfang, sie können allerdings die beteiligten Daten und Modelle Risiken mit potenziell schwerwiegenden Folgen für das geistige Eigentum aussetzen. Deshalb entscheiden sich viele Unternehmen für Private Cloud-Lösungen mit selbstgehosteten Open-Source-Modellen, um Zeit und Ressourcen zu sparen, die Datentransparenz zu verbessern und die Flexibilität zu steigern. Aber auch bei privaten Modellen müssen Unternehmen in der Lage sein, die Datennutzung zu verwalten und die Modellnutzung mitzuverfolgen bzw. zu überwachen, um sicherzustellen, dass KI-Modelle den angestrebten Nutzen erbringen und die Entstehung neuer Risiken vermeiden. Ein Fokus auf private Lösungen bietet mehr Kontrolle über KI und Sicherheit und macht die Umgebung weniger komplex. Diese Umstellung hat eine Rückkehr zur Private Cloud ausgelöst: 68 % der Unternehmen betrachten die hybride Multi-Cloud als Schlüssel für ihre GenAI-Strategie, und mehr als 50 % planen den Einsatz einer dedizierten privaten Infrastruktur.[1]

[1] „Collaboration Insights—IDC's Future Enterprise Resiliency and Spending Survey", IDC, 2023

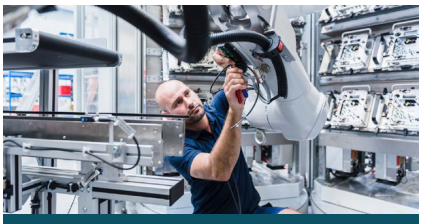
Flexibilität und das Beste der Cloud für KI

IDC prognostiziert, dass bis 2026 60 % aller Unternehmen nicht die angestrebte Leistung in ihren GenAI-Initiativen erzielen werden, weil es ihnen nicht gelingt, Verbindungen zwischen Daten, KI-Modellen und Anwendungen herzustellen.[2] Nur 10 % der KI-Projekte werden die Produktionsreife erreichen.[3] Aber warum geschieht das? Und was können Unternehmen tun, um es zu vermeiden? Es ist von entscheidender Bedeutung, Innovationen zu fördern, und Unternehmen müssen über die richtigen Tools und die nötige Infrastruktur verfügen, um die Erfolgsraten ihrer KI-Projekte zu erhöhen.

Neu aufkommende Technologien steigern die Komplexität der IT und verlangen ausgeprägte Fachkenntnisse für ihr Management. So finden zwar beispielsweise zahlreiche Innovationen im Open-Source-Bereich statt, doch die Einführung dieser Technologien im Unternehmen ist oftmals eine Mammutaufgabe. Die Abstimmung einer Vielzahl von Softwarekomponenten, damit sie reibungslos mit der zugrundeliegenden Hardware zusammenarbeiten können, kann teuer und zeitraubend sein. Talentmangel kann eine bedeutende Hürde für die Einführung neuer Anwendungsfälle darstellen, während leistungsschwache Technologie Daten-Pipelines ausbremst und die Daten Sicherheitsbedrohungen aussetzt.

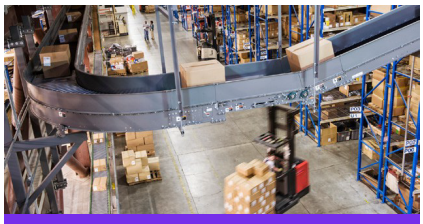
Ebenso werden Unternehmen durch die unsachgemäße Nutzung von KI mit Risiken in Verbindung mit ihrem Ruf, Datenschutz, der Customer Experience und dem Unternehmenswachstum konfrontiert.

Unternehmen setzen zunehmend auf die Hybrid Cloud für KI, wobei Datenschutz und Kontrolle kritische Faktoren sind. Eine Full-Stack-Private Cloud vereinfacht das Management fragmentierter Technologien und steigert die Produktivität von KI-Anwendern. Diese sofort nutzbare Lösung, die als HPE Private Cloud AI bekannt ist, unterstützt den gesamten KI-Entwicklungslebenszyklus und ermöglicht eine reibungslose Nutzung sowohl für den IT-/Cloud-Betrieb als auch für Data-Science-/KI-Teams. Ebenso ermöglicht HPE Private Cloud AI Konsistenz, wenn sich die KI-Infrastruktur im Laufe der Zeit verändert.



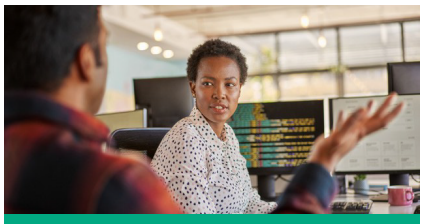
Helfen Sie Data Scientists und Entwicklern, Innovationen mit beispielloser Flexibilität und Leistung schneller einzuführen, um mehr Pilotprojekte erfolgreich in die Produktion zu überführen

Gesteigerte Produktivität



Sorgen Sie für Kontrolle, Governance und Verwaltbarkeit der Umgebungen, die Ihre KI unterstützen

Garantierte Kontrolle



Profitieren Sie vom Besten der Cloud – Cloud-Technologien, Wirtschaftlichkeit und Flexibilität

Experimentieren und skalieren

Abbildung 2. Was Unternehmen benötigen, um das volle Potenzial von KI auszuschöpfen

[2] „Generative AI will drive a foundational shift for companies—IDC“, Computer World, 2024

[3] „Reasons Why Generative AI Pilots Fail To Move Into Production“, Forbes, 2024

Warum HPE Private Cloud AI?

Bei unseren Marktanalysen erkennen wir eine Lücke, mit deren Schließung Hewlett Packard Enterprise eine einzigartige Chance hat, Unternehmen bei der Optimierung des Einsatzes von KI zu helfen. Unsere Antwort lautet HPE Private Cloud AI.

HPE Private Cloud AI ist die erste gemeinsam entwickelte, sofort nutzbare Lösung von NVIDIA AI Computing by HPE – einer neuen gemeinsamen Initiative, die Unternehmen dabei helfen soll, ihre KI-Ambitionen zu verwirklichen. NVIDIA AI Computing by HPE vereint Mitarbeiter, Technologie und Wirtschaftlichkeit, um KI-Bereitstellungen zu beschleunigen, Schutz vor Risiken zu bieten und die KI-Kosten langfristig zu optimieren. Die Lösung zielt speziell auf KI-Modelle ab und lässt sich mühelos im Einklang mit dem Wachstum und der Nutzung von KI-Anwendungsfällen skalieren. Mit dieser bewährten Grundlage steigert HPE Private Cloud AI die Produktivität von Data Scientists und hilft mit einer flexiblen, vorab getesteten und KI-optimierten Private Cloud bei der Überwindung der häufigsten Herausforderungen bei der Operationalisierung von KI.

Der Schwerpunkt vieler heute verfügbarer KI-Lösungen liegt auf Herausforderungen an Tag 0 und Tag 1 (z. B. der Integration des Technologie-Stacks). Dies erzielt nur einen begrenzten Nutzen, da Design und Einrichtung nur einen Bruchteil des KI-Lebenszyklus ausmachen. In der Regel mangelt es diesen Lösungen an Unterstützung für Herausforderungen, die ab Tag 2 entstehen. HPE und NVIDIA wollen das ändern. Wir stärken Ihre KI- und IT-Teams mit einem umfassenden Ökosystem aus unternehmenseigenen und Open-Source-Tools, damit Sie KI-Workloads schnell bereitstellen, Infrastrukturkonfiguration und -management vereinfachen und sich die nötige Freiheit sichern können, um KI-Projekte zu erproben und zu skalieren. Natürlich bleiben Ihre Daten dabei stets geschützt.

Hier kommt HPE GreenLake zum Tragen. HPE Private Cloud AI bietet das Beste der Cloud im Self-Service-Modell, unterstützt durch die HPE GreenLake Cloud. Damit erhalten Sie eine zentrale, plattformbasierte Kontrollebene mit einem Portfolio von Cloud-Services zur Automatisierung, Orchestrierung und Verwaltung von Benutzern und Daten in hybriden Umgebungen. Sie fangen mit einem einzelnen kleinen KI-Pilotprojekt an und weiten den Einsatz schnell auf mehrere Anwendungsfälle oder einen höheren Durchsatz in einer einzigen Lösung aus. Darüber hinaus lässt sich die Lösung vor Ort, in einer Colocation oder in der Cloud bereitstellen, wobei Sie die vollständige Kontrolle über finanzielle Risiken behalten.

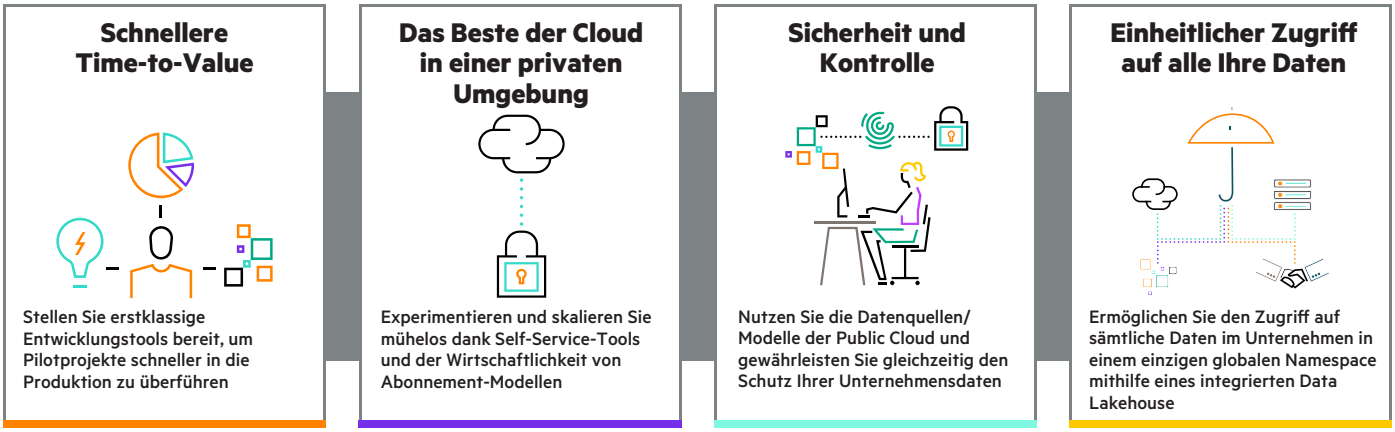


Abbildung 3. HPE und NVIDIA bieten wichtige Funktionen für die KI-Optimierung

Beschleunigung Ihrer KI-Initiativen mit einer sofort nutzbaren Lösung

HPE Private Cloud AI bildet eine umfassend kuratierte Lösung für Ihren KI-Erfolg – von speziell entwickelter Infrastruktur und den richtigen Tools für jede Phase der KI-Entwicklung bis hin zu einer Bibliothek von Modellen mit der größten Relevanz für Ihr Unternehmen. Nutzungserlebnis und Benutzeroberfläche bleiben dabei stets gleich. Dieses Flaggschiff-Angebot von NVIDIA AI Computing by HPE ermöglicht eine bessere KI-Optimierung als alle bisherigen Lösungen.

In dieser Lösung werden KI-Computing, Networking und Software von NVIDIA mit robusten HPE ProLiant Gen12 Inferenz-Servern, HPE AI Datenspeicher und HPE GreenLake Cloud kombiniert, um Unternehmen aller Größenordnungen eine schnelle und flexible Möglichkeit für die Entwicklung und Bereitstellung von GenAI-Anwendungen an die Hand zu geben.

KI-optimierte Hardware wird in kleinen bis mittelgroßen Konfigurationen als einzelnes Rack bereitgestellt. Kleine Konfigurationen eignen sich ideal für grundlegende LLM-Inferenz, mittlere Konfigurationen können RAG für LLMs unterstützen. Auch große Konfigurationen mit mehreren Racks sind verfügbar und ermöglichen die Feinabstimmung selbst der komplexesten Modelle.

Die Softwareebene bietet eine spezialisierte Reihe von KI-Tools, bei denen NVIDIA AI Enterprise-Software zum Einsatz kommt, um Ihre KI-Anforderungen langfristig zu erfüllen. In einer einfachen Konfiguration bietet die HPE AI Essentials Software eine kuratierte Reihe von Tools von HPE und NVIDIA, die Daten-Pipelines und Anwendungsfälle beschleunigen. Die Integration mit NVIDIA NIM Inferenz-Microservices hilft Ihnen bei der Erstellung von Daten-Pipelines, der Entwicklung und Feinabstimmung Ihrer Modelle und einer beeindruckend schnellen Bereitstellung von KI-Anwendungen. Tools der Enterprise-Klasse unterstützen die Zusammenarbeit mit rollenbasierter Zugriffskontrolle, Datenversionierung und -herkunft sowie Entwicklungsfunktionen für die Feinabstimmung von Modellen.

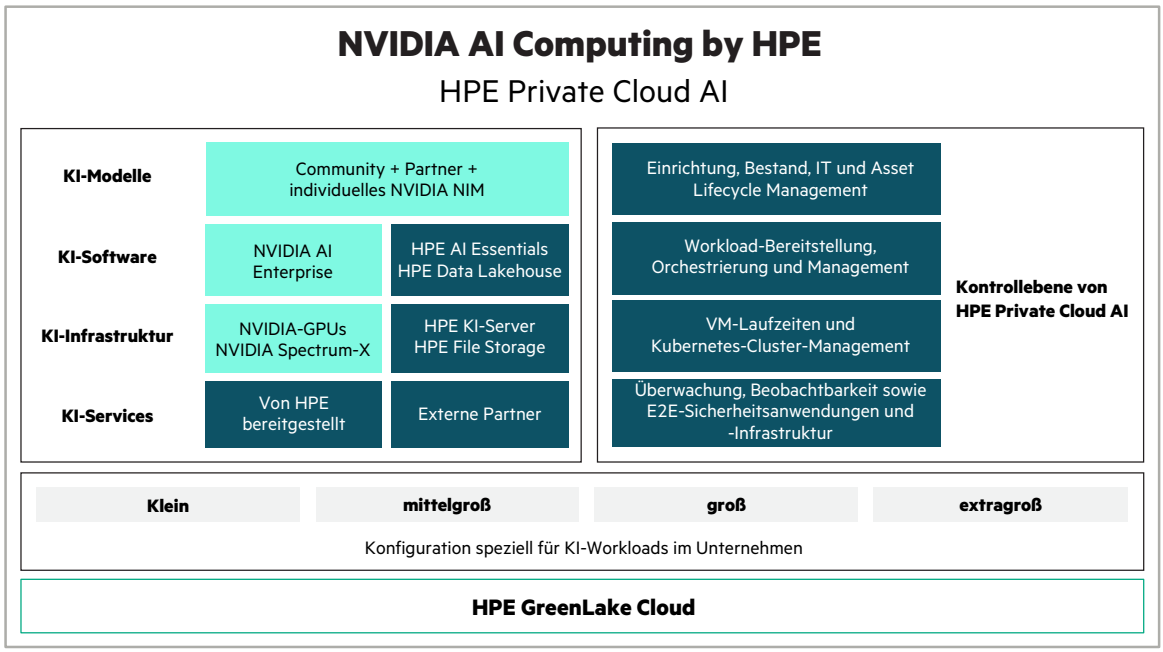


Abbildung 4. Architektur von HPE Private Cloud AI

Konfigurationen	klein	mittelgroß	groß	extragroß
Ideal für	Inferenz	Inferenz + RAG	Inferenz + RAG + Feinabstimmung	Inferencing + RAG + Feinabstimmung
Computing	4 oder 8 NVIDIA L40S-GPUs	8 oder 16 NVIDIA L40S-GPUs	16 oder 32 NVIDIA H100 NVL-GPUs	12 oder 24 NVIDIA GH200 NVL2
Datenspeicher	30 TB bis 248 TB	109 TB bis 390 TB	670 TB bis 1088 TB	670 TB bis 1088 TB
Networking	100 GbE NVIDIA Networking	200 GbE NVIDIA Networking	400 GbE NVIDIA Networking	800 GbE NVIDIA Networking
Stromversorgung	Rack mit bis zu 8 kW	Rack mit bis zu 17,7 kW	bis zu 2x 25 kW	bis zu 2x 25 kW

Abbildung 5. Infrastrukturkonfigurationen für HPE Private Cloud AI

Die Zukunft der KI in der Private Cloud

Eines ist sicher: KI wird unser Leben und unsere Arbeit weiterhin von Grund auf verändern, indem sie unser Leben einfacher und sicherer macht und neue Herausforderungen und ethische Fragen aufdeckt, die uns nie zuvor begegnet sind.

Diese Fortschritte werden die Finanzdienstleistungsbranche revolutionieren, indem sie Aufgaben wie die Prüfung komplexer Finanzdokumente (z. B. Kreditanträge) beschleunigen und dabei unser Geld mit fortschrittlichen Funktionen für die Betrugserkennung und -verhinderung besser schützen. Das Gesundheitswesen kann sich auf personalisierte Behandlungen und schnelle Diagnosen freuen. Gleichzeitig werden Ärzte dank leistungsstarker virtueller Assistenten, die umfassendes medizinisches Wissen in Kombination mit privaten Patientendaten nutzen können, von Verwaltungsaufgaben befreit. Branchen wie Einzelhandel und öffentlicher Sektor werden ihre Customer Experience und Effizienz durch Automatisierung und dynamische Prognosen, die Risiken reduzieren und eine höhere Kundenzufriedenheit ermöglichen, verbessern können. Und diese Anwendungsbereiche sind erst der Anfang.

KI wird in den kommenden Jahrzehnten ein schnell wachsender Bestandteil der IT-Landschaft bleiben. Eine Private Cloud-Umgebung ist von entscheidender Bedeutung, um die Komplexität der KI-Einführung zu verringern und Risiken zu reduzieren, während Sie experimentieren, Innovationen einführen und die Grenzen des Möglichen mit KI sprengen.

HPE und NVIDIA sind bereit, Sie auf diesem Weg zu begleiten. HPE Private Cloud AI ist anders als alle anderen heute verfügbaren Angebote, denn bei der Entwicklung haben wir an Ihre heutigen und künftigen Anforderungen gedacht – ganz gleich, ob Sie bereits ein erfahrener KI-Anwender sind oder gerade erst loslegen. HPE AI Services stehen global zur Unterstützung Ihrer Transformation bereit. Die Experten von HPE und NVIDIA arbeiten bei der Planung, Einführung und Verwaltung Ihrer KI-Umgebung mit Ihnen zusammen – von der Strategie und Technologieauswahl über Design und Machbarkeitsstudien bis hin zu Bereitstellung, laufenden Produktion und Verwaltung.

Erfahren Sie, wie eine Private Cloud Ihnen eine bessere Kontrolle und Sicherheit bei der KI-Nutzung ermöglichen kann, damit Sie das Potenzial von KI voll ausschöpfen können.

Weitere Informationen unter

HPE.com/Private-Cloud-AI

Chat mit Vertrieb

Hewlett Packard Enterprise

a00141386DEE