Größerer Nutzen von GenAI-Investitionen mit Dell AI Factory

Der Kosten-Nutzen-Faktor der Implementierung einer Dell AI Factory-Lösung gegenüber AWS und Azure

Generative KI (GenAI) kann Unternehmen jeder Art und Größe dabei unterstützen, ihre Geschäftsziele zu erreichen. Die Auswahl einer richtig dimensionierten Lösung, die hinsichtlich Leistung, Budget und Sicherheitsanforderungen den Anforderungen entspricht, kann jedoch eine Herausforderung darstellen. Es mag zwar sinnvoll erscheinen, große Sprachmodelle (Large Language Models, LLMs) in der Cloud zu hosten, um maximale Flexibilität zu erzielen, doch die Entscheidung für eine KI-Bereitstellung außerhalb des eigenen Rechenzentrums kann im Laufe der Zeit zu Budgetproblemen führen und langfristig höhere Kosten verursachen. In einer Umfrage aus dem Jahr 2024 gaben 46 % der Führungskräfte an, dass die Kosten für die Implementierung von KI ein Problem darstellen – ein deutlicher Anstieg gegenüber nur 3 % im Vorjahr. Unternehmen beginnen aufgrund der Kosten, die durch „Tokenkosten, unerwartete Zusatzkosten und KI-Wildwuchs" entstehen, KI-Initiativen zu stoppen oder zu verschieben.[1]

Um Unternehmen dabei zu helfen, die Gesamtkosten für die Bereitstellung und das Management von GenAI-Workloads zu verstehen, einschließlich Modellfeinabstimmung und Inferenz, haben wir uns die ungefähren Kosten einer On-Premise-Lösung von Dell™ AI Factory mit PowerEdge™ R660- und PowerEdge XE9680-Hardware unter Verwendung von zwei Zahlungsoptionen – einer CAPEX-Lösung mit traditionellem Zahlungsmodell und einer Dell APEX-Abonnementlösung – sowie vergleichbaren Amazon Web Services (AWS) SageMaker- und Microsoft Azure Machine Learning-Lösungen über einen Zeitraum von vier Jahren angesehen. Unseren Berechnungen zufolge waren die Dell AI Factory-Lösungen über vier Jahre hinweg die kosteneffizienteste der Lösungen, die wir verglichen haben. Die Dell APEX-Abonnementlösung reduzierte die Kosten im Vergleich zu AWS um 71 % und im Vergleich zu Azure um 60 %. Dieselbe On-Premise-Lösung von Dell AI Factory ohne Abonnement (CAPEX-Modell) würde die Kosten über vier Jahre im Vergleich zur AWS-Lösung um 71 % und im Vergleich zur von uns kalkulierten Azure-Cloud-Lösung um 61 % senken. Lesen Sie weiter, um zu erfahren, wie Ihr Unternehmen mit einer On-Premise-Lösung von Dell AI Factory für GenAI das Beste aus seiner Investition herausholen kann.

Erhalten Sie mit einer Dell AI Factory-Lösung mehr für Ihre Investition

Bis zu 71 % größere Ersparnis als mit einer Konkurrenzlösung von AWS

Amortisation nach einem Jahr

Bis zu 61 % größere Ersparnis als mit einer Konkurrenzlösung von Azure

Amortisation nach etwa 1,5 Jahren

Gesamtbetriebskosten (TCO) für eine On-Premise-Lösung von Dell AI Factory über einen Zeitraum von 4 Jahren im Vergleich zu Umgebungen von AWS und Azure | *niedrigerer Graph = geringere Kosten*

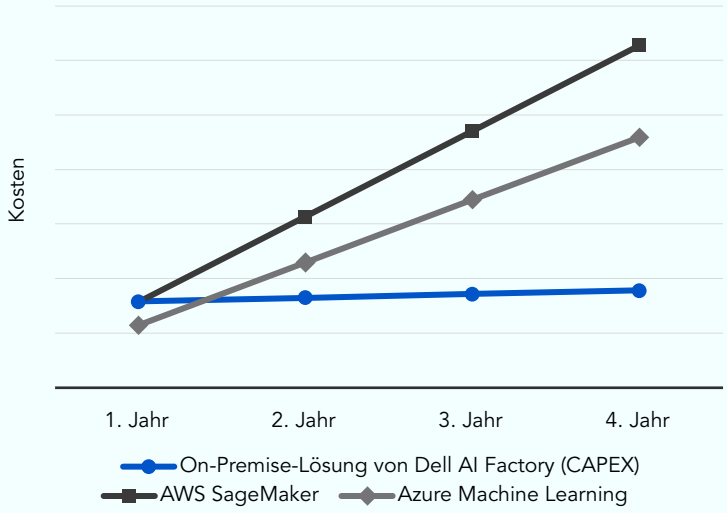


Abbildung 1: Relative Kosten einer On-Premise-Lösung (CAPEX) von Dell AI Factory im Vergleich zu AWS SageMaker und Azure Machine Learning-Lösungen über einen Zeitraum von vier Jahren.

Informationen zur Dell AI Factory

Dell AI Factory ist ein umfassender Ansatz, der entwickelt wurde, um den sich ändernden Anforderungen moderner KI-Anwendungen gerecht zu werden. Die von Dell bereitgestellte AI Factory kombiniert modernste Infrastruktur, Software und Services und bietet Flexibilität zur Beschleunigung von KI-Initiativen. Unternehmen können die gesamte Lösung, einschließlich aller Komponenten, für einen optimierten Ansatz übernehmen oder ihre Konfiguration anpassen, indem sie die speziell auf bestimmte GenAI-Ergebnisse zugeschnittene Infrastruktur und Services von Dell nutzen. Der offene Ökosystemansatz stellt sicher, dass die Dell AI Factory-Lösungen mit verschiedenen Technologien und Workflows kompatibel sind.

Die Dell AI Factory bietet eine breite Palette an Infrastrukturoptionen für moderne Unternehmensanforderungen, darunter die Option zur Nutzung von Dell APEX-Abonnements. Damit ist sie ein wertvolles Framework für Unternehmen, die KI einsetzen möchten, um ihre Daten in positive Geschäftsergebnisse umzuwandeln. Weitere Informationen zu den Vorteilen der Dell AI Factory für Ihr Unternehmen finden Sie unter dell.com/ai.

Übersicht über TCO-Szenarien und -Lösungen

Neue Workloads bedeuten oft neue Investitionen. Viele KI-Workloads erfordern leistungsstarke Komponenten zusätzlich zu den großen Storage-Mengen, die bereits Ihre Daten enthalten. Bei der Implementierung von KI-Workloads müssen Sicherheit, Zeit, Performance und Skalierbarkeit, Benutzerfreundlichkeit und Kosten gegeneinander abgewogen werden. Um eine Vorstellung davon zu vermitteln, wie viel KI-Lösungen kosten, haben wir ein KI-Szenario mit dem Open-Source-Modell Llama 3 8B erstellt und die Kosten für die Ausführung der Workload in vier verschiedenen Umgebungen verglichen. Unser Szenario umfasste vier spezifische Aufgaben in einer GenAI-Workload: Programmierung und Entwicklung von maschinellem Lernen von Data Scientists, Datenverarbeitungsaufgaben, Aufgaben zur Modellfeinabstimmung und Inferenzierungsaufgaben. Diese Aufgaben sorgen dafür, dass das Modell präzise und mit den neuesten unternehmensgenerierten Daten auf dem neuesten Stand bleibt, um optimale Modellergebnisse zu erzielen. Tabelle 1 zeigt die allgemeinen Spezifikationen der vier von uns untersuchten Umgebungen. Hinweis: Wir haben alle Recherchen und Preisermittlungen am 27. März 2025 abgeschlossen. Die Preise können sich nach diesem Datum ändern.

Tabelle 1: Lösungsdetails für den TCO-Vergleich.

Aufgabe	Server/Instanz	GPUs pro Server/ Instanz	Zusätzliche Käufe
On-Premise-Lösung von Dell AI Factory			
Clustermanagement	3 x PowerEdge R660	N. z.	2 x PowerSwitch S5232-ON-Netzwerkinfrastruktur und 1 x PowerSwitch N3200-ON OOB-Management
Laptops			
Datenverarbeitung	2 x PowerEdge XE9680 (jeweils mit 30 TB Storage)	8x NVIDIA H100	
Modellfeinabstimmung			
Schlussfolgerung			
Gemanagte On-Premise-Abonnementlösung von Dell APEX			
Clustermanagement	3 x PowerEdge R660	N. z.	2 x PowerSwitch S5232-ON-Netzwerkinfrastruktur und 1 x PowerSwitch N3200-ON OOB-Management
Laptops			
Datenverarbeitung	2 x PowerEdge XE9680 (jeweils mit 30 TB Storage)	8x NVIDIA H100	
Modellfeinabstimmung			
Schlussfolgerung			

Aufgabe	Server/Instanz	GPUs pro Server/ Instanz	Zusätzliche Käufe
AWS SageMaker-Lösung			
Clustermanagement	N. z.	N. z.	7 TB EBS-Storage pro Monat für ml.r5.16xlarge-Instanzen und 1 TB für die eingehende und 15 TB für die ausgehende S3-Datenübertragung
Laptops	20x ml.t3.medium	N. z.	
Datenverarbeitung	2x ml.r5.16xlarge	N. z.	
Modellfeinabstimmung	ml.p5.48xlarge	8x NVIDIA H100	
Schlussfolgerung	ml.p5.48xlarge	8x NVIDIA H100	
Azure Machine Learning-Lösung			
Clustermanagement	N. z.	N. z.	10.000.000 Azure Block Blob Storage-Datenübertragungsvorgänge
Laptops	20x D2 v2	N. z.	
Datenverarbeitung	M64	N. z.	
Modellfeinabstimmung	1x ND96isr H100 v5	8x NVIDIA H100	
Schlussfolgerung	1x ND96isr H100 v5	8x NVIDIA H100	

Bitte beachten Sie, dass in dieser Studie Preise für NVIDIA H100-GPUs verwendet werden. Obwohl NVIDIA H200-GPUs veröffentlicht hat und PowerEdge XE9680-Server diese unterstützen, haben wir uns für einen Vergleich der Gesamtbetriebskosten (TCO) für ähnlich konfigurierte Lösungen mit H100-GPUs entschieden. Die genauen Spezifikationen der von uns verglichen Lösungen finden Sie im wissenschaftlichen Hintergrund des Berichts.

Bei dieser Analyse haben wir versucht, ein allgemein anwendbares Beispielszenario zu erstellen, um Kostenunterschiede zwischen Umgebungen zu schätzen. Wir haben uns für das GenAI-Modell Llama 3 8B entschieden, da es sich um ein weit verbreitetes Open-Source-Modell handelt. Wir haben die Kosten für die Entwicklungsnotebooks der Data Scientists für maschinelles Lernen, Datenverarbeitungsaufgaben, kontinuierliche Modellfeinabstimmung und Echtzeitinferenz einbezogen. Nicht berücksichtigt wurden Kosten für Storage, die über das hinausgehen, was die Server oder Instanzen für ihre Aufgaben benötigten.

Bei den On-Premise-Lösungen von Dell gingen wir davon aus, dass die Entwicklungsnotebooks und Clustermanagementaufgaben auf dem Dell PowerEdge R660-Cluster ausgeführt würden, während die Verarbeitungs-, Feinabstimmungs- und Inferenzaufgaben auf dem Dell PowerEdge XE9680-Cluster durchgeführt würden.

Für die Cloud-Lösungen haben wir Instanzen ausgewählt, die den Anforderungen einer Aufgabe entsprechen. Notebookinstanzen waren sehr klein, während wir Verarbeitungsinstanzen einen erheblichen Arbeitsspeicher gaben. Da die Public Cloud-Services für jede Aufgabe eine neue Instanz aktivieren, verfügt jede dieser Aufgaben für die Dauer ihrer Ausführung über eine dedizierte Instanz mit acht GPUs. Daher haben wir die Anzahl der Aufgaben berechnet, die die PowerEdge XE9680-Server bei gleichem GPU-pro-Aufgabe-Verhältnis ausführen können. Außerdem haben wir eine Schätzung der Kosten für die Datenübertragung zum und vom Objektspeicher des Cloud-Anbieters hinzugefügt, um die Kosten für das Verschieben von Daten durch die Cloud zu berücksichtigen.

Um unterschiedliche geschäftliche Realitäten zu berücksichtigen und einen fairen Vergleich zu erzielen, haben wir die folgenden Annahmen getroffen:

- Alle Kosten verstehen sich ohne Steuern, da die spezifischen Steuersätze je nach geografischem Standort variieren.
- Die gesamte Software ist Open-Source-Software mit Lizenzen, die eine kommerzielle Nutzung ermöglichen.
- Die Verwaltungskosten für die Cloud-Lösungen werden nicht berücksichtigt. Bei den Vor-Ort-Lösungen werden die laufenden Kosten für die Systemverwaltung zur Wartung der Hardware und zur Unterstützung der Data Scientists berücksichtigt.
- Bei den On-Premise-Lösungen berücksichtigen wir die Kosten für den physischen Platz im Rechenzentrum sowie für Strom und Kühlung.
- Bei den CAPEX-Anschaffungen für die Dell AI Factory haben wir alle Betriebskosten für Kapital/Abschreibungen ausgeschlossen.

Weitere Informationen zu unseren Annahmen und Berechnungen finden Sie im wissenschaftlichen Hintergrund des Berichts.

Vergleich der Kosten für GenAI: On-Premise-Lösungen von Dell AI Factory gegenüber der Cloud

Annahmen für GenAI-Kostenvergleiche

- Wir gehen davon aus, dass es 22 Arbeitstage pro Monat gibt, wobei Workloads 24 Stunden ausgeführt werden, um die Nutzung zu maximieren.
- Daher bietet jeder Server 528 Stunden Laufzeit pro Monat.
- Datenverarbeitungsaufgaben können die gesamten 528 Stunden x zwei Dell PowerEdge XE9680-Server = 1.056 Stunden Laufzeit ausführen.
- 20 DatenanalystInnen arbeiten 8 Stunden am Tag an 22 Tagen im Monat, also insgesamt 3.520 Stunden.

Da für die Verarbeitungsaufgaben CPU und Arbeitsspeicher benötigt werden, hosten wir sie für die gesamten 1.056 Serverbetriebsstunden auf den PowerEdge XE9680-Servern. Wir teilen die Aufgaben für die Modellfeinabstimmung und die Inferenzierung auf die beiden Server auf, wobei wir davon ausgehen, dass die Workload mehr Zeit für die Feinabstimmung als für die Inferenzierung benötigt. Daher haben wir 792 Stunden pro Monat für Feinabstimmungsaufgaben und 264 Stunden pro Monat für Inferenzierungsaufgaben berechnet.

Schließlich nehmen wir für die Notebooknutzung durch die 20 Data Scientists an, dass alle einen typischen 8-Stunden-Arbeitstag an 5 Tagen pro Woche haben, was insgesamt 3.520 Stunden pro Monat entspricht. Die Anzahl der Data Scientists, die Ihr Unternehmen beschäftigt, um Ihr Modell zu pflegen und zu verfeinern, wird von mehreren Faktoren abhängen, zum Beispiel davon, auf wie viele verschiedene Arten Sie Ihren Datensatz interpretieren möchten oder wie viele Anwendungen Ihr Datensatz speist. Wir haben uns für eine Zahl am oberen Ende der Skala entschieden, um Höchstkosten zu repräsentieren, die für viele Unternehmen gelten würden. Da diese Instanzen in der Public Cloud sehr klein sind und im Vergleich zur Lösung als Ganzes sehr wenig Kosten verursachen, hat die Anzahl der Data Scientists keinen großen Einfluss auf die Gesamtkosten unserer Lösung. Mithilfe dieser Betriebszeitberechnungen konnten wir die Anzahl der Stunden ermitteln, die jeder Instanztyp pro Monat auf den beiden Cloud-Lösungen laufen würde. Die endgültigen Gesamtkosten aller Lösungen finden Sie im wissenschaftlichen Hintergrund des Berichts.

Preisdetails zur On-Premise-Lösung von Dell AI Factory

Dell hat ein Angebot unter Verwendung des von Dell empfohlenen Preises für die On-Premise-Lösung von Dell AI Factory erstellt. Dieses Angebot enthielt die Kosten für Server und Switches, ProDeploy Plus für Vor-Ort-Installationsservices für die Server und einen 5-Jahres-Plan für ProSupport for Infrastructure zur Bereitstellung von Support- und Wartungsservices für die Ausstattung. Hinweis: Wir haben uns für einen 5-Jahres-Supportplan entschieden, da wir unsere TCO auf vier Jahre begrenzt haben, die meisten Server aber drei bis fünf Jahre funktionieren und über die vier Jahre hinaus Service benötigen. Anschließend berechneten wir die Energiekosten für Strom und Kühlung sowie die Kosten für den Rack-Platz im Rechenzentrum für einen Zeitraum von vier Jahren sowie die Verwaltungskosten für die Wartung der Ausstattung für vier Jahre.

Preisdetails zur AWS SageMaker-Cloud-Lösung

AWS unterteilt seinen SageMaker-Service in mehrere Unterservices, die Aufgaben wie Verarbeitung und Training sowie Notebooks von Data Scientists abdecken. Beachten Sie, dass der AWS SageMaker-Unterservice während der Feinabstimmung eines vortrainierten Modells als SageMaker Training bezeichnet wird. Um SageMaker-Preise zu ermitteln, haben wir den AWS-Preisrechner und den Rechner für Einsparungspläne für maschinelles Lernen verwendet.[2,3] Für unsere TCO haben wir die Preise für Instanzen für Notebooks, Verarbeitung, Modellfeinabstimmung und Inferenz wie folgt berechnet:

Tabelle 2: AWS SageMaker-Umgebungsinstanzen und Laufzeitstunden pro Monat.

Instanzmodell	Anzahl der Instanzen	Aufgabe	Laufzeit (Stunden/ Monat)/Instanz
ml.t3.medium	20	Notebook für Data Scientists	176
ml.r5.16xlarge	2	Datenverarbeitung	1.056
ml.p5.48xlarge	1	Modellfeinabstimmung	792
ml.p5.48xlarge	1	Inferenz	264

Annahmen zu Preisdetails:

Wir haben zwei ml.r5.16xlarge-Instanzen für die Datenverarbeitung ausgewählt, um mindestens 1 TB Arbeitsspeicher pro Aufgabe sicherzustellen. Dies basiert auf Untersuchungen, die darauf hinweisen, dass Verarbeitungsaufgaben arbeitsspeicherintensiv sind.[4,5]

- Wir haben 3,5 TB pro Monat an EBS-Storage zu jeder ml.r5.16xlarge-Instanz hinzugefügt, da diese nicht mit Festplatten ausgestattet sind.
- Wir haben zwar die Kosten für den Storage, der das Hauptdatenvolumen hostet, nicht geschätzt, aber wir haben die Kosten für 1 TB für die eingehende und 15 TB für die ausgehende S3-Datenübertragung pro Monat geschätzt, um die Teilmengen der Daten zu berücksichtigen, die die Trainings- und Inferenzaufgaben verwenden werden.
- Die ml.p5.48.large-Instanzen waren mit Direct Attached NVMe-Storage ausgestattet, sodass für diese Instanzen kein EBS-Storage hinzugefügt wurde.

Hinweis: SageMaker umfasst einen Elastic Fabric Adapter (EFA), der hohe Durchsatzraten bietet.[6] Wir sind zwar der Meinung, dass das Netzwerk in der Dell Lösung für unser Szenario angemessen ist, aber Sie können sich für eine Netzwerkkonfiguration mit mehr Bandbreite entscheiden. Daher ist es möglich, dass die AWS-Lösung je nach Netzwerkauswahl mehr Aufgaben verarbeiten kann als die Dell Lösung.

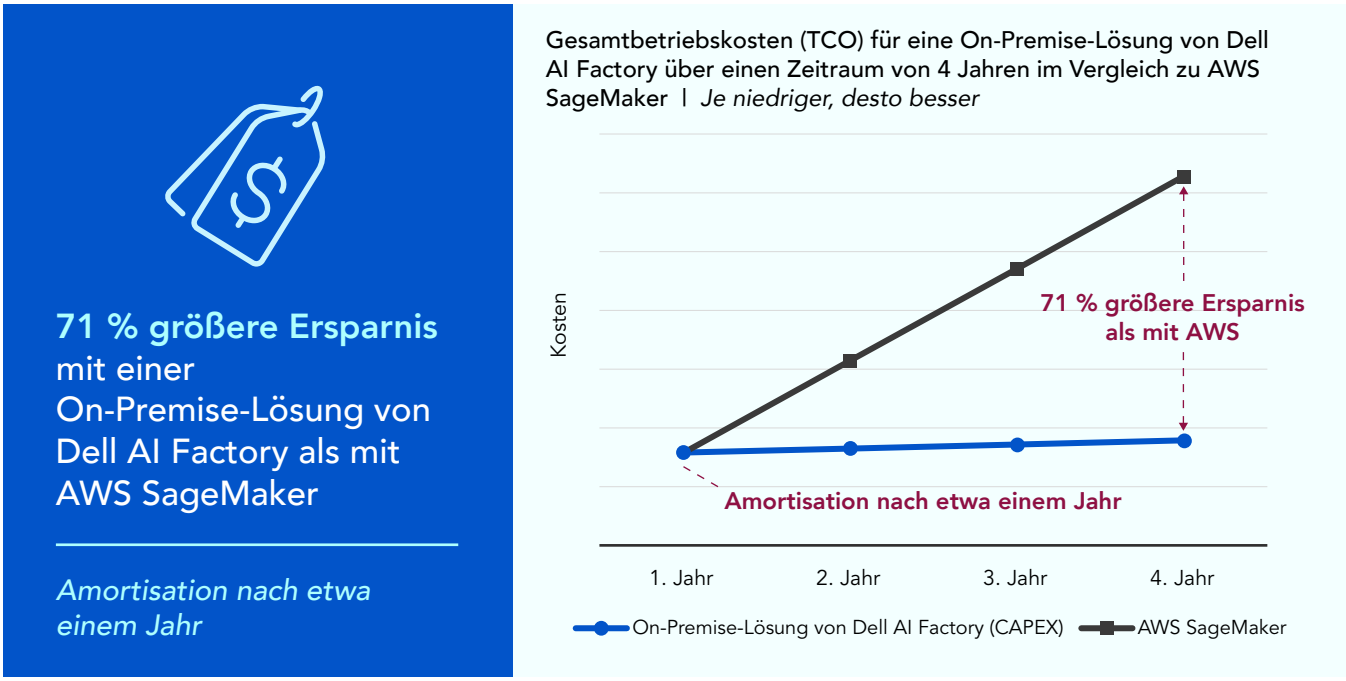
AWS bietet sowohl On-Demand-Preise als auch SageMaker-Sparpläne. On-Demand-Preise sind am teuersten, während die Sparpläne Kosten mit einer Laufzeit von drei Jahren um bis zu 64 % senken.[7] AWS bietet keine spezifischen Preise für eine vierjährige Vertragslaufzeit an. Um die bestmöglichen Kosten für unsere vierjährige Gesamtbetriebskostenberechnung zu ermitteln, haben wir die AWS-Konfiguration daher anhand des Preises für eine dreijährige Vertragslaufzeit berechnet und auf vier Jahre umgerechnet.[8] (Für eine alternative Betrachtung der AWS-Preise haben wir auch die Kosten für vier Jahre berechnet, indem wir drei Jahre zum Preis für eine dreijährige Vertragslaufzeit und ein Jahr zum Preis für eine einjährige Vertragslaufzeit zugrunde gelegt haben. Diese zusätzlichen Ergebnisse finden Sie im wissenschaftlichen Bericht.) Darüber hinaus bietet AWS den KundInnen die Möglichkeit, Kosten im Voraus zu bezahlen, um eine größere Kostensenkung zu erreichen. Diese Option haben wir bei unseren TCO-Berechnungen gewählt. Bitte beachten Sie, dass wir unsere AWS-Lösung für die Region „US East (Ohio)" berechnet haben und die Preise je nach Region variieren können.

Sparen Sie mit einer On-Premise-Lösung von Dell AI Factory im Vergleich zu AWS SageMaker

Anhand der oben genannten Annahmen für beide Lösungen haben wir einen Vergleich der TCO über vier Jahre berechnet. Unsere Berechnungen zeigen, dass die Wahl der On-Premise-Lösung von Dell AI Factory zur Ausführung von GenAI-Workloads im Vergleich zur Ausführung derselben Workloads auf AWS SageMaker echte Einsparungen bieten könnte.

Wie Abbildung 2 zeigt, haben wir berechnet, dass die On-Premise-Lösung von Dell AI Factory 71 Prozent weniger kosten würde als eine ähnliche AWS SageMaker-Lösung. Da die AWS-Lösung über einen Zeitraum von vier Jahren 3,5-mal so viel kostet, können NutzerInnen davon ausgehen, dass sie mit einer lokalen Dell AI Factory-Lösung im Vergleich zu den Kosten für das AWS-Hosting bereits nach einem Jahr die Gewinnschwelle erreichen würden.

Abbildung 2: Relative Kosten einer On-Premise-Lösung von Dell AI Factory und einer AWS SageMaker-Lösung über vier Jahre.



Preisdetails zur Azure Machine Learning-Cloud-Lösung

Für die Azure Machine Learning-Serviceumgebung haben wir Instanzen für dieselben vier Aufgaben wie die AWS-Umgebung ausgewählt: Entwicklungsnotebooks für Data Scientists, Datenverarbeitung, Feinabstimmung und Inferenz. Wir haben unsere Preise über den Azure-Preisrechner ermittelt und uns für die Option eines 4-Jahres-Sparplans mit Reservierung entschieden.[9] Die Instanzen, deren Preise wir berechnet haben, lauten wie folgt:

Tabelle 3: Azure Machine Learning-Umgebungsinstanzen und Laufzeitstunden pro Monat.

Instanzmodell	Anzahl der Instanzen	Aufgabe	Laufzeit (Stunden/ Monat/ Instanz)
D2 v2	20	Notebook für Data Scientists	176
M64	1	Datenverarbeitung	1.056
ND96isr H100 v5	1	Modellfeinabstimmung	792
ND96isr H100 v5	1	Inferenz	264

Annahmen zu Preisdetails für Azure Machine Learning

- Alle Azure Machine Learning-Instanzen verfügen über angeschlossenen Block-Storage, sodass wir keinen zusätzlichen Storage für die Azure-Umgebung berechnet haben. Wie bei unseren AWS-Berechnungen haben wir jedoch auch in diesem Fall ungefähr 10.000.000 Block Blob Storage-Datenübertragungsvorgänge für den Datentransfer in und aus den Machine Learning-Instanzen angenommen. Der Rechner für diese Transaktionen umfasst mehrere spezifische Transaktionen, z. B. Schreibvorgänge, Lesevorgänge und mehr. Wir haben jeweils 10.000.000 ausgewählt.

Azure bietet „Pay as you go"-Preise, Azure-Sparpläne und Azure-Reservierungsoptionen für den Machine Learning-Service.[10] Azure bietet zwar wie AWS eine Vorauszahlungsoption an, jedoch scheinen sich dadurch weder die monatlichen Kosten zu ändern noch Rabatte gewährt zu werden. Um die Preisgestaltung der AWS-Umgebung bestmöglich abzubilden, haben wir uns für den 3-Jahres-Reservierungsplan entschieden und die Kosten anteilig für vier Jahre berechnet. (Wie bei AWS haben wir auch die Kosten und Einsparungen für eine dreijährige Reservierung plus eine einjährige Reservierung berechnet, die Sie in den wissenschaftlichen Grundlagen dieses Berichts einsehen können.) Wie bei AWS haben wir unsere Microsoft Azure-Lösung in der Region „USA Ost 2" bewertet. Bitte beachten Sie, dass die Preise je nach Region variieren können.

Mit einer On-Premise-Lösung von Dell AI Factory anstelle von Azure ML amortisieren sich die Kosten in weniger als zwei Jahren

Anhand der oben genannten Annahmen haben wir die Kosten einer vierjährigen Azure-Lösung berechnet und mit unseren vierjährigen TCO-Schätzungen für die On-Premise-Lösung von Dell AI Factory verglichen. Auch hier zeigen unsere Berechnungen, dass die On-Premise-Lösung von Dell AI Factory für GenAI-Workloads im Vergleich zu einer vergleichbaren Azure ML-Lösung erhebliche Einsparungen über vier Jahre bieten kann.

Tatsächlich schätzen wir, dass die On-Premise-Lösung von Dell AI Factory 61 % weniger kostet als eine vergleichbare Azure Machine Learning-Lösung (siehe Abbildung 3). Diese Ergebnisse zeigen, dass Sie mit einer On-Premise-Lösung von Dell AI Factory für GenAI, bei der Sie Ihre Hardware vor Ort behalten, Ihr GenAI-Budget optimieren können. Da Azure ML über einen Zeitraum von vier Jahren mehr als 2,5-mal so viel kostet, können Kunden der On-Premise-Lösung von Dell AI Factory im Vergleich zu den Azure-Preisen damit rechnen, dass sich die Investition nach etwa anderthalb Jahren amortisiert hat.

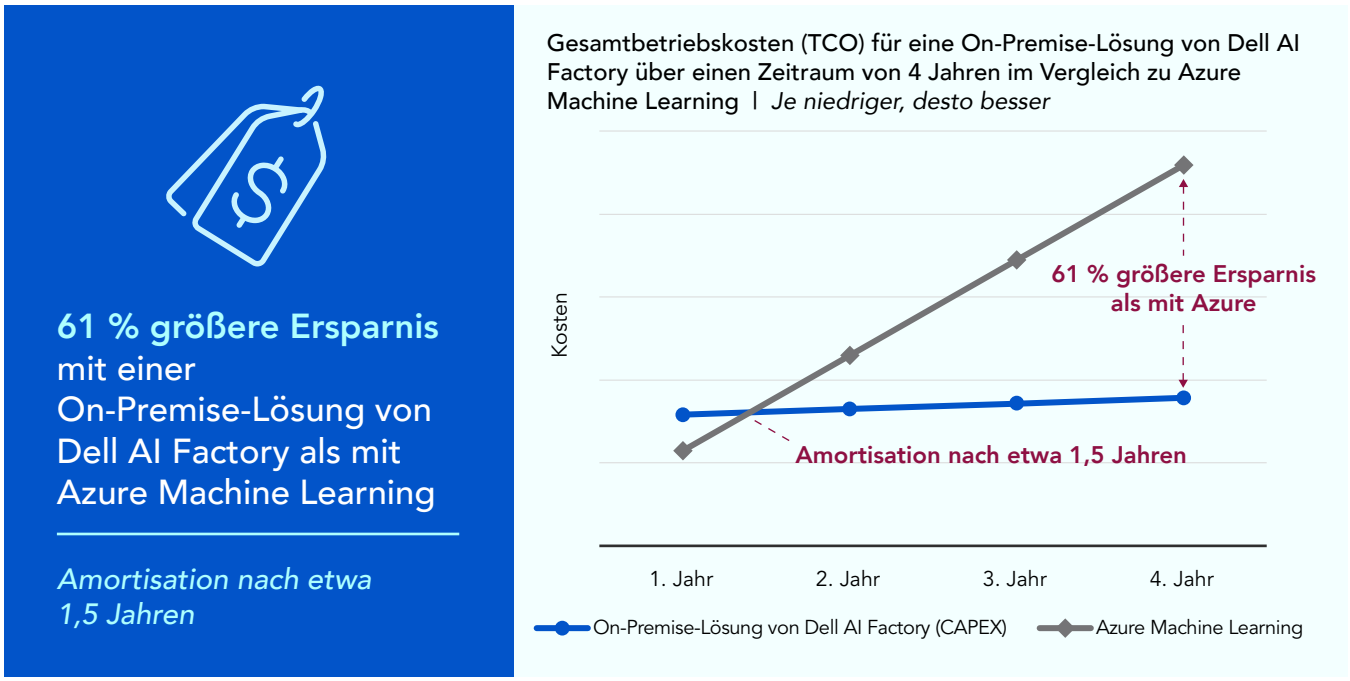


Abbildung 3: Relative Kosten einer On-Premise-Lösung von Dell AI Factory und einer Azure Machine Learning-Lösung über vier Jahre.

Sparen Sie mit Dell APEX-Abonnements

Einige Unternehmen halten die langfristige Bindung, die mit einer herkömmlichen On-Premise-Lösung einhergeht, möglicherweise für unerschwinglich. Aus diesem Grund bieten wir Dell APEX-Abonnements an. Dell kann Hardware im Rechenzentrum Ihres Unternehmens installieren, sodass sie wie bei der herkömmlichen Lösung vor Ort verbleibt. Sie können sich für eine Laufzeit von 3, 4 oder 5 Jahren und für Rechenressourcen zu einem festgelegten Verbrauchstarif für eine gleichbleibende monatliche Zahlung entscheiden. Wenn Sie mehr als Ihr zugesagtes Volumen benötigen, können Sie die verbleibenden Ressourcen gegen Aufpreis nutzen. Wenn Ihr Abonnement endet, können Sie den Service kündigen und die Hardware zurückgeben, das Abonnement unverändert verlängern oder zu einer Lösung wechseln, die Ihren Anforderungen besser entspricht.[11]

Für unseren Vergleich der TCO haben wir ein Angebot von Dell für dieselbe Hardware erhalten, die wir in unserer herkömmlichen On-Premise-Lösung (CAPEX) von Dell AI Factory berücksichtigt haben, aber auch ein 4-Jahres-Abonnement für Dell APEX mit einer garantierten Nutzungsrate von 75 % hinzugefügt. Die Nutzungsraten der Dell APEX-Abonnements für Server basieren auf der Zeit, die ein Server mit mehr als 5 % CPU-Aktivität in einem Monat verbringt.

Annahmen zu Dell APEX-Abonnements

- Etwa 726 Stunden pro Monat mit einer garantierten Nutzungsrate von 75 % ergeben maximal 544,5 Stunden Serverzeit pro Monat, bevor zusätzliche Ressourcen benötigt werden. Zur Konsistenz mit den anderen Berechnungen haben wir 528 Stunden pro Monat verwendet.
- Das Angebot enthielt auch Pläne für ProDeploy Plus und ProSupport am nächsten Werktag, sodass wir keine Administratorkosten für die Ersteinrichtung einbeziehen.
- Wir haben die gleichen Kosten für Strom, Kühlung und Rack-Platz im Rechenzentrum wie bei unserer herkömmlichen Lösung berücksichtigt.

Wir haben festgestellt, dass Dell APEX-Abonnements, die die Sicherheits- und Kontrollvorteile einer herkömmlichen On-Premise-Lösung mit dem Komfort und der Flexibilität eines verwalteten Services kombiniert, Unternehmen über vier Jahre im Vergleich zu den Cloud-Lösungen, für die wir Preise ermittelt haben, erhebliche Einsparungen ermöglichen könnte.

Wie Abbildung 4 zeigt, Kosten Dell APEX-Abonnements 71 % weniger als die AWS SageMaker-Lösung. Die AWS-Lösung kostet über einen Zeitraum von vier Jahren 3,5-mal mehr als Dell APEX-Abonnements. Mit Dell APEX-Abonnements zahlen Sie vom ersten Tag an weniger und sparen letztendlich über einen Zeitraum von vier Jahren erheblich.

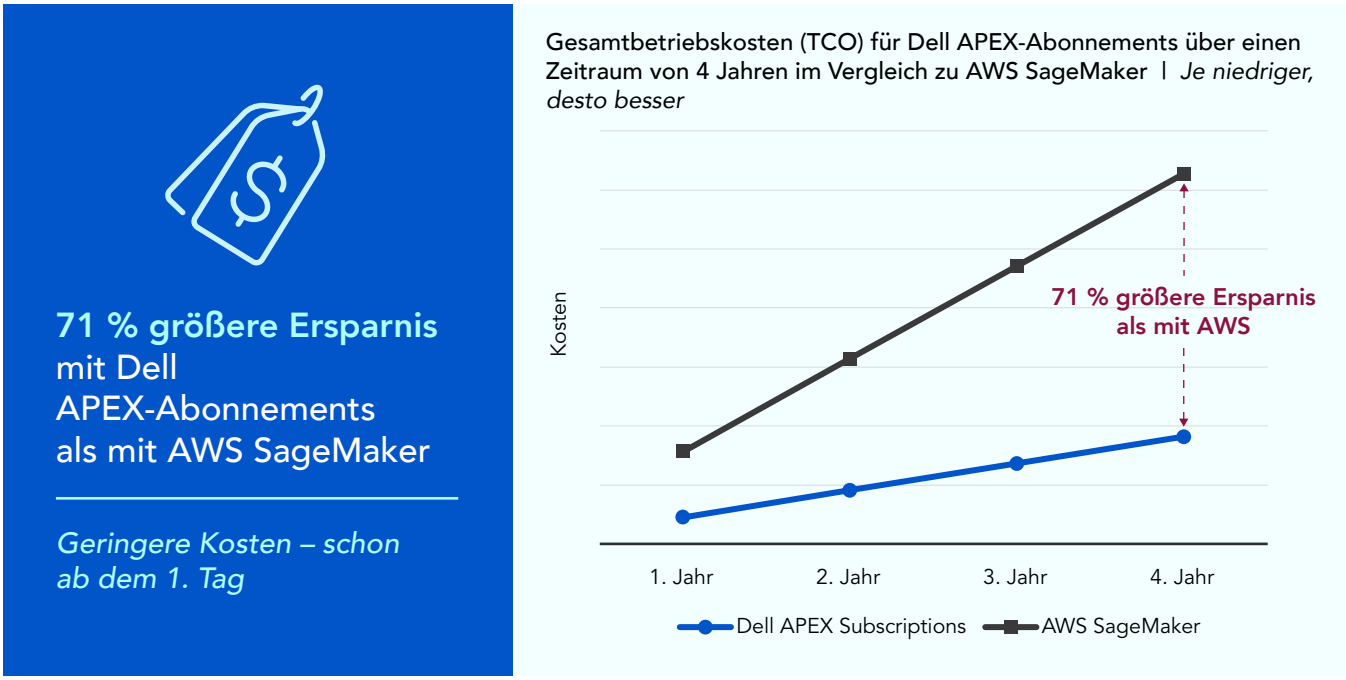


Abbildung 4: Relative Kosten von Dell APEX-Abonnements und einer AWS SageMaker-Lösung über vier Jahre.

Wie Abbildung 5 zeigt, bieten Dell APEX-Abonnements auch im Vergleich zur Azure Machine Learning-Lösung erhebliche Kosteneinsparungen und senken die Gesamtbetriebskosten über vier Jahre um 60 Prozent. Das bedeutet, dass die Azure Machine Learning-Lösung über einen Zeitraum von vier Jahren 2,5-mal so viel kosten würde wie die Nutzung von Dell APEX-Abonnements. Diese Ergebnisse zeigen, dass budgetbewusste Unternehmen, die GenAI implementieren möchten, ihre Anforderungen mit Dell APEX-Abonnements gut erfüllen können, anstatt diese potenziell sensiblen Workloads in der Cloud zu hosten. Darüber hinaus zahlen Kunden, wie im vorherigen Vergleich, vom ersten Tag mit Dell APEX-Abonnements an weniger, was zu deutlich geringeren Kosten über die Laufzeit von vier Jahren führt.

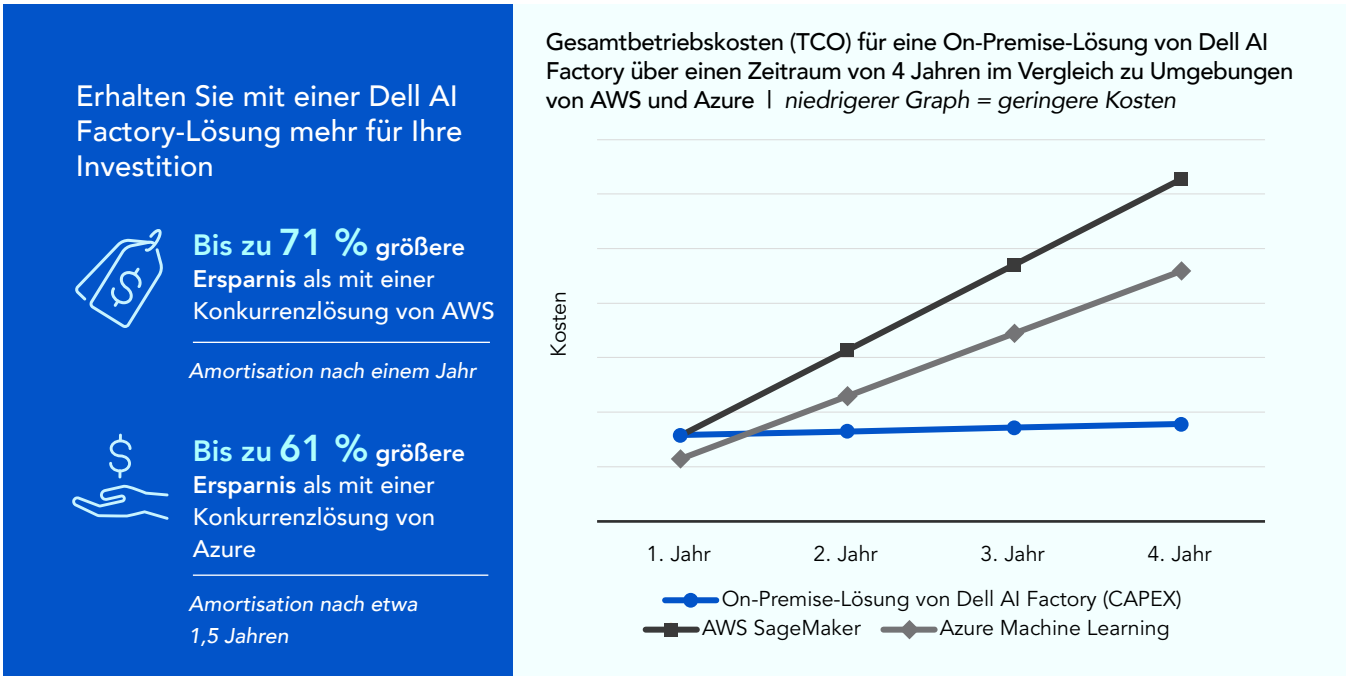


Abbildung 5: Relative Kosten von Dell APEX-Abonnements und einer Azure Machine Learning-Lösung über vier Jahre.

Zusätzliche Überlegungen zur On-Premise-Ausführung von GenAI-Workloads im Vergleich zur Cloud

Die Speicherung großer Mengen an Nutzerdaten in der Public Cloud, damit sie von LLMs auf Plattformen von Drittanbietern erfasst und verfeinert werden können, kann erhebliche Sicherheitsrisiken mit sich bringen, darunter:

- Daten werden öffentlichen Schnittstellen ausgesetzt, auf die AngreiferInnen zugreifen könnten. CrowdStrike hat beispielsweise eine solche Sicherheitslücke entdeckt, die es ihnen ermöglichte, AWS S3-Buckets basierend auf DNS-Anfragen zu finden.[12]
- Die Komplexität steigt, was zu Fehlkonfigurationen führen kann, da Ihre IT-Teams mit mehreren Services und Cloud-Anbieter jonglieren, die Standardwerte und Einstellungen regelmäßig ändern.
- Menschliche Fehler bei der Verwendung cloudbasierter APIs, die sensible Daten offenlegen könnten, nehmen zu.[13]

Die Bereitstellung von LLMs in privaten Netzwerken kann diese Risiken mindern, da interne Lösungen eine bessere Kontrolle über Datenströme, Netzwerkisolierung, API-Kontrollen, die Einhaltung von Datenvorschriften, die Leistungsoptimierung und vieles mehr bieten. Darüber hinaus haben NutzerInnen, die LLMs lokal ausführen, mehr Kontrolle über den gesamten Stack, von der Hardware, auf der das LLM ausgeführt wird, bis zum Modell und den Daten, die die Lösung ermöglichen. AdministratorInnen können zusätzliche Trainings nutzen, um sicherzustellen, dass lokale LLMs spezifische Bestimmungen einhalten. In der Cloud haben NutzerInnen weniger Kontrolle über die zugrundeliegende Infrastruktur und Implementierung.[14] Darüber hinaus können On-Premise-Lösungen die Kosten vorhersehbar halten, sodass sie nicht von Monat zu Monat variieren.

Daten-Storage und -Übertragung sind ein großer Teil der LLM-Anwendungsanforderungen. Das Training eines LLM erfordert große Datenmengen, die irgendwo gespeichert und dann zur Verarbeitung zwischen Storage- und Rechenressourcen verschoben werden müssen. Wenn die Geräte, Datenbanken und Benutzerdaten, die Ihr LLM speisen, bereits ihre Daten lokal speichern, könnten die Kosten für die Übertragung dieser Daten in die Cloud und die benötigte Netzwerkbandbreite hoch sein.

Llama 3-Modelle

Llama 3, das für Large Language Model Meta AI steht, ist eine kostenlose und vielseitige Sprachverarbeitungstechnologie, die von Meta entwickelt wurde. Es handelt sich um ein vortrainiertes Large Language Model (LLM) mit zwei Hauptmodellgrößenvarianten basierend auf der Anzahl der Parameter (8 Mrd. und 70 Mrd.), die eine Vielzahl von Anwendungsfällen unterstützen.[15] Meta hat Llama 3 mit einem „... neuen hochwertigen menschlichen Evaluierungssatz trainiert. Dieser Evaluierungssatz enthält 1.800 Prompts, die 12 wichtige Anwendungsfälle abdecken: um Rat fragen, Brainstorming, Klassifizierung, Beantwortung geschlossener Fragen, Codierung, kreatives Schreiben, Extraktion, Verkörperung einer Figur/Persönlichkeit, Beantwortung offener Fragen, Argumentation, Umschreiben und Zusammenfassen."[16]

Weitere Informationen zu Llama 3 finden Sie unter https://ai.meta.com/blog/meta-llama-3/.

Fazit

Unsere Recherche hat ergeben, dass Sie mit einem On-Premise-Hosting von GenAI-Workloads, entweder mit einer herkömmlichen Dell Lösung oder mit verwalteten Dell APEX-Abonnements, Ihre GenAI-Kosten über einen Zeitraum von vier Jahren im Vergleich zum Hosting dieser Workloads in der Cloud erheblich senken können. Tatsächlich haben wir festgestellt, dass eine On-Premise-Lösung von Dell AI Factory die Kosten im Vergleich zu einer vergleichbaren AWS SageMaker-Lösung um bis zu 71 % und im Vergleich zu einer vergleichbaren Azure ML-Lösung um bis zu 61 % senken kann. Diese Ergebnisse zeigen, dass Unternehmen, die GenAI implementieren und die daraus resultierenden geschäftlichen Vorteile nutzen möchten, viele Vorteile in einer On-Premise-Lösung von Dell AI Factory finden können, ganz gleich, ob sie sich dafür entscheiden, sie selbst zu erwerben und zu verwalten, oder Dell APEX-Abonnements nutzen. Durch die Wahl einer On-Premise-Lösung von Dell AI Factory könnte Ihr Unternehmen gegenüber dem Hosting von GenAI in der Cloud erheblich sparen. Gleichzeitig haben Sie die Kontrolle über die Sicherheit und den Datenschutz Ihrer Daten sowie alle Updates und Änderungen an der Umgebung und stellen gleichzeitig sicher, dass Ihre Umgebung konsistent verwaltet wird.

---

1. CIO, „How to get gen AI Spend under Control", abgerufen am 7. April 2025, https://www.cio.com/article/3478467/how-to-get-gen-ai-spend-under-control.html.
2. AWS, „AWS Pricing Calculator", abgerufen am 16. April 2025, https://calculator.aws/#/.
3. AWS, „Machine Learning Savings Plans", abgerufen am 16. April 2025, https://aws.amazon.com/savingsplans/ml-pricing/.
4. StackOverflow, „Why should preprocessing be done on CPU rather than GPU?", abgerufen am 16. April 2025, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu.
5. Hugging Face, „Model Memory Requirements", abgerufen am 16. April 2025, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
6. AWS, „Training large language models on Amazon SageMaker: Best practices", abgerufen am 16. April 2025, https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/.

7. AWS, „Machine Learning Savings Plans“, abgerufen am 16. April 2025, https://aws.amazon.com/savingsplans/ml-pricing/.
8. Hinweis: AWS hat bestätigt, dass die ml.p5.48xlarge-Instanz bei einer Laufzeit von drei Jahren enthalten ist. Zum Zeitpunkt dieser Recherche war sie nicht im Sparplanrechner aufgeführt. Wir haben die Kosten für die ml.p5-Instanz anhand des Prozentsatzes der Einsparungen geschätzt, der für die Version p5.48xlarge ohne maschinelles Lernen aufgeführt ist, wie unter https://aws.amazon.com/savingsplans/compute-pricing/ aufgeführt.
9. Microsoft, „Azure Pricing Calculator“, abgerufen am 16. April 2025, https://azure.microsoft.com/en-us/pricing/calculator/.
10. Microsoft, „Azure Machine Learning pricing“, abgerufen am 16. April 2025, https://azure.microsoft.com/en-us/pricing/details/machine-learning/.
11. Dell, „Dell APEX-Abonnements“, abgerufen am 16. April 2025, https://www.dell.com/en-us/dt/apex/subscriptions.htm.
12. CrowdStrike, „12 Cloud Security Issues: Risks, Threats, and Challenges“, abgerufen am 16. April 2025, https://www.crowdstrike.com/cybersecurity-101/cloud-security/cloud-security-risks-threats-challenges/.
13. CrowdStrike, „12 Cloud Security Issues: Risks, Threats, and Challenges“.
14. DataCamp, „The Pros and Cons of Using LLMs in the Cloud Versus Running LLMs Locally“, abgerufen am 16. April 2025, https://www.datacamp.com/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally.
15. Meta, „Introducing Meta Llama 3: The Most Captly Available LMM to date“, abgerufen am 16. April 2025, https://ai.meta.com/blog/meta-llama-3/.
16. Meta, „Introducing Meta Llama 3: The Most Capably Available LMM to date“.

Lesen Sie den wissenschaftlichen Hintergrund dieses Berichts ▶

▶ Lesen Sie die Originalversion dieses Berichts in englischer Sprache

Principled Technologies®

Facts matter.®

Dieses Projekt wurde in Auftrag gegeben von Dell Technologies.