Kostenfallen durch komplexe KI erkennen und beseitigen

Bessere Beherrschung der KI durch starke Private-Cloud-Strategie

Künstliche Intelligenz (KI) wird oft als der ultimative Game Changer angepriesen – eine technologische Kraft, die die Arbeitsweise von Branchen, Unternehmen und Einzelpersonen in allen Aspekten ihres Bestehens bzw. Lebens neu gestalten wird.

Doch für viele Unternehmen war der Weg von prestigeträchtigen KI-Pilotprojekten zum Einsatz in der Praxis mühsam. Versteckte Kosten, komplexe Infrastrukturanforderungen und schwieriger Betrieb bringen oft sogar bestens geplante KI-Initiativen zum Scheitern und führen zu Frustration statt zu Veränderung.

Obwohl fast die Hälfte (44 %) der weltweiten IT-Führungskräfte glaubt, dass ihre Unternehmen umfassend darauf vorbereitet sind, die Vorteile der KI zu nutzen, wie aus einer Anfang 2024 veröffentlichten HPE Umfrage[1] hervorgeht, heißt es in einem aktuellen RAND-Bericht[2], dass letztendlich mehr als 80 % der KI-Projekte scheitern. Darüber hinaus wurden einer Umfrage von WEKA Resources zufolge 51 % dieser KI-Initiativen in der Produktion noch nicht im erforderlichen Maßstab umgesetzt.[3]

„Wir sind uns alle der transformativen Kraft der KI bewusst, aber KI ist nicht einfach", so Antonio Neri, CEO von Hewlett Packard Enterprise, während seiner Keynote auf der HPE Discover 2024-Konferenz im Juni. „Sie ist kompliziert und birgt viele Risiken. Es ist verlockend, bei der KI voreilig einzusteigen – niemand möchte zurückbleiben. Dennoch sind Innovationen um jeden Preis sehr gefährlich."[4]

Warum der Einsatz von KI so schwierig ist

Ein Großteil der mit KI verbundenen Risiken resultiert aus den folgenden vier zentralen Herausforderungen, die, wenn sie nicht erkannt und nicht angegangen werden, jedes Projekt zum Scheitern bringen können:

1. Unzureichende Datenaufbereitung und -verwaltung

Unzureichend aufbereitete Daten sind möglicherweise die Achillesferse jeder KI-Initiative. Zur Datenaufbereitung gehört das Bereinigen, Organisieren und Kuratieren von unformatierten Daten, um sicherzustellen, dass sie genau, relevant und für das Trainieren und Optimieren von KI-Modellen passend formatiert sind. Dennoch gaben weniger als sechs von zehn Befragten in der HPE Umfrage an, dass ihr Unternehmen die wichtigsten Phasen der Datenaufbereitung für die Verwendung in KI-Modellen vollständig bewältigen könne – vom Zugriff auf (59 %) und Speichern der Daten (57 %) bis hin zur Verarbeitung (55 %) und Wiederherstellung (51 %).5 Diese Diskrepanz birgt die Gefahr, dass der Prozess der KI-Modellerstellung verlangsamt wird und die Wahrscheinlichkeit steigt, dass das Modell ungenaue Erkenntnisse und einen negativen Return on Investment liefert.

„Die Daten, die Sie KI-Anwendungen zuführen, und die generativen Modelle, die diese anschließend verwenden, bestimmen fast vollständig, wie gut die KI-Workloads Geschäftsaufgaben erledigen", sagt Alexander Ollman, Produktmanager von HPE Private Cloud AI.

2. Ineffektive Skalierung und Infrastruktur

Die Skalierung von KI-Workloads erfordert erhebliche Investitionen in Hardware und Software. Eine Überbereitstellung der Infrastruktur führt zur Verschwendung von Ressourcen, während eine Unterbereitstellung Engpässe zur Folge hat, die die Bereitstellung verlangsamen.

Wie Ollman anmerkt, ist die Bereitstellung einer einzelnen KI-Instanz zwar einfach, die Skalierung auf Tausende – oder sogar Zehntausende – Benutzer erfordert jedoch erhebliche GPU-Ressourcen, Bandbreite und eine robuste Infrastruktur. Die Herausforderung besteht darin, Ressourcen dynamisch zu verwalten, um Spitzenlasten zu bewältigen, ohne während Ausfallzeiten eine Überlastung zu verursachen. Daher sind effektive Beobachtungs- und Managementtools von entscheidender Bedeutung, um die Skalierung zu optimieren und die Sicherheit zu gewährleisten. Dennoch werden diese Faktoren oft übersehen, sagt Ollman, was zu Ineffizienzen und steigenden Kosten führt.

3. Schwierige Governance und Sicherheit

Die Wahrung des Datenschutzes, die Gewährleistung der Compliance mit gesetzlichen Vorschriften und die Festlegung angemessener Schutzmaßnahmen für KI-Ausgaben können in der heutigen globalen politischen Landschaft eine gewaltige Herausforderung darstellen. Ohne einen soliden Governance-Rahmen riskieren Unternehmen einen Reputationsschaden und Bußgelder. Dies könnte für sie problematisch werden, da zunehmend Vorschriften, wie etwa das umfassende EU-Gesetz zur künstlichen Intelligenz, erlassen werden.

[1, 5] „Global report finds organizations overlook huge blind spots in their AI overconfidence", HPE, 30. April 2024

[2] „The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed", RAND, 13. August 2024.

[3] „2024 Global AI Trends Report", WEKA, August 2024.

[4] „Intelligence has no limits", HPE Discover Las Vegas, HPE, 2024.

4. Fragmentierte Tools, isolierte Teams

Die Bereitstellung von KI erfordert häufig die Verwaltung eines Labyrinths aus Tools, Teams und Plattformen, was zu Ineffizienzen und Verzögerungen führen kann.

„Wenn mehrere Teams auf mehreren Infrastrukturinstanzen arbeiten, kann das für ein Unternehmen erhebliche Kosten verursachen", meint Ollman. In großen Unternehmen konkurrieren die Teams um Ressourcen, so sein Kommentar, und IT-Manager verbringen häufig sehr viel Zeit damit, Tickets zu bearbeiten und ständig wechselnde Anforderungen zu erfüllen. Zentralisierte Tools, die den Teams die Selbstbedienung ermöglichen und gleichzeitig Transparenz für die IT gewährleisten, können die resultierenden Verzögerungen reduzieren. Durch die Rationalisierung der Abläufe und Minimierung des ständigen Wechsels können Unternehmen versteckte Kosten vermeiden und die Effizienz ihrer KI-Bereitstellungen verbessern, sagt Ollman.

Vereinfachung der KI-Reise

HPE und NVIDIA haben diese Hürden erkannt und bieten Lösungen an, die Unternehmen bei deren Überwindung unterstützen. So wird KI von einem gewaltigen Unterfangen zu einer überschaubaren Investition. Das HPE Private Cloud AI-Angebot ist für diese Strategie von zentraler Bedeutung.

Mit HPE Private Cloud AI können Unternehmen KI-Workloads in nur wenigen Schritten bereitstellen und verwalten. Die Lösung kombiniert die vertrauenswürdige Infrastruktur von HPE mit den leistungsstarken GPUs und der KI-Software von NVIDIA und erstellt so ein vorintegriertes System, bei dem Geschwindigkeit und Einfachheit im Vordergrund stehen.

Als Teil des Portfolios von NVIDIA® AI Computing by HPE umfasst HPE Private Cloud AI:

- **Sofort nutzbare Bereitstellung:** HPE Private Cloud AI ist mit KI-Modellen, Software, Infrastruktur und Services vorintegriert und kann von HPE entweder in Ihrer Umgebung oder an einem ausgewählten Standort innerhalb weniger Stunden bereitgestellt werden. Dies kann die Produktivität der Data Science und die Time-to-Value für generative KI-Lösungen der Enterprise-Klasse beschleunigen.
- **Sicherer und einheitlicher Datenzugriff:** Die Full-Stack Private Cloud-Lösung von HPE und NVIDIA vereinfacht das Datamanagement und reduziert Kosten und Komplexität durch die Integration, Organisation und Governance von Unternehmensdaten für nahtlosen Zugriff, Integrität, Sicherheit, Datenschutz und Compliance. Die integrierte HPE Ezmeral Data Fabric Software vereinfacht außerdem das Datamanagement und reduziert Kosten und Komplexität, indem sie Unternehmensdaten mit einem einzigen globalen Namespace integriert, organisiert und verwaltet. Dies ermöglicht einen nahtlosen Datenzugriff bei gleichzeitiger Wahrung der Datenintegrität und trägt dazu bei, die Compliance von Branchenstandards und -vorschriften sicherzustellen.
- **Vertrauen und Kontrolle:** Umfangreiche mehrschichtige Kontrollen schützen geistiges Eigentum, vertrauliche Daten und angepasste Modelle der Benutzer und halten hohe Leistung, Zuverlässigkeit und Nutzung der KI-Infrastruktur aufrecht.
- **Flexible Evergreen Cloud-Erlebnisse:** Sofort einsatzbereite Hardware- und Softwarelösungen lassen sich mit Ihren wachsenden Anforderungen skalieren und bieten gleichzeitig die gleiche Leistung und Unterstützung für die neuesten Fortschritte im Bereich der KI. Erweiterte Überwachung und Managementtools, integriert in die HPE GreenLake Cloud, bieten zudem Echtzeit-Einblicke in Leistung, Nachhaltigkeitskennzahlen, Daten-Governance und Kosten. Solche Leistungsmerkmale ermöglichen es Ihnen, Engpässe zu identifizieren und zu beheben, Kosten und Stromverbrauch zu senken und eine optimale Nutzung der NVIDIA AI Computing-Ressourcen sicherzustellen.

Die langjährige Partnerschaft von HPE mit NVIDIA war für unsere gemeinsame KI-Strategie von entscheidender Bedeutung. Sie wurde gefestigt, als wir Anfang 2024 NVIDIA AI Computing by HPE vorstellten, ein Portfolio gemeinsam entwickelter KI-Lösungen und -Integrationen, die es Unternehmen ermöglichen, die Einführung generativer KI zu beschleunigen. Das Portfolio beginnt mit HPE Private Cloud AI und baut darauf auf, indem es NVIDIA KI-Computing, -Netzwerke und -Software in HPE AI Speicher- und Computing-Systeme und HPE GreenLake integriert.

Nachhaltigkeit im Vordergrund

Mit der zunehmenden Verbreitung von KI wird auch die Nachhaltigkeit zu einem dringenden Anliegen. KI-Workloads verbrauchen enorme Mengen Energie, wodurch erhebliche Wärme erzeugt wird und effiziente Kühllösungen erforderlich sind.

HPE begegnet dieser Herausforderung mit fortschrittlichen Flüssigkeitskühlungstechnologien, die die Energieeffizienz verbessern und gleichzeitig die Betriebskosten senken. „Wärme kostet Geld", sagt Ollman. „Unsere Flüssigkeitskühlsysteme senken die Kosten und sorgen für Spitzenleistung der KI-Infrastruktur."

Neri hat das Engagement von HPE für verantwortungsvolle Innovation unterstrichen: „HPE ist mit mehr als zwei Jahrzehnten Erfahrung und über 300 Patenten Vorreiter bei der Anwendung der neuesten Flüssigkeitskühlungstechnologie", schrieb er in einem Beitrag im HPE Newsroom.[6] „Heute sind sieben der zehn besten Supercomputer von HPE auf der Green-500-Liste vertreten, was teilweise auf unsere führende Position im Bereich der Flüssigkeitskühlung zurückzuführen ist."

HPE GreenLake bietet außerdem eine Reihe leistungsstarker Dashboards zur Nachverfolgung von Kosten, Verbrauch und CO2-Fußabdruck von Kundensystemen, einschließlich derjenigen, die KI-Funktionen unterstützen. Gemeinsam ermöglichen sie eine genauere Kostenvorhersage als in der Public Cloud, denn sie informieren Sie darüber, wie viel Ihnen für die Nutzung des Netzwerks verrechnet wird, und sie unterstützen Sie bei der Kapazitätsplanung, so Ollman.

Eine Zukunftsvision

Der Weg zur Einführung von KI ist kompliziert, aber HPE und NVIDIA ebnen Ihnen den Weg, damit Sie ihn erfolgreich beschreiten können, und unterstützen Innovationen, die Ihnen beim Erreichen Ihrer Geschäftsziele helfen können.

Mit vertrauenswürdigen Technologiepartnern ist Ihr Unternehmen in der Lage, die Herausforderungen der KI-Komplexität zu bewältigen und Möglichkeiten in Realitäten zu verwandeln.

Weitere Informationen

HPE.com/AI

[6] „Accelerating the generative AI industrial revolution at largest-ever HPE Discover event", Antonio Neri, HPE Newsroom, 20. June 2024

Visit HPE.com

Jetzt chatten

© Copyright 2025 Hewlett Packard Enterprise Development LP. Die Informationen in diesem Dokument können sich jederzeit ohne vorherige Mitteilung ändern. Neben der gesetzlichen Gewährleistung gilt für Produkte und Services von Hewlett Packard Enterprise (HPE) ausschließlich die Herstellergarantie, die in den Garantieerklärungen für die jeweiligen Produkte und Services explizit genannt wird. Aus dem vorliegenden Dokument sind keine weiterreichenden Garantieansprüche abzuleiten. Hewlett Packard Enterprise haftet nicht für hierin enthaltene technische oder redaktionelle Fehler oder Auslassungen.

NVIDIA ist eine Marke und/oder eingetragene Marke der NVIDIA Corporation in den USA und anderen Ländern. Alle Marken von Dritten sind Eigentum ihrer jeweiligen Rechteinhaber.

a50011916DEE, Rev. 2

HEWLETT PACKARD ENTERPRISE

hpe.com