Entwicklung auf einer integrierten KI-Plattform mit Red Hat AI Enterprise

Der rasante Aufstieg von KI

Der bereits weit verbreitete Einsatz von KI ist in Zahlen messbar:

- 55 % der Unternehmen nutzen gen KI.[1]
- Über 50 % der Unternehmen führen Proofs of Concept durch oder implementieren ausgewählte Use Cases für agentische KI.[2]
- 30 % der KI-gestützten Anwendungen werden Ende 2026 voraussichtlich agentische KI nutzen.[2]

Neue KI-Chancen führen zu höherer Komplexität

KI bietet Unternehmen die Möglichkeit, die Art und Weise, wie sie agieren und konkurrieren, neu zu definieren. Sie ist in kurzer Zeit zu einem wichtigen Bestandteil der heutigen Geschäftslandschaft geworden. Doch trotz ihrer zunehmenden Verbreitung sehen sich viele Unternehmen Herausforderungen gegenüber, die eine optimale Ausschöpfung des Potenzials von KI verhindern.

Dazu gehören unter anderem steigende Modellkosten, komplexe Anpassungsanforderungen, strenge Deployment-Beschränkungen und das Schritthalten mit dem hohen Innovationstempo.

Unternehmen konzentrieren sich zunehmend auf Innovationen mit gen KI, agentischen KI-Workflows und souveräner KI. Dabei gestaltet sich jedoch die Bewältigung dieser Herausforderungen immer komplexer. Dies erfordert einen strategischen Wechsel zu KI-Plattformen, mit denen Unternehmen Inferenzkosten senken, Skalierung und Monitoring optimieren, souveräne On-Premise-KI entwickeln und sich schnell an Veränderungen anpassen können.

Bewältigen Sie die Komplexität von KI mit Unternehmensplattformen und -tools

Ganz gleich, ob Unternehmen gerade erst mit der Einführung von KI beginnen oder bereits etablierte KI-Initiativen im gesamten Unternehmen skalieren – Red Hat® AI unterstützt sie bei der Beschleunigung von KI-Innovationen und der Reduzierung der Betriebskosten für die Entwicklung und Bereitstellung von KI-Lösungen in Hybrid Cloud- und Multi Cloud-Umgebungen.

Red Hat AI bietet kosteneffiziente Lösungen mit optimierten Modellen und effizienter Inferenz, optimierter Integration mit privaten Daten und beschleunigter Bereitstellung von agentischer KI auf einer skalierbaren, flexiblen Plattform. So können Unternehmen den Lifecycle sowohl von prädiktiven als auch von gen KI-Modellen in großem Umfang verwalten und überwachen – von Deployments mit einem einzelnen Server bis hin zu horizontal skalierten verteilten Plattformen.

Basierend auf bewährten Open Source-Technologien, bietet Red Hat AI Unterstützung für eine breite Palette führender Open Source KI-Modelle, damit Unternehmen Innovationen beschleunigen und den Zugang zu wertvollen Tools und Technologien demokratisieren können.

Hinzu kommt ein KI-Partnernetzwerk, das getestete und zertifizierte Produkte und Services mit Schwerpunkt auf Performance, Stabilität und Beschleuniger-Support für verschiedene Infrastrukturen anbietet. Dies hilft Unternehmen bei der Bewältigung wichtiger geschäftlicher und technischer Herausforderungen mit KI.

1 IDC Tech Buyer-Präsentation. „Worldwide Generative AI Industry Use Case Early Adoption Trends, 2025: Executive Summary." Dokument Nr. US53280825, April 2025. (Kauf erforderlich)

2 IDC-Umfrage. „Agentic AI Impact on Digital Infrastructure Strategies." Dokument Nr. US53418526, Okt. 2025. (Kauf erforderlich)

Red Hat AI bietet:

- Lückenloses Monitoring und Management des KI-Modell-Lifecycles
- Flexibilität zur Nutzung beliebiger KI-Modelle auf beliebigen Hardwarebeschleunigern
- Eine validierte Sammlung optimierter KI-Modelle
- Flexible und kostengünstige Inferenz in Hybrid Cloud- oder Multi Cloud-Umgebungen
- Zuverlässige Integration in die privaten Datenbestände von Unternehmen
- Optimierte Bereitstellung agentischer KI-Workflows
- Die Fähigkeit einer bedarfsgerechten Skalierung mit effizienter Ressourcennutzung

Red Hat AI Enterprise ist eine einheitliche KI-Plattform, die als Teil des Red Hat AI Portfolios angeboten wird und Unternehmen bei der Bewältigung aktueller Herausforderungen im Zuge der Einführung von KI unterstützt.

6 Gründe für die Wahl von Red Hat AI Enterprise als Ihre KI-Plattform

Red Hat AI Enterprise bietet die Tools und Funktionen, die erforderlich sind, um Unternehmen in jeder Phase ihres Wachstums zu unterstützen. Die Lösung ermöglicht die Entwicklung und Bereitstellung effizienter und kostengünstiger KI-Modelle, -Agenten und -Anwendungen in Hybrid- und Multi Cloud-Umgebungen auf einer einheitlichen KI-Plattform.

Diese Plattform bietet Unternehmen eine einsatzbereite Entwicklungsumgebung und zentrale KI-Funktionen. Dazu gehören Modell-Tuning, leistungsstarke Inferenz, Management agentischer KI-Workflows sowie die Flexibilität, beliebige Modelle zu unterstützen, beliebige Hardware zu verwenden und diese standortunabhängig bereitzustellen. So werden Unternehmen den sich ändernden Anforderungen in Bezug auf Datenspeicherorte gerecht.

Optimierung von KI-Deployments mit flexibler und effizienter Inferenz

Angesichts der wachsenden Beschleunigung von KI-Innovationen hat auch die Komplexität dieser Deployments zugenommen. Dies kann zu ineffizienter Ressourcennutzung, höheren Betriebskosten und einer geringeren Qualität der Ergebnisse führen.

Red Hat AI Enterprise bietet eine optimierte Modellinferenz in verschiedenen Umgebungen, einschließlich On-Premise-, Hybrid- und Multi Cloud-Umgebungen, oder am Edge eines Netzwerks. Unternehmen können damit das Deployment von Modellen optimieren und gleichzeitig den Ressourceneinsatz reduzieren, sodass die Inferenzkosten sinken und die Genauigkeit gewahrt bleibt.

Möglich wird dies durch eine optimierte Runtime auf Basis von vLLM, einem Inferenzserver, der die Ausgabe von gen KI-gestützten Anwendungen beschleunigt, indem er den GPU-Speicher besser nutzt und so Durchsatz und Latenz optimiert. Darüber hinaus beinhaltet die Lösung fortschrittliche Komprimierungstechniken, die zur Kostenoptimierung beitragen, indem sie die Größe und den Rechenaufwand eines Modells verringern.

Wenn ein Unternehmen zur Skalierung bereit ist – von wenigen Modellen auf Dutzende und von Dutzenden von Nutzenden auf Hunderte –, bietet die Plattform die Option für den Einsatz von llm-d. Diese Technologie, die auf den Kerninnovationen von vLLM aufbaut und mit zunehmendem Volumen eine verbesserte Performance bietet, stellt ein verteiltes Inferenz-Framework für kosteneffiziente, vorhersagbare Performance in großem Umfang für gen KI-Modelle bereit.

Verbesserung der Genauigkeit von KI-Modellen mit privaten Unternehmensdaten

Trotz der Fähigkeiten öffentlich verfügbarer gen KI-Modelle benötigen Unternehmen KI-Modelle, die mit ihren eigenen besonderen Daten trainiert werden, um domainspezifische Use Cases zu ermöglichen.

Red Hat AI Enterprise bietet ein optimiertes und konsistentes IT-Erlebnis für Unternehmen, das eine effiziente Modellanpassung fördert und AI Engineers sowie Data Scientists dabei unterstützt, die Genauigkeit und Relevanz von Modellantworten zu verbessern. Unternehmen können Modelle mit organisatorischen Daten für RAG Use Cases (Retrieval-Augmented Generation) verbinden.

Diese Plattform unterstützt auch Fine Tuning sowie kontinuierliches und bestärkendes Lernen und bietet einen modularen, flexiblen Ansatz für die Verarbeitung und Analyse von Dokumenten, die Generierung synthetischer Daten sowie Modellbewertung und -aufgaben.

Beschleunigung der Innovation mit agentischen KI-Workflows

Die KI-Technologie hat sich in relativ kurzer Zeit rasant weiterentwickelt. Angefangen mit prädiktiver KI, die zahlreiche neue Use Cases eröffnet, gefolgt von der Verbreitung von gen KI-Modellen, kommt jetzt die Fokussierung auf agentische KI-Workflows zur Bereitstellung der nächsten Entwicklungsphase hinzu, die KI-Strategien zusätzlich ausweiten und Unternehmen noch mehr geschäftlichen Mehrwert bringen wird.

Red Hat AI Enterprise bietet eine agile, stabile Plattform, mit der Unternehmen dank agentischer KI erfolgreich einen Mehrwert schaffen können. Erreicht wird dies durch eine einheitliche API-Schicht (Application Programming Interface), sofort einsatzbereite Komponenten zur Optimierung agentischer KI-Workflows für Use Cases in Unternehmen, dedizierte Benutzererlebnisse und eine flexible, skalierbare Basis, die die Bereitstellung und Verwaltung agentischer KI-Systeme unterstützt.

Die Plattform bietet auch Unterstützung für das Model Context Protocol (MCP), das eine entscheidende Komponente für agentisches Deployment darstellt und als standardisierter Übersetzer zwischen einer Vielzahl von Tools und Funktionen und Large Language Models (LLM) fungiert.

Flexible und konsistente Skalierung von KI in der Hybrid Cloud

Unternehmen aus zahlreichen Branchen konnten in ersten Tests mit Use Cases die Vorteile von KI belegen. Dennoch haben zu viele von ihnen Schwierigkeiten, diese Vorteile unternehmensweit zu skalieren und gleichzeitig die Kosteneffizienz sowie die Einhaltung von gesetzlichen Vorschriften, Datenschutz und Sicherheit zu gewährleisten. Dies erweist sich als besonders komplex in Unternehmen mit unterschiedlichen IT-Umgebungen wie Hybrid- und Multi Cloud-Umgebungen oder KI-Deployments am Edge eines Netzwerks.

Red Hat AI Enterprise bietet die erforderliche Kontrolle und Konsistenz für das Trainieren, Tuning, Bereitstellen, Verwalten und Skalieren von prädiktiven KI- und gen KI-Modellen, und zwar dort, wo es für die KI-Workload-Strategie eines Unternehmens am sinnvollsten ist – unabhängig von der Hardware oder IT-Umgebung.

Diese Plattform sorgt für Flexibilität, da sie beliebige KI-Modelle auf beliebigen Kombinationen aus Hardware, Original Equipment Manufacturers (OEM), Cloud-Anbietern und Rechenzentren unterstützt. In Kombination mit einer umfassenden Suite von Beobachtbarkeits- und Monitoring-Funktionen können Unternehmen die Performance optimieren, die Kosten kontrollieren, die Governance aufrechterhalten und den Fokus auf Sicherheit richten, wenn sie zur Anpassung an sich ändernde geschäftliche Anforderungen skalieren.

Wahrung der Datensouveränität mit On-Premise KI-Deployments

Weltweit werden fortlaufend gesetzliche Regelungen zur Nutzung von KI sowie zur Speicherung und Verarbeitung von Daten eingeführt. Als Reaktion darauf suchen viele Unternehmen nach Wegen zur Entwicklung von On-Premise KI-Deployments, um die Souveränität und den Schutz ihrer Daten zu wahren.

Durch On-Premise-Speicherung und -Verarbeitung ihrer sensiblen, privaten Daten können Unternehmen nicht nur jederzeit die vielen neuen Bestimmungen einhalten, sondern mit einer geringeren Angriffsfläche und einem verstärkten Fokus auf Sicherheit auch den Schutz dieser Daten gewährleisten.

Red Hat bietet einen umfassenden und mehrschichtigen Ansatz für KI-Sicherheit, mit dem Unternehmen KI-Innovationen zuverlässig nutzen können. Red Hat AI Enterprise bietet diesen Sicherheitsansatz auf einer bewährten Open Hybrid Cloud-Basis, die dazu beiträgt, die besonderen Herausforderungen in Bezug auf Sicherheit und Schutz des KI-Lifecycles zu bewältigen.

Sie ermöglicht echte Souveränität, indem sie beliebige Modelle, Hardware und sicherheitsorientiertes GPU-Sharing in den verschiedenen Cloud- und Edge-Umgebungen einschließlich Air-Gap-Umgebungen unterstützt. Red Hat arbeitet eng mit souveränen Cloud-Anbietern zusammen, um Unternehmen die Bereitstellung von KI in privaten, sicherheitsorientierten Cloud-Umgebungen zu ermöglichen.

Red Hat AI Enterprise beinhaltet außerdem Tools, die durch Erklärbarkeit von Modellen, Fairness und Durchsetzung von Ausgaberichtlinien prüfbares Vertrauen fördern, sowie Funktionen für Lifecycle-Nachverfolgbarkeit mit reproduzierbaren Pipelines und Auditvorbereitung.

Entwicklung von KI auf einer vertrauenswürdigen Kubernetes-Plattform

Red Hat AI Enterprise bietet diesen kompletten Satz an KI-Funktionen als Teil eines integrierten, einsatzbereiten End-to-End-KI-Stacks, der auf der bewährten Basis von Kubernetes aufbaut.

Unternehmen, die bereits Kubernetes und Containerisierung nutzen, profitieren auf diese Weise von einem vertrauten und konsistenten IT-Erlebnis in On-Premise-Rechenzentren, Hybrid Cloud- oder Multi Cloud-Umgebungen sowie KI-Deployments am Edge. Unternehmen können moderne, cloudnative, KI-gestützte Anwendungen auf einer einzelnen, zentralisierten Plattform entwickeln, trainieren, bereitstellen und verwalten und dabei dieselben Tools, Frameworks und Kompetenzen nutzen, mit denen ihre Teams bereits vertraut sind.

Mehr über die Vorteile von Red Hat AI Enterprise erfahren

Erfahren Sie mehr darüber, wie Red Hat AI Enterprise Unternehmen die Bereitstellung von effizienter und kostengünstiger KI in der Hybrid Cloud erleichtert – mit einer einheitlichen Plattform, die jede Phase der KI-Einführung unterstützt.

Über Red Hat

Red Hat, weltweit führender Anbieter von Open Source-Softwarelösungen für Unternehmen, folgt einem communitybasierten Ansatz, um zuverlässige und leistungsstarke Linux-, Hybrid Cloud-, Container- und Kubernetes-Technologien bereitzustellen. Red Hat unterstützt Kunden bei der Entwicklung cloudnativer Anwendungen, der Integration neuer und bestehender IT-Anwendungen sowie der Automatisierung, Sicherung und Verwaltung komplexer Umgebungen. Als bewährter Partner der Fortune 500-Unternehmen stellt Red Hat vielfach ausgezeichnete Support-, Trainings- und Consulting-Services bereit, die unterschiedlichen Branchen die Vorteile der Innovation mit Open Source erschließen. Als Mittelpunkt eines globalen Netzwerks aus Unternehmen, Partnern und Communities unterstützt Red Hat Unternehmen bei der Steigerung ihres Wachstums und auf ihrem Weg in die digitale Zukunft.

EUROPA, NAHOST, UND AFRIKA (EMEA)
00800 7334 2835
de.redhat.com
europe@redhat.com

TÜRKEI
00800 448820640

ISRAEL
1 809 449548

VAE
8000-4449549

facebook.com/redhatinc
@RedHatDACH
linkedin.com/company/red-hat

de.redhat.com