Investitionen in GenAI: Kosten-Nutzen-Analyse von Dell On-Premise-Bereitstellungen im Vergleich zu ähnlichen AWS- und Azure-Bereitstellungen

In der Technologiewelt ist generative KI (GenAI) der neueste Trend. Während Unternehmen auf der ganzen Welt damit beginnen, zu untersuchen, wie GenAI ihre Geschäftsziele voranbringen kann, stehen sie vor zahlreichen Hürden bei der Implementierung einer GenAI-Lösung, die ihre spezifischen Anforderungen erfüllt. Zu den größten Herausforderungen gehört die genaue Bestimmung, wie viel eine maßgeschneiderte GenAI-Lösung kosten wird und ob sie vor Ort oder in der Cloud bereitgestellt werden soll. Gartner-Analyst Frances Karamouzis hat es so formuliert: „Kosten sind eine der größten Bedrohungen für den Erfolg von KI und generativer KI. Mehr als die Hälfte der Organisationen gibt ihre Bemühungen aufgrund von Fehlern bei der Schätzung und Berechnung der Kosten auf.“[1]

Um Unternehmen einen direkten Überblick über die Gesamtkosten für die Bereitstellung und das Management von GenAI-Workloads zu geben, einschließlich Modellfeinabstimmung und Inferenz, haben wir uns die ungefähren Kosten von zwei On-Premise-Lösungen von Dell™ mit PowerEdge™ R660- und PowerEdge XE9680-Hardware – einer herkömmlichen Lösung und einer abonnementbasierten Dell APEX Pay-Per-Use-Lösung – sowie vergleichbaren Amazon Web Services (AWS) SageMaker- und Microsoft Azure Machine Learning-Lösungen über einen Zeitraum von drei Jahren angesehen. Unseren Berechnungen zufolge war die Dell APEX Pay-Per-Use-Lösung über drei Jahre hinweg die kosteneffizienteste der Lösungen, die wir verglichen haben. Die wettbewerbsfähigen Cloud-Lösungen von AWS und Azure kosten bis zu 3,81-mal so viel wie die abobasierte Dell APEX Pay-Per-Use-Lösung. Im Vergleich zur herkömmlichen On-Premise-Lösung von Dell würden die von uns berechneten Preise für die Cloud-Lösungen von AWS und Azure bis zu 2,88-mal so hoch ausfallen. Lesen Sie weiter, um zu erfahren, wie GenAI Ihrem Unternehmen helfen kann und wie wir unsere Gesamtbetriebskosten (Total Cost of Ownership, TCO) berechnet haben.

Zahlen Sie weniger für GenAI mit einer Dell APEX Pay-Per-Use-Lösung und einer Dell On-Premises-GenAI-Lösung

Wettbewerbsfähige AWS- und Azure-Cloud-Lösungen kosten mehr:

Bis zu 3,81-mal höhere Kosten als bei einer Dell APEX Pay-Per-Use-Lösung

Bis zu 2,88-mal höhere Kosten als bei der herkömmlichen Dell On-Premise-Lösung

Die Vorteile der Verwendung vortrainierter Modelle für generative KI

Modelle der künstlichen Intelligenz (KI) sind Systeme, die Aspekte der menschlichen Intelligenz oder des menschlichen Verhaltens imitieren sollen. Unternehmen und Privatpersonen nutzen GenAI-Tools (wie ChatGPT und DALL-E), um Inhalte wie Text, Audio, Videos, Bilder, Code und Simulationen sowie komplexere Ausgaben wie personalisierte Marketinginhalte, kundenspezifische Anwendungen und Software zu erzeugen.[2] Ein Unternehmen, das generative KI in seinen Betrieb integrieren möchte, kann zwischen der Verwendung eines vortrainierten Modells oder der Erstellung eines eigenen Modells von Grund auf wählen. Ein vortrainiertes Modell, z. B. Llama 2, wurde bereits auf einem Basisdatensatz für die vorgesehene Verwendung des Modells trainiert. Von dort aus können Unternehmen das Modell mithilfe ihrer spezifischen Daten optimieren und so den Prozess der Erstellung eines Modells beschleunigen, das auf ihre Daten und Anwendungsfälle zugeschnitten ist.[3] Einer Quelle zufolge könnte die Verwendung vortrainierter Modelle die Zeit bis zu einem funktionsfähigen KI-Modell um bis zu ein Jahr verkürzen und gleichzeitig Einsparungen in Höhe von Hunderttausenden von Dollar ermöglichen.[4]

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 2 13B 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 andere Arbeiten 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 29. März 2024 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
Herkömmliche On-Premise-Lösung			
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	8x NVIDIA H100	
Modellfeinabstimmung			
Schlussfolgerung			
Verwaltete On-Premise Dell APEX Pay-Per-Use-Lösung			
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	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	k. A.	
Modellfeinabstimmung	4x ND96amsr A100 v4	8x NVIDIA A100	
Schlussfolgerung	4x ND96amsr A100 v4	8x NVIDIA A100	

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 2 13B 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 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 werden die Kosten für den Platzbedarf im physischen Rechenzentrum sowie für die Stromversorgung und Kühlung berücksichtigt. Diese Kosten werden in die Instanzkosten für die Cloud-Lösungen einbezogen.

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

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

Annahmen für GenAI-Kostenvergleiche

- Wir gehen davon aus, dass es 22 Arbeitstage pro Monat gibt, wobei Workloads über Nacht 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 Data Scientists 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 EntwicklerInnen 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 herkömmlichen On-Premise-Lösung von Dell

Wir haben uns an Dell gewandt und um ein von Dell empfohlenes Preisangebot für unsere herkömmliche On-Premise-Lösung gebeten. 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 drei Jahre begrenzt haben, die meisten Server aber drei bis fünf Jahre funktionieren und über die drei 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 drei Jahren sowie die Verwaltungskosten für die Wartung der Ausstattung für drei 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.[5,6] 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.[7,8]
- Wir haben 7 TB pro Monat an EBS-Storage zu den ml.r5.16xlarge-Instanzen 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.[9] 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.[10] Wir haben die AWS-Konfiguration mit dem Preis für eine Laufzeit von drei Jahren berechnet.[11] 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.

Vergleich der herkömmlichen On-Premise-Lösung von Dell mit AWS SageMaker

Anhand der oben genannten Annahmen für beide Lösungen haben wir einen Vergleich der TCO über drei Jahre berechnet. Unsere Berechnungen zeigen, dass die Wahl der herkömmlichen Dell PowerEdge-On-Premise-Lösung

zur Ausführung von GenAI-Workloads im Vergleich zur Ausführung derselben Workloads auf AWS SageMaker echte Einsparungen bieten könnte.

Wie Abbildung 1 zeigt, haben wir berechnet, dass die AWS SageMaker-Lösung bis zu 2,88-mal so viel kosten könnte wie die On-Premise-Lösung von Dell. Hinweis: Weitere Informationen hierzu und zu allen nachfolgenden Kosten finden Sie im wissenschaftlichen Hintergrund des Berichts.

AWS SageMaker könnte fast dreimal so viel kosten wie eine ähnliche Dell Lösung

Amortisation nach etwa einem Jahr

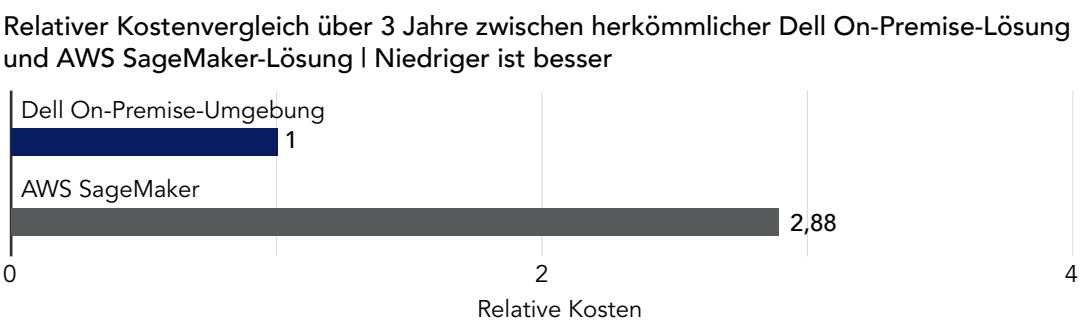

Abbildung 1: Relative Kosten einer GenAI Dell On-Premise-Lösung und einer AWS SageMaker Lösung über drei Jahre.

Angesichts der über 2,8-mal höheren Kosten über einen Zeitraum von drei Jahren können NutzerInnen davon ausgehen, dass sich die Kosten für Hosting, Kühlung und Management einer On-Premise-Lösung im Vergleich zu den Kosten für AWS-Hosting bereits nach einem Jahr fast amortisiert haben.

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 3-Jahres-Sparplans mit Reservierung entschieden.[12] 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
ND96asmr A100 v4	4	Modellfeinabstimmung	792
ND96asmr A100 v4	4	Inferenz	264

Annahmen zu Preisdetails für Azure Machine Learning

- Der Azure Machine Learning-Service bot keine Instanz mit den NVIDIA H100-GPUs an. Daher haben wir vier A100-GPU-Instanzen ausgewählt, um eine ähnliche Leistung zu erreichen.[13]
- 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.

Azure bietet „Pay as you go"-Preise, Azure-Sparpläne und Azure-Reservierungsoptionen für den Machine Learning-Service.[14] Azure bot im Gegensatz zu AWS keine günstigere Vorabzahlungsoption an. Um die Preisgestaltung der AWS-Umgebung bestmöglich abzubilden, haben wir uns für den 3-Jahres-Reservierungsplan entschieden.

Vergleich der On-Premise-Lösung von Dell mit Azure ML

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

Tatsächlich schätzen wir, dass die Gesamtkosten der Azure Machine Learning-Lösung über drei Jahre 2,72-mal so hoch wären wie bei der herkömmlichen On-Premise-Lösung von Dell (siehe Abbildung 2). Diese Ergebnisse zeigen, dass Sie Ihr GenAI-Budget mit einer herkömmlichen Dell Lösung auf ein vernünftiges Maß beschränken können, sodass Sie die Einsparungen für Innovationen an anderer Stelle nutzen können.

Azure Machine Learning könnte fast dreimal so viel kosten wie eine ähnliche Dell Lösung

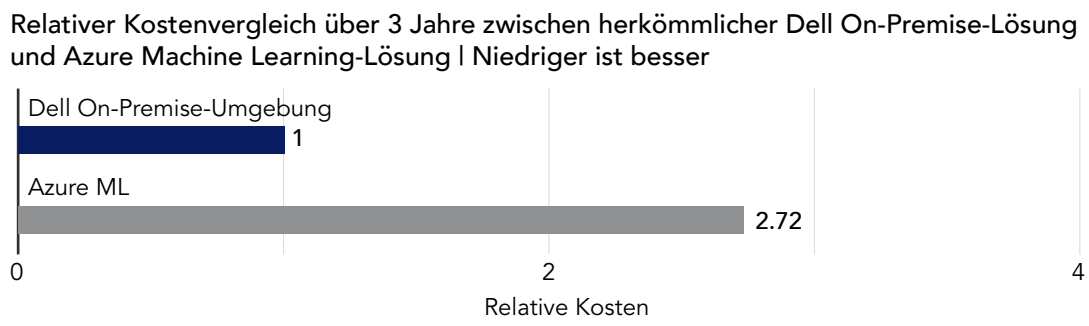

Abbildung 2: Relative Kosten einer On-Premise-Lösung für generative KI von Dell und einer Azure Machine Learning-Lösung über drei Jahre.

Ähnlich wie bei der AWS-Lösung, die über einen Zeitraum von drei Jahren mehr als 2,7-mal so teuer ist, können Dell On-Premise-KundInnen damit rechnen, dass sich die Kosten im Vergleich zu den Azure-Preisen nach einem Jahr fast amortisiert haben.

Sparen Sie noch mehr, indem Sie eine Dell APEX Pay-Per-Use-Lösung abonnieren

Einige Unternehmen halten die langfristige Bindung, die mit einer herkömmlichen On-Premise-Lösung einhergeht, möglicherweise für unerschwinglich. Aus diesem Grund bietet Dell eine Dell APEX Pay-Per-Use-Lösung 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 entscheiden und eine Dell APEX Pay-Per-Use-Lösung für Rechenressourcen zu einem festgelegten Verbrauchstarif für eine gleichbleibende monatliche Zahlung nutzen. 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.[15]

Für unseren Vergleich der TCO haben wir ein Angebot von Dell für dieselbe Hardware erhalten, die wir in unserer herkömmlichen On-Premise-Umgebung berücksichtigt haben, aber auch ein 3-Jahres-Abonnement für eine Dell APEX Pay-Per-Use-Lösung mit einer garantierten Nutzungsrate von 75 % hinzugefügt. Die Nutzungsraten der Dell APEX Pay-Per-Use-Lösung für Server basieren auf der Zeit, die ein Server mit mehr als 5 % CPU-Aktivität in einem Monat verbringt. Unsere Annahmen lauten wie folgt:

Annahmen zur Dell APEX Pay-Per-Use-Lösung

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

Wie Abbildung 3 zeigt, könnte die AWS SageMaker-Lösung 3,81-mal so viel kosten wie die Dell APEX Pay-Per-Use-Lösung.

AWS SageMaker könnte fast viermal so viel kosten wie eine ähnliche Dell APEX Pay-Per-Use-Lösung

Amortisation nach weniger als einem Jahr

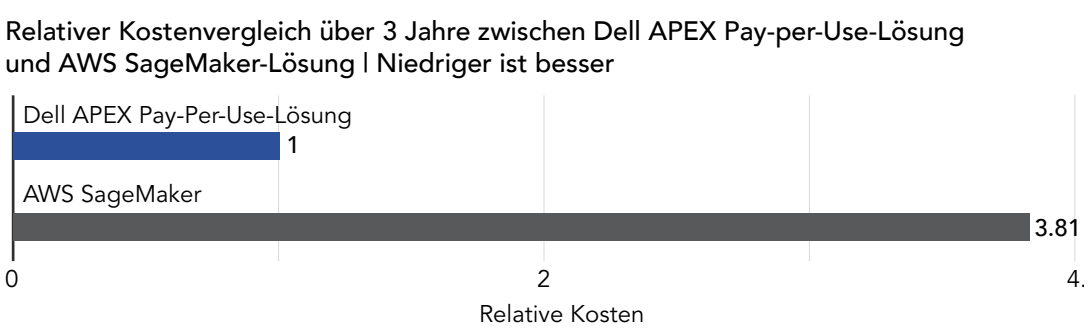

Abbildung 3: Relative Kosten einer GenAI Dell APEX Pay-Per-Use-Lösung und einer AWS SageMaker-Lösung über drei Jahre.

Da die Kosten über einen Zeitraum von drei Jahren mehr als 3,8-mal höher sind als bei einer Dell APEX Pay-Per-Use-Lösung, können NutzerInnen davon ausgehen, dass sich ihre Investition bereits vor Ablauf eines Jahres amortisiert hat. Die Kosteneinsparungen einer Dell APEX-Pay-Per-Use-Lösung waren ähnlich wie bei der Azure Machine Learning-Lösung, die 3,60-mal so viel kostet wie die Dell APEX Pay-Per-Use-Lösung (siehe Abbildung 4).

Azure Machine Learning könnte fast viermal so viel kosten wie eine ähnliche Dell APEX Pay-per-Use-Lösung

Amortisation nach weniger als einem Jahr

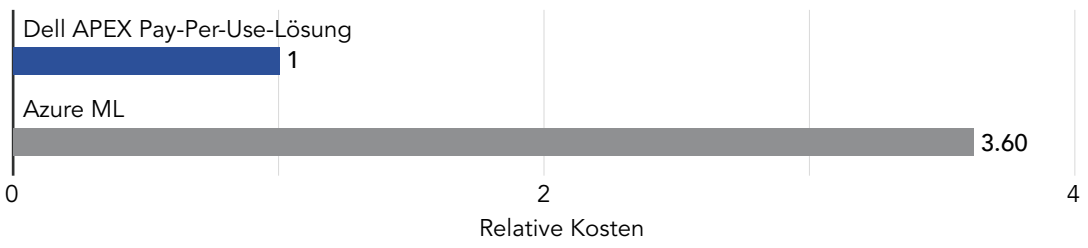

Abbildung 4: Relative Kosten einer GenAI Dell APEX Pay-Per-Use-Lösung und einer Azure Machine Learning-Lösung über drei Jahre.

Diese Ergebnisse zeigen, dass budgetbewusste Unternehmen, die GenAI implementieren möchten, ihre Anforderungen gut erfüllen können, indem sie sich für eine Dell APEX-Pay-Per-Use-On-Premise-Lösung entscheiden, anstatt diese potenziell sensiblen Workloads in der Cloud zu hosten. Darüber hinaus können KundInnen, wie beim AWS-Vergleich, davon ausgehen, dass sich die Investition im Vergleich zur Azure-Lösung bereits vor Ablauf eines Jahres amortisiert hat.

Zusätzliche Überlegungen zu GenAI-Workloads

Weitere Vorteile der Wahl von On-Premise-Lösungen im Vergleich zur Cloud

Cloud-Hosting bietet zwar Vorteile wie Skalierbarkeit und Flexibilität, aber neben den Kosten sollten Sie vor dem Hosten Ihrer LLMs in einer Public Cloud mehrere Aspekte berücksichtigen. Eines der größten Bedenken ist das Risiko, das mit dem Hosting großer Mengen an Nutzerdaten auf einer Drittanbieterplattform verbunden ist. Viele LLMs sammeln Nutzerdaten, um ihre Modelle zu verbessern, und diese Daten müssen irgendwo aufbewahrt werden, damit die Modelle darauf zugreifen können. Die Speicherung dieser sensiblen Daten in der Cloud kann zum Beispiel folgende Risiken haben:

- 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.[16]
- 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.[17]

Private LLMs reduzieren diese Risiken, da NutzerInnen in der Regel mehr Kontrolle über ihre Rechenzentren und damit ihre Datenströme, Netzwerkisolierung, API-Kontrollen und mehr haben. 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.[18] 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 gespeichert und dann vom Storage zur Verarbeitung an Compute-Ressourcen übertragen 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.

Das Dell AI-Portfolio

Bei der Implementierung von KI-Workloads geht es natürlich um mehr als nur die Ermittlung der Gesamtkosten der Lösung. Sie müssen bestimmen, welches Modell am besten für Ihre Anforderungen geeignet ist, und eine Lösung entwickeln, die die Menge der vorhandenen Daten verarbeiten kann. Nach der Auswahl des richtigen Modells und der Entscheidung für eine Lösung stellen sich Personalfragen. Sie müssen MitarbeiterInnen schulen oder einstellen, um

Llama 2-Modelle

Llama 2, 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 drei Hauptmodellgrößenvarianten basierend auf der Anzahl der Parameter (7 Mrd., 13 Mrd. und 70 Mrd.), die jeweils unterschiedliche Leistungsmerkmale bieten.[22] Meta hat Llama 2 mit einem Ansatz zum bestärkenden Lernen trainiert, um NutzerInnen familienfreundliche Ergebnisse zu liefern, mit dem Ziel, sich mit menschlichen Entscheidungen und Präferenzen vertraut zu machen.[23] Weitere Informationen zu Llama 2 finden Sie unter https://llama.meta.com/llama2/.

sicherzustellen, dass Ihr IT-Personal Data Science versteht und in der Lage ist, Ihre KI-Lösung angemessen zu trainieren und auszuführen.

Dell bietet ein umfassendes KI-Portfolio, das Ihnen mehr als nur Kosteneinsparungen bei der Hardware bietet. Sie erhalten auch die Tools, die Sie benötigen, um Herausforderungen während Ihrer KI-Implementierung zu bewältigen. Das KI-Portfolio von Dell umfasst Hardware, Datenmanagementservices, professionelle Dienstleistungen, KI-Trainings, Referenzarchitekturen und viele Partnerschaften mit anderen KI-orientierten Unternehmen.[19] Dell kann Sie bei der erfolgreichen Entwicklung, Planung und Implementierung einer KI-Lösung unterstützen, die Ihren Anforderungen entspricht. Weitere Informationen dazu, wie das Dell KI-Portfolio im Vergleich zu den Angeboten von MitbewerberInnen abschneidet, finden Sie in unseren Berichten, in denen das Dell KI-Portfolio mit denen von Supermicro[20] und HPE verglichen wird.[21]

Entscheidung

Der Einstieg in die Welt der GenAI kann Ihrer Organisation viele Vorteile bringen, aber zunächst muss überlegt werden, wie diese GenAI-Workloads am besten implementiert werden können. Ganz gleich, ob Ihre KI-Ziele darin bestehen, einen Chatbot für OnlinebesucherInnen zu erstellen, Marketingmaterialien zu generieren, das Troubleshooting zu unterstützen oder andere Aufgaben zu bewältigen – die Implementierung einer KI-Lösung erfordert eine sorgfältige Planung und Entscheidungsfindung. Eine wichtige Entscheidung besteht darin, ob GenAI in der Cloud gehostet oder Ihre Daten lokal aufbewahrt werden sollen. Herkömmliche On-Premise-Lösungen können überragende Sicherheit und Kontrolle bieten, was bei der Verarbeitung großer Mengen potenziell sensibler Daten von erheblicher Bedeutung ist. Aber belastet die Unterstützung einer GenAI-Lösung vor Ort das IT-Budget einer Organisation?

Unsere Recherche hat ergeben, dass das Wertversprechen genau das Gegenteil ist. Wenn Sie GenAI-Workloads lokal hosten, entweder mit einer herkömmlichen Dell Lösung oder mit einer verwalteten Dell APEX-Pay-Per-Use-Lösung, können Sie Ihre GenAI-Kosten über einen Zeitraum von drei Jahren im Vergleich zum Hosting dieser Workloads in der Cloud erheblich senken. Tatsächlich haben wir festgestellt, dass eine vergleichbare AWS SageMaker-Lösung bis zu 3,8-mal so viel und eine Azure ML-Lösung bis zu 3,6-mal so viel kosten würde wie GenAI auf einer Dell APEX Pay-Per-Use-Lösung. 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 finden können, ganz gleich, ob sie sich dafür entscheiden, sie selbst zu erwerben und zu verwalten oder eine abobasierte Dell APEX Pay-Per-Use-Lösung auszuwählen. Durch die Wahl einer On-Premise-Lösung von Dell 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. LinkedIn, Beitrag von Jacquecine Burrell, abgerufen am 19. April 2024, https://www.linkedin.com/posts/jacquecine-burrell-88094714_meet-frances-karamouzis-at-gartner-data-activity-7173514796506554368-9Sie.
2. Deloitte, „The State of Generative AI in the Enterprise", abgerufen am 3. April 2024, https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-generative-ai-in-enterprise.html.
3. OneAI, „The Future is Pre-trained: The Shortcut to AI Mastery", abgerufen am 3. April 2024, https://oneai.com/learn/pre-trained-ai-model.
4. NVIDIA, „What Is a Pretrained AI Model?", abgerufen am 3. April 2024, https://blogs.nvidia.com/blog/what-is-a-pretrained-ai-model/.
5. AWS, „AWS Pricing Calculator", abgerufen am 3. April 2024, https://calculator.aws/#/.
6. AWS, „Machine Learning Savings Plans", abgerufen am 3. April 2024, https://aws.amazon.com/savingsplans/ml-pricing/.
7. StackOverflow, „Why should preprocessing be done on CPU rather than GPU?", abgerufen am 3. April 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu.
8. Hugging Face, „Model Memory Requirements", abgerufen am 3. April 2024, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
9. AWS, „Training large language models on Amazon SageMaker: Best practices", abgerufen am 3. April 2024, https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/.
10. AWS, „Machine Learning Savings Plans", abgerufen am 3. April 2024, https://aws.amazon.com/savingsplans/ml-pricing/.
11. 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.
12. Microsoft, „Azure Pricing Calculator", abgerufen am 3. April 2024, https://azure.microsoft.com/en-us/pricing/calculator/.
13. Comet, „Comparison of NVIDIA A100, H100 + H200 GPUs", abgerufen am 3. April 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.
14. Microsoft, „Azure Machine Learning pricing", abgerufen am 3. April 2024, https://azure.microsoft.com/en-us/pricing/details/machine-learning/.
15. Dell, „Dell APEX Flex on Demand", abgerufen am 3. April 2024, https://www.delltechnologies.com/partner/en-us/partner/apex-flex-on-demand.htm.
16. CrowdStrike, „12 Cloud Security Issues: Risks, Threats, and Challenges", abgerufen am 3. April 2024, https://www.crowdstrike.com/cybersecurity-101/cloud-security/cloud-security-risks-threats-challenges/.
17. CrowdStrike, „12 Cloud Security Issues: Risks, Threats, and Challenges".
18. DataCamp, „The Pros and Cons of Using LLMs in the Cloud Versus Running LLMs Locally", abgerufen am 3. April 2024, https://www.datacamp.com/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally.
19. Dell, „Dell AI Solutions", abgerufen am 29. April 2024, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#footnote-ref1&tab0=0.
20. Principled Technologies, „Finding the path to AI success with the Dell AI portfolio", abgerufen am 29. April 2024, https://facts.pt/q9p46K9.
21. Principled Technologies, „Meeting the challenges of AI workloads with the Dell AI portfolio", abgerufen am 29. April 2024, https://facts.pt/zPmSx4c.
22. Meta, „Llama 2: open source, free for research and commercial use", abgerufen am 3. April 2024, https://llama.meta.com/llama2/.
23. Pavan Belagatti, „Unpacking Meta's Llama 2: The Next Leap in Generative AI", abgerufen am 3. April 2024, https://www.singlestore.com/blog/a-complete-beginners-guide-to-llama2/

Wissenschaftlicher Hintergrund des Berichts

In diesem Abschnitt finden Sie alle Ergebnisse unserer Tests. Außerdem beschreiben wir hier die Lösungen, die wir zum Testen verwendet haben, sowie unsere Testmethodik.

Unsere Recherche endete am 29. März 2024. Die Ergebnisse in diesem Bericht spiegeln Konfigurationen wider, deren Zusammenstellung und Preisgestaltung wir am 29. März 2024 oder früher abgeschlossen haben. Es ist unvermeidlich, dass diese Konfigurationen und ihre Kosten zum Zeitpunkt der Veröffentlichung dieses Berichts nicht mehr aktuell sind.

Systeminformationen

Dell On-Premise-Lösungen

Die herkömmlichen und verwalteten On-Premise-Lösungen von Dell umfassen beide die folgende Hardware:

- 3 x PowerEdge R660-Haupt-Node
- 2 x PowerEdge XE9680-GPU-Worker-Node
- 2 x PowerSwitch S5232-ON-Netzwerkinfrastruktur
- 1 x PowerSwitch N3200-ON OOB-Management

Tabelle 1: Detaillierte Konfigurationsinformationen für jeden PowerEdge XE9680-GPU-Worker-Node.

Konfigurationsinformationen	Dell PowerEdge XE9680-GPU-Worker-Node
Anzahl der Nodes in der Lösung	2
Gehäuse	
Gehäuse	XE9680-6-HE-Gehäuse mit 8 GPUs und ausschließlich 8 x 2,5"-NVMe-Laufwerken
Prozessor	
Anzahl der Prozessoren	2
Hersteller und Modell	Intel® Xeon® Platinum 8468
Anzahl der Cores (pro Prozessor)	48 Cores und 96 Threads
GPU(s)	
Anzahl der GPUs	1 8-GPU-Baugruppe
Hersteller und Modell	NVIDIA® HGX H100, 8 GPUs, SXM, 80 GB, 700 W, GPUs, Baugruppe
Arbeitsspeichermodul(e)	
Gesamtarbeitsspeicher im System (GB)	1.024
Anzahl der Arbeitsspeichermodule	16
Typ	RDIMM, 4.800 MT/s, Dual Rank
Größe (GB)	64
Storage-Controller	
Hersteller und Modell	BOSS-N1-Controller-Karte + mit 2 M.2 480 GB (RAID 1)
Lokaler Storage (Typ A)	
Gesamtgröße der Laufwerke im System (TB)	44,8

Konfigurationsinformationen	Dell PowerEdge XE9680-GPU-Worker-Node
Anzahl der Laufwerke	7
Laufwerksgröße (TB)	6,4
Informationen zum Laufwerk (Geschwindigkeit, Schnittstelle, Typ)	Enterprise NVMe™, kombinierte Verwendung, AG-Laufwerk U.2 Gen 4 mit Träger
Netzwerkadapter	
Anzahl und Typ der Anschlüsse	2 x 10/25 GbE
Hersteller und Modell	Intel E810-XXV, 2 Anschlüsse, 10/25 GbE SFP28, OCP NIC 3.0
Netzwerkadapter	
Anzahl und Typ der Anschlüsse	2 x 10/100 GbE
Hersteller und Modell	Mellanox ConnectX-6 DX, 2 Anschlüsse, 100 GbE, QSFP56, Netzwerkadapter, gesamte Höhe
Lüfter	
Anzahl der Lüfter	6
Hersteller und Modell	Lüfter mit sehr hoher Performance
Netzteile	
Anzahl der Netzteile	1
Hersteller und Modell	Vollständig redundant, 5 + 1 (oder 3 + 3 FTR), Hot-Plug-PSU, 2.800 W MM HLAC (200–240 V Wechselstrom) Titanium, C22-Connector
Jeweilige Wattleistung (W)	2800
ProSupport und ProDeploy	
ProSupport (5 Jahre)	ProSupport und Vor-Ort-Service am nächsten Arbeitstag
ProDeploy Plus	ProDeploy Plus, Dell Server XE Serie, 5 HE/6 HE
Integrierte Systemverwaltung	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage™ Enterprise Advanced Plus

Tabelle 2: Detaillierte Konfigurationsinformationen für jeden PowerEdge R660-Haupt-Node.

Konfigurationsinformationen	Dell PowerEdge R660 Haupt-Node
Anzahl der Nodes in der Lösung	3
Gehäuse	
Gehäuse	2,5"-Gehäuse mit bis zu 10 Festplatten (SAS/SATA), PERC 11, 1 CPU
Prozessor	
Anzahl der Prozessoren	1
Hersteller und Modell	Intel Xeon Gold 6426Y
Anzahl der Cores (pro Prozessor)	16 Cores und 32 Threads
Arbeitsspeichermodul(e)	
Gesamtarbeitsspeicher im System (GB)	192

Konfigurationsinformationen	Dell PowerEdge R660 Haupt-Node
Anzahl der Arbeitsspeichermodule	12
Typ	16
Größe (GB)	RDIMM, 4.800 MT/s, Single Rank
Storage-Controller	
Hersteller und Modell	BOSS-N1-Controller-Karte + mit 2SED M.2 960 GB (RAID 1)
Lokaler Storage (Typ A)	
Gesamtgröße der Laufwerke im System (TB)	5,76
Anzahl der Laufwerke	6
Laufwerksgröße (TB)	960
Informationen zum Laufwerk (Geschwindigkeit, Schnittstelle, Typ)	vSAS-SSD, leseoptimiert, 12 Gbit/s 512e 2,5"-Hot-Plug, AG-Laufwerk SED, 1 DWPD
Netzwerkadapter	
Anzahl und Typ der Anschlüsse	2 x 10/25 GbE
Hersteller und Modell	NVIDIA ConnectX-6 Lx, 2 Anschlüsse, 10/25 GbE, SFP28, ohne Verschlüsselung,OCP NIC 3.0
Lüfter	
Anzahl der Lüfter	4
Hersteller und Modell	Lüfter mit sehr hoher Performance
Netzteile	
Anzahl der Netzteile	1
Hersteller und Modell	Dual, redundant (1 + 1), Hot-Plug-Netzteil, 1.100 W MM (100–240 V AC) Titanium
Jeweilige Wattleistung (W)	1.100
ProSupport und ProDeploy	
ProSupport (5 Jahre)	ProSupport und Vor-Ort-Service am nächsten Arbeitstag
ProDeploy Plus	ProDeploy Plus PowerEdge R Serie (1 HE/2 HE)
Integrierte Systemverwaltung	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage Enterprise Advanced Plus

AWS SageMaker-Lösungsinstanzen

Tabelle 3: Detaillierte Konfigurationsinformationen für die AWS-Instanzen.

Konfigurationsinformationen	ml.t3.medium (Notebooks)	ml.r5.16xlarge (Verarbeitung)	Ml.p5.48xlarge (Inferenz und Feinabstimmung)
Anzahl der Instanzen	20	2	2
Cloud Service Provider (CSP)	AWS	AWS	AWS
Region	USA Ost (Ohio)	USA Ost (Ohio)	USA Ost (Ohio)
Prozessor			
Anzahl der vCPUs	2	64	192

Konfigurationsinformationen	ml.t3.medium (Notebooks)	ml.r5.16xlarge (Verarbeitung)	Ml.p5.48xlarge (Inferenz und Feinabstimmung)
Arbeitsspeichermodul(e)			
Gesamtarbeitsspeicher im System (GiB)	4.	512	2.048
Storage-Controller			
Hersteller und Modell			
Lokaler Storage (Typ A)			
Anzahl der Laufwerke	1	1	1
Laufwerksgröße (GB)	5 GB	Standard	Standard
Informationen zum Laufwerk (Geschwindigkeit, Schnittstelle, Typ)	EBS	EBS	EBS
GPU			
Anzahl der GPUs	N. z.	N. z.	8
Hersteller und Modell	N. z.	N. z.	NVIDIA H100
Zusätzliche Funktionen	N. z.	N. z.	3.200 Gbit/s Netzwerkbandbreite[1]

Azure Machine Learning-Lösungsinstanzen

Tabelle 4: Detaillierte Konfigurationsinformationen für die Azure-Instanzen.

Konfigurationsinformationen	D2 v2 (Notebooks)	M64 (Verarbeitung)	ND96amsr A100 v4 (Inferenz und Feinabstimmung)
Anzahl der Instanzen	20	1	8
Cloud Service Provider (CSP)	Azure	Azure	Azure
Region	USA Ost 2	USA Ost 2	USA Ost 2
Prozessor			
Anzahl der vCPUs	2	64	96
Arbeitsspeichermodul(e)			
Gesamtarbeitsspeicher im System (GiB)	7	1.000	1.900
Lokaler Storage (Typ A)			
Anzahl der Laufwerke	1	1	1
Laufwerksgröße (GB)	100	7.168	6.400
Informationen zum Laufwerk (Geschwindigkeit, Schnittstelle, Typ)	Temporary	Temporary	Temporary
Anzahl der GPUs	N. z.	N. z.	8
Hersteller und Modell	N. z.	N. z.	NVIDIA A100
Weitere Informationen	N. z.	N. z.	Jede GPU verfügt über eine dedizierte NVIDIA Mellanox HDR InfiniBand-Verbindung mit einer Bandbreite von 200 GB/s[2]

Einführung

Um ein Beispiel für die Kosten von KI-Lösungen zu liefern, haben wir ein KI-Szenario mit dem Open-Source-Modell Llama 2 13B erstellt und die Kosten für die Ausführung der Workload in vier verschiedenen Umgebungen verglichen. Wir haben die Kosten für vier Lösungen dimensioniert und geschätzt:

- Herkömmliche Dell On-Premise-Lösung
- Verwaltete On-Premise-Lösung mit einer Dell APEX Pay-Per-Use-Lösung
- AWS SageMaker-Lösung
- Microsoft Azure Machine Learning-Lösung

Sowohl die herkömmlichen als auch die verwalteten On-Premise-Lösungen von Dell verwenden dieselbe Hardware. Beim herkömmlichen Plan kauft das Unternehmen die Hardware im Voraus. Bei der Dell APEX Pay-Per-Use-Lösung installiert Dell Hardware im Rechenzentrum des/der KundIn und stellt dem Unternehmen monatlich eine Rechnung basierend auf „zugesagter Kapazität" und „Pufferkapazität".

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 2 13B entschieden, da es sich um ein weit verbreitetes Open-Source-Modell handelt. Wir haben unser Szenario um eine einzige, relativ kleine KI-Workload entwickelt. Wir haben die Kosten für die Entwicklungsnotebooks der Data Scientists für maschinelles Lernen, Datenverarbeitungsaufgaben, kontinuierliche Modellfeinabstimmung und Echtzeitinferenz einbezogen.

Wir haben die Hardware für jede Lösung basierend auf Annahmen über Arbeitsstunden und erforderliche Hardwarefunktionen dimensioniert. Für diese Recherche haben wir öffentliche Quellen genutzt. Wir haben Onlinepreisrechner für AWS SageMaker und Azure Machine Learning verwendet und Angebote von Dell Technologies für die Kosten mithilfe der von Dell empfohlenen Preise für die beiden Dell Lösungen angefordert und erhalten. Wir haben keine der Lösungen für diesen Bericht praktisch getestet.

Unsere Ergebnisse

Im Hauptbericht zeigen wir die normalisierten Vergleiche jeder der beiden On-Premise-Lösungen von Dell mit den Cloud-Lösungen von AWS SageMaker und Azure Machine Learning. Die Tabellen 5 und 6 zeigen die Kostenbasis für diese Normalisierungen. Der normalisierte Wert ergibt sich aus der Division der einzelnen Werte durch die Kosten für die in der Tabelle gezeigte Dell On-Premise-Lösung. Tabelle 5 zeigt, dass Cloud-Lösungen über einen Zeitraum von drei Jahren bis zu 2,88-mal so viel kosten wie die herkömmliche On-Premise-Lösung von Dell. Aus der Amortisationsrechnung geht hervor, dass die Kosten für drei Jahre der herkömmlichen lokalen Dell Lösung in etwa den Kosten für nur ein Jahr der beiden Cloud-Lösungen entsprechen.

Tabelle 5: Normalisierte TCO über drei Jahre für die herkömmliche On-Premise-Lösung von Dell im Vergleich zu SageMaker- und Azure Machine Learning-Lösungen.

	Kosten der herkömmlichen On-Premise von Dell über drei Jahre	AWS SageMaker mit einer Laufzeit von drei Jahren	Azure Machine Learning mit einer Laufzeit von drei Jahren
Gesamt	817.880,00$	2.357.549,00 $	2.231.805,00 $
Normalisiert	1	2.88	2.72
Amortisation in Monaten für die Dell Lösung im Vergleich zu Cloud-Lösungen	N. z.	12,5	13.3

Tabelle 6 zeigt, dass die Cloud-Lösungen über einen Zeitraum von drei Jahren bis zu 3,81-mal so viel kosten wie eine Dell APEX-Pay-Per-Use-Lösung und dass drei Jahre der herkömmlichen On-Premise-Lösung von Dell weniger kosten als nur ein Jahr der beiden Cloud-Lösungen.

Tabelle 6: Normalisierte TCO über drei Jahre für eine Dell APEX Pay-Per-Use-Lösung im Vergleich zu SageMaker- und Azure Machine Learning-Lösungen.

	Kosten der Dell APEX Pay-Per-Use-Lösung über drei Jahre	AWS SageMaker mit einer Laufzeit von drei Jahren	Azure Machine Learning mit einer Laufzeit von drei Jahren
Gesamt	618.648,00$	2.357.549,00 $	2.231.805,00 $
Normalisiert	1	3.81	3.60
Amortisation in Monaten für die Dell Lösung im Vergleich zu Cloud-Lösungen	N. z.	9,5	10

Überlegungen zum Speicher

Wir haben keine Kosten für Storage berücksichtigt, der über das hinausgeht, was die Server oder Instanzen für ihre Aufgaben benötigen.

Die On-Premise-Lösungen von Dell umfassen 106,88 TB Storage:

- 5,76 TB SSD-Kapazität auf jedem der drei PowerEdge R660-Haupt-Nodes
- 44,8 TB SSD-Kapazität auf jedem der beiden PowerEdge XE9680-GPU-Worker-Nodes

Clustermanagement- und Notebookaufgaben teilen den Storage auf den Haupt-Nodes. Verarbeitungs-, Modellfeinabstimmungs- und Inferenzierungsaufgaben teilen den Storage auf den GPU-Worker-Nodes. Wir haben die Dell PowerEdge-Cluster mit etwas zusätzlichem Storage im Vergleich zu den Cloud-Lösungen bereitgestellt, um Platz für Managementaufgaben auf den Haupt-Nodes und bei Bedarf etwas Spielraum für Wachstum zu gewährleisten.

Die AWS SageMaker-Lösung umfasst insgesamt 63,444 TB Storage:

- 7.000 GB EBS gp2-Storage, der für jede der beiden ml.r5.16xlarge-Verarbeitungsinstanzen erworben wurde
- 8 x 3.084 GB NVMe-SSDs für jede der ml.p5.48xlarge-Instanzen
- 1 x 5 GB temporärer EBS-Storage für jede Notebookinstanz, die in der Instanz enthalten ist

Die Azure Machine Learning-Lösung umfasste insgesamt 64,82 TB temporären Storage:

- 7.168 GiB temporärer Storage für die M64-Verarbeitungsinstanz
- 6.400 GiB temporärer Storage für jede der acht ND96amsr A100 v4-Instanzen
- 1 x 100 GiB temporärer Storage für jede Notebookinstanz, die in der Instanz enthalten ist

Die Storage-Anforderungen variieren für Notebookinstanzen, erfordern aber in der Regel nicht viel für diese Art von Workload. Daher haben wir uns entschieden, die Cloud-Instanzen mit dem standardmäßig bereitgestellten Storage zu belassen. Wir haben die Datenübertragungskosten für die EBS-Datenübertragung in AWS SageMaker und für die Blob-Storage-Übertragung in den Azure Machine Learning-Lösungen berücksichtigt. Wir haben keine Datenübertragungskosten für die Dell On-Premise-Lösungen berücksichtigt, bei denen integrierte SSDs verwendet werden.

Nutzungsstunden

Wir haben die Lösungen basierend auf den folgenden Schätzungen von Stunden pro Monat für Notebooks, Datenverarbeitung, Modellfeinabstimmung und Inferenznutzung dimensioniert. Wir haben diese Stunden verwendet, um Stunden der Instanznutzung für die Cloud-Lösungen zu berechnen und diese Instanzen und die Server für die On-Premise-Lösungen zu dimensionieren.

Bei der Dimensionierung der Lösungen wurde davon ausgegangen, dass es 22 Arbeitstage pro Monat gibt, wobei Workloads über Nacht ausgeführt werden, um die Nutzung zu maximieren. Daher stehen für jeden Server und jede Cloud-Instanz pro Monat 528 Stunden Laufzeit zur Verfügung. (Siehe Tabelle 7.)

Tabelle 7: Nutzungsstunden für die vier Aufgaben.

Aufgabe	Gesamtstunden pro Monat	Nutzungsberechnungen
Laptop	3.520	Wir haben jede Lösung so dimensioniert, dass sie 20 DatenexpertInnen unterstützt, mit jeweils einer Notebookinstanz, 8 Stunden am Tag, 22 Tage im Monat, insgesamt 176 Stunden pro Monat, insgesamt 3.520 für alle 20 DatenexpertInnen. Mindestanforderungen sind kleine Cloud-Notebookinstanzen mit 2vCPU und mindestens 4 GiB Speicher.
Verarbeitung	1.056	Datenverarbeitungsaufgaben würden während der 528 Betriebsstunden auf zwei Dell PowerEdge XE9680-Servern mit einer Laufzeit von insgesamt 1.056 Stunden ausgeführt und würden 1.056 Stunden Laufzeit auf Cloud-Instanzen erfordern. Mindestanforderungen für die Cloud-Instanzen waren 64 vCPU, 1.000 GiB Arbeitsspeicher und 7.000 GB Storage. Wir haben die Dell PowerEdge-Server so dimensioniert, dass sie diese Anforderungen sowie die der Feinabstimmung und Inferenzierungsaufgaben unterstützen.
Modellfeinabstimmung	792	Die kombinierte Laufzeit von 1.056 Stunden für Feinabstimmungs- und Inferenzierungsaufgaben erfordert zwei Dell PowerEdge XE9680-Server. Bei allen Aufgaben werden acht H100-Grafikprozessoren und bis zu einem halben TB Arbeitsspeicher verwendet. Die Cloud-Lösungen erfordern dieselbe Anzahl an Stunden auf Instanzen mit acht H100-GPUs oder gleichwertig. Für AWS haben wir zwei H100-Instanzen verwendet. Für Azure Machine Learning haben wir acht A100-Instanzen ersetzt.
Inferenz	264	

Details zu Notebooks

Data Scientists würden kleine Cloud-Notebookinstanzen mit 2 vCPU und mindestens 4 GiB Speicher benötigen. Zwar würden einige Notebookaufgaben mit mehr Speicher möglicherweise besser funktionieren, aber wir haben uns für die AWS ml.t3.medium-Instanz basierend auf den Vorschlägen des AWS SageMaker TCO-Leitfadens[3] entschieden und dann eine ähnliche Instanz für Azure ausgewählt. Bei Dell On-Premise-Lösungen haben wir einen Core eines gleichwertigen Prozessors und 4,3 GB Arbeitsspeicher pro Notebook angenommen, während die Notebooks auf den 3 x PowerEdge R660-Managementservern zusammen mit Managementaufgaben ausgeführt werden.

AWS SageMaker- und Azure Machine Learning-Notebookinstanzen

Für die AWS SageMaker- und Azure Machine Learning-Lösungen haben wir Notebookinstanzen mit 2 vCPUs und mindestens 4 GiB Arbeitsspeicher ausgewählt.

Tabelle 8: Wichtige Konfigurationsinformationen für die AWS SageMaker- und Azure Machine Learning-Notebookinstanzen.

Instanztyp	Instanz	Anzahl der vCPUs pro Instanz	Arbeitsspeicher (GiB) pro Instanz
Notebooks mit SageMaker	ml.t3.medium	2	4
Notebooks mit Azure ML	D2 v2	2	7

Notebooks mit On-Premise-Lösungen von Dell

Die Lösungen von Dell unterstützen diese Notebook-Workloads auf den drei PowerEdge R660-Haupt-Nodes, die zusammen 576 GB Arbeitsspeicher und 48 Prozessorkerne haben, was ausreicht, um sowohl Clustermanagementaufgaben als auch die 20 Notebook-Workloads zu unterstützen. Unsere Annahmen für diese Dimensionierungsentscheidung lauten wie folgt:

- Managementaufgaben beanspruchen weniger als die Hälfte der Prozessorkapazität dieser Haupt-Nodes und weniger als 500 GB Arbeitsspeicher, wobei die verbleibende Kapazität für die Ausführung dieser Aufgaben verfügbar ist.
- Während der 176 Stunden, die jede Notebook-Workload pro Monat ausgeführt wird, würde ein einzelner Prozessorkern verwendet, was den 2 vCPUs für die SageMaker- und Azure ML-Notebooks entspricht, wobei ein Verhältnis von 1 Thread zu 1 vCPU angenommen wird, und 4,3 GB Arbeitsspeicher, was den 4 GiB entspricht, die wir bei der Dimensionierung der Cloud-Notebooks definiert haben. Alle 20 Notebook die gleichzeitig ausgeführt werden, würden weniger als die Hälfte der 48 Cores und 15 % des Arbeitsspeichers auf diesen Systemen verbrauchen.
- Alle Arbeiten an allen Systemen erfolgen in weniger als 72,3 % der Gesamtstunden pro Monat, basierend auf 22 Arbeitstagen pro Woche, wobei 24 Stunden pro Tag verfügbar sind.

Details zur Verarbeitung

AWS SageMaker- und Azure Machine Learning-Verarbeitungsinstanzen

Verarbeitungsaufgaben laufen am besten auf der CPU statt auf der GPU und profitieren von einem hohen Arbeitsspeicher-zu-Kern-Verhältnis.[4] Daher haben wir uns bei den AWS SageMaker- und Azure Machine Learning-Verarbeitungsinstanzen auf arbeitsspeicheroptimierte Instanzen konzentriert. Unser Ziel waren mindestens 64 vCPUs, 1.000 GiB Arbeitsspeicher und 7.000 GB Storage, was die Azure Machine Learning M64-Verarbeitungsinstanz mit 64 vCPUs, 1.000 GiB Arbeitsspeicher und 7.168 GiB temporärem Storage nahezu erfüllte. AWS SageMaker-Verarbeitungsinstanzen mit optimiertem Arbeitsspeicher boten keine Instanz an, die unserer Spezifikation entspricht. Daher wählten wir stattdessen ein Paar speicheroptimierter ml.r5.16xlarge-Verarbeitungsinstanzen von SageMaker mit insgesamt 1.024 GiB Speicher. Zusammen übertreffen diese SageMaker-Instanzen unsere Anforderungen mit der doppelten vCPU-Kapazität und fast der doppelten Storage-Kapazität (mit EBS-Storage) unserer Ziele und der Azure Machine Learning-Instanz.

Tabelle 9: Wichtige Konfigurationsinformationen zu den AWS SageMaker- und Azure Machine Learning-Verarbeitungsinstanzen.

Informationen zur Instanzkonfiguration	SageMaker ml.r5.16xlarge (Verarbeitung)	Azure Machine Learning M64 (Verarbeitung)
Anzahl der Instanzen	2	1
Cloud Service Provider (CSP)	AWS	Azure
Anzahl der vCPUs	64 (128 bei 2 Instanzen)	64
Gesamtarbeitsspeicher im System (GiB)	512 (1.024 bei zwei Instanzen)	1.000
Anzahl der Laufwerke	1 (2 bei zwei Instanzen)	1
Laufwerksgröße (GiB)	7.000 (14.000 bei zwei Instanzen)	7.168
Informationen zum Laufwerk (Geschwindigkeit, Schnittstelle, Typ)	EBS	Temporary

Dell On-Premise-Lösungen

Die beiden Dell PowerEdge XE9680 Worker-Nodes verfügen über mehr Kapazität als die Workloads für Feinabstimmung und Inferenz erfordern, sodass die parallel ausgeführten Verarbeitungsworkloads unterstützt werden können. Die Verarbeitungsworkloads würden sich auf CPUs und die Feinabstimmung und Inferenz des Modells auf GPUs stützen. Um unseren Zielspezifikationen gerecht zu werden, würden die Verarbeitungsworkloads die Hälfte der kombinierten 2 TB Arbeitsspeicher der beiden Server, 32 CPU-Cores und etwa 7 TB des 44,8 TB Storage der beiden Server verwenden.

Details zur Modellfeinabstimmung und Inferenz

Die Lösungen erfordern GPU-Instanzen oder Server für die Feinabstimmung und Inferenz von Workloads mit acht NVIDIA HGX H100-GPUs oder einer gleichwertigen GPU-Kapazität und 512 GB Arbeitsspeicher.

AWS SageMaker- und Azure Machine Learning-Instanzen

Die ml.p5.48xlarge-Inferenzierungsinstanz ist die einzige SageMaker-Instanz, die über die NVIDIA HGX H100-GPUs verfügt.[5] Zum Zeitpunkt unserer Recherche verfügte die beste Azure-VM über acht NVIDIA A100-GPU-VMs. Daher haben wir die Azure-Umgebung auf die vierfache Anzahl der Instanzstunden festgelegt, um ungefähr die gleiche Performance wie die H100s in der AWS SageMaker-Umgebung zu erreichen.[6] Beide Instanztypen verfügen über mehr als unsere Mindestanforderung von insgesamt 512 GB Storage.

Tabelle 10: : Wichtige Konfigurationsinformationen für die AWS SageMaker- und Azure Machine Learning-Instanzen für Feinabstimmung und Inferenz.

Inferenz- und Trainingsinstanzen	SageMaker ml.p5.48xlarge	Azure Machine Learning ND96amsr A100 v4
Anzahl der Instanzen	2 (1 für Feinabstimmung und 1 für Inferenz)	8 (4 für Feinabstimmung und 4 für Inferenz)
Cloud Service Provider (CSP)	AWS	Azure
Anzahl der vCPUs	192	96
Gesamtarbeitsspeicher im System (GiB)	2.048	1.900
Anzahl der Laufwerke	8	1
Laufwerksgröße (GiB)	3.084	6.400
Informationen zum Laufwerk (Geschwindigkeit, Schnittstelle, Typ)	NVMe-SSD	Temporary
Anzahl der GPUs	8	8
Hersteller und Modell	NVIDIA H100	NVIDIA A100

Dell On-Premise-Lösungen

Für die On-Premise-Lösungen von Dell haben wir die beiden PowerEdge XE9680-GPU-Worker-Nodes so dimensioniert, dass sie die Workloads für Feinabstimmung und Inferenz mithilfe von GPU-Ressourcen verarbeiten und die Verarbeitungsworkloads mithilfe von ungenutzten CPU- und Arbeitsspeicherressourcen unterstützen. Die beiden PowerEdge XE9680-GPU-Worker-Nodes verfügen jeweils über zwei Intel Xeon Platinum 8468-Prozessoren mit 48 Cores, 1.024 GB Arbeitsspeicher, 44,8 TB Storage und eine NVIDIA HGX H100-8-GPU-Baugruppe.

Kostenanalyse

Bei den Cloud-Lösungen haben wir die Lizenzierungskosten für die Instanzen einbezogen, die wir für Notebooks, Verarbeitung, Feinabstimmung und Inferenz benötigen. Bei den On-Premise-Lösungen von Dell bieten wir zwei Zahlungsoptionen für die Hardware: ein Vorabkaufmodell und einen monatlichen Zahlungsplan für die Dell APEX Pay-Per-Use-Lösung. Bei beiden On-Premises-Lösungen haben wir die Kosten für die Serververwaltung für die Hardware und das Betriebssystem sowie die Kosten für den Rechenzentrumsbetrieb für den Rack-Platz und die Energiekosten für Strom und Kühlung hinzugefügt, Kosten, die für die beiden Cloud-Lösungen nicht relevant sind.

Wir haben einige Kosten außer Acht gelassen, zum Beispiel:

- Wir haben bei allen vier Lösungen ähnliche Arbeitskosten wie die Installation und Wartung von Open-Source-Software, die Datenübertragung und das Backup sowie die Gehälter der Data Scientists ausgelassen.
- Die Softwarekosten haben wir bei keiner der Lösungen berücksichtigt. SageMaker- und Azure Machine Learning-Instanzen umfassen einige Software und Services wie Jupyter Notebook auf den Notebookinstanzen und Verarbeitungs-APIs mit den Verarbeitungsinstanzen. Wir sind davon ausgegangen, dass alle zusätzlichen Softwareprogramme und Tools, die die Data Scientists dort installieren, als Open-Source-Software verfügbar sind. Mit den On-Premise-Lösungen von Dell würden Data Scientists ausschließlich Open-Source-Software und -Tools wie Jupyter Notebook, Python und PyTorch verwenden, und auf den Servern würde Ubuntu oder ein anderes Open-Source-Betriebssystem laufen.
- Wir haben keine Umsatzsteuern berücksichtigt, da diese von Land zu Land und von Unternehmen zu Unternehmen variieren. Migrationskosten oder Kosten am Ende der Nutzungsdauer sind nicht enthalten.

Bei jeder dieser generativen KI-Lösungen haben wir uns auf einen Lebenszyklus von drei Jahren konzentriert. Wir haben uns für eine Dauer von drei Jahren entschieden, da die Preisrechner von AWS und Azure die Laufzeit auf eine Dauer von maximal drei Jahren begrenzt haben. Drei Jahre sind auch ein angemessener Lebenszyklus für eine On-Premise-Lösung für generative KI, die hochmoderne Hardware erfordert.[7] Unternehmen könnten die von ihnen gekaufte Dell Hardware danach für andere Zwecke verwenden und hätten zwei verbleibende Jahre der enthaltenen fünf Jahre Dell ProSupport für eine hohe Produktivität.

Kosten für Dell On-Premise-Lösungen über drei Jahre

Bei den On-Premise-Lösungen von Dell haben wir die folgenden Kosten über einen Zeitraum von drei Jahren berücksichtigt:

- Der von Dell empfohlene Preis für Dell Server und Switches, einschließlich ProSupport und Vor-Ort-Service am nächsten Tag und ProDeploy Plus für die Server
- Systemadministration zur Wartung und Sicherung der Hardware und des Betriebssystems
- Energiekosten für Strom und Kühlung
- Rechenzentrumskosten für Rack-Platz

Die beiden Lösungen enthielten dieselbe Hardware und würden die gleichen Kosten für die Systemverwaltung, die Energiekosten für Strom und Kühlung sowie die Rechenzentrumskosten für den Rack-Platz verursachen.

Tabelle 11: Kosten für eine herkömmliche On-Premises-Lösung über drei Jahre.

Herkömmliche Dell Technologies On-Premise-Lösung	Kosten über drei Jahre (aufgerundet auf ganze Dollar)
Dell Hardware mit 5 Jahren ProSupport und ProDeploy Plus (für Server)	680.251$
Systemadministration	6.531$
Energiekosten für Strom und Kühlung	88.978$
Rechenzentrumskosten für Rack-Platz	42.120$
Gesamt	817.880$

Für die herkömmliche On-Premise-Lösung haben wir vom Dell Technologies Vertrieb ein Angebot für den empfohlenen Preis der On-Premise-Lösung von Dell angefordert. Am 20. März 2024 hat Dell 680.251 USD als Kaufpreis für die Hardware angegeben.

Das Angebot umfasste fünf Jahre ProSupport und Vor-Ort-Service am nächsten Werktag für die Server und Switches sowie ProDeploy Plus für die Server. Dell definiert den empfohlenen Preis als einen Preis, der als Ausgangspunkt für potenzielle KäuferInnen dient. Er stellt die Kosten dar, zu denen die Lösung für Unternehmen sofort verfügbar ist, selbst wenn sie keine bestehenden KundInnen sind, und fungiert im Wesentlichen als empfohlener Einzelhandelspreis für ihre Produkte.

Tabelle 12: Kosten für eine Dell APEX Pay-Per-Use-Lösung über drei Jahre.

Dell APEX Pay-Per-Use-Lösung	Kosten über drei Jahre (aufgerundet auf ganze Dollar)
Dell Hardware mit 5 Jahren ProSupport und ProDeploy Plus (für Server)	481.019$
Systemadministration	6.531$
Energiekosten für Strom und Kühlung	88.978$
Kosten für Rack-Platz im Rechenzentrum	42.120$
Gesamt	618.648$

Dell APEX Pay-Per-Use-On-Premise-Lösung

Basierend auf diesem empfohlenen Preis hat Dell Technologies eine Kostenschätzung für eine Dell APEX Pay-Per-Use-Lösung über drei Jahre in Höhe von 481.019 USD vorgelegt. Dell hat PT ein Angebot für die Dell ApEX Pay-Per-Use-Lösung für dieselbe Hardware wie oben für eine Laufzeit von 36 Monaten und eine Kapazitätszusage von 75 % übermittelt. Wir haben dieses Angebot am 2. April 2024 erhalten. Die geschätzte Kapazität von 75 % war die nächstgelegene Kapazitätsoption, die die 528 Betriebsstunden abdeckt, für die wir die Lösungen dimensionieren.

Systemadministration

ServeradministratorInnen überwachen und stellen die Leistung, Verfügbarkeit, Funktionalität und Sicherheit der Hardware und des Betriebssystems sicher und installieren in diesem Fall das Betriebssystem. Diese Dienste sind bei der On-Premise-Lösung erforderlich, bei den Cloud-Lösungen jedoch nicht, da sie in deren Servicevereinbarung enthalten sind. Wir haben diese Zeit- und Kostenschätzungen für die On-Premise-Lösung niedrig gehalten, da wir davon ausgehen, dass sie größtenteils automatisiert sind, und weil ProSupport for Infrastructure einige der Aufgaben zur Überwachung und physischen Reparatur von den On-Premise-SystemadministratorInnen auslagert.

Wir haben die Kosten für diese Serververwaltung auf 6.531,00 USD für drei Jahre geschätzt, basierend auf der Gesamtvergütung für mittelrangige SystemadministratorInnen[8], die in der Lage sind, 300 Server und die zugehörigen Switches und Betriebssysteme mithilfe automatisierter Tools und Prozesse zu warten, und die von ProSupport- und ProDeploy Plus-Services unterstützt werden.

Dell ProDeploy Plus for Infrastructure

Wir haben keine separate Schätzung für die Bereitstellung vorgenommen, sondern uns auf Dell ProDeploy Plus for Infrastructure verlassen, einen Service, den wir in das Hardwareangebot aufgenommen haben, um die Hardware- und Softwarebereitstellung vor Ort zu ermöglichen.[9] Ein Bericht von Principled Technologies zeigt, dass ProDeploy Plus for Infrastructure wertvolle Zeit für interne AdministratorInnen einsparen kann, indem von Dell Technologies zertifizierte TechnikerInnen für die Installation und Konfiguration von Dell Produkten eingesetzt werden.[10]

Dieser Service deckt möglicherweise nicht alle Planungsaufgaben, das Auspacken und Einbauen der Switches oder die Installation des Betriebssystems ab. Diese Aufgaben würden wenig Zeit in Anspruch nehmen und sind in der Schätzung der Systemverwaltungszeit für die Wartung und Sicherung der Lösung enthalten.

Dell ProSupport for Infrastructure

Wir haben 5 Jahre Dell ProSupport und Vor-Ort-Service am nächsten Tag einbezogen. Der 5-Jahres-Support ist länger als der 3-Jahres-Zeitraum unserer Analyse, bietet dem kaufenden Unternehmen jedoch den Mehrwert eines längeren Lebenszyklus für die Dell Hardware.

Energiekosten für Strom und Kühlung

Bei den Cloud-Lösungen waren die Energiekosten für Strom und Kühlung in den Preisen enthalten. Bei den On-Premise-Lösungen haben wir die Strom- und Kühlungskosten für drei Jahre mithilfe des Dell Enterprise Infrastructure Planning-Tools geschätzt.[11] Um eine Schätzung zu erhalten, haben wir die Spezifikationen für die Server und Switches eingegeben, die im Verkaufspreisangebot von Dell Technologies enthalten sind. Wir haben zwei weitere Faktoren einbezogen, die sich auf die Berechnungen auswirkten:

- Multiplikator von 1,58 für die Energieeffizienz (Power Use Effectiveness, PUE) der Stromkosten, um die kombinierten Strom- und Kühlungskosten zu ermitteln. Laut dem Uptime Institute, einem Unternehmen, das Rechenzentrumskosten erfasst und nachverfolgt, lag dieser PUE-Wert im Jahr 2023 im Branchendurchschnitt.[12]
- 12,74 Cent pro Kilowattstunde Energiekosten basierend auf dem von der US-amerikanischen Energy Information Administration (EIA) gemeldeten durchschnittlichen Einzelhandelspreis für Strom im Jahr 2023.[13]

Wir haben die Kosten für Strom und Kühlung getrennt für die Geräte berechnet, die im Leerlauf und mit Rechenlasten laufen. Wir haben die Ergebnisse basierend auf den 528 Laufzeitstunden (ca. 72,3 % eines durchschnittlichen Monats) abgewogen, für die wir die Lösungen dimensioniert haben.

Tabelle 13: Energiekosten für Strom und Kühlung über drei Jahre.[14]

Workloads	Energiekosten für On-Premise-Lösungen über drei Jahre	Gewichtungsmultiplikator	Gewichtete Energiekosten für Strom und Kühlung
Rechenleistung	114.439,69$	72,3 %	82.739,90 $
Spannungslos	22.516,69$	27,7 %	6.237,12 $
Gewichtete Energiekosten für Strom und Kühlung über drei Jahre			88.977,02$

Kosten für Rack-Platz im Rechenzentrum

Dem Unternehmen würden zusätzliche Kosten für die Unterbringung der Server und Racks im Rechenzentrum entstehen. Diese Kosten umfassen die Kosten für das Rack, die Vernetzung und andere Kosten für die Rechenzentrumseinrichtung. Wir schätzten diese Kosten auf 1.800 USD pro Rack und Monat für Racks mit einer nutzbaren Kapazität von 28 HE.[15] Die Server und Switches in dieser Lösung füllen 18 HE, 65 % eines dieser Racks. Die Kosten für diese Nutzung über drei Jahre belaufen sich auf 42.120 USD.

Wir haben zwar die Datenübertragung für die beiden Cloud-Lösungen für den Zugriff auf den AWS S3-Storage und den Azure Block Storage einbezogen, diese Kosten bei On-Premise-Lösungen jedoch nicht berücksichtigt, da bei diesen die integrierten Festplatten oder lokalen Speicher-Arrays für die Speicherung verwendet werden würden.

AWS SageMaker-Lösung

Wir haben die AWS SageMaker-Lösung so konfiguriert, dass sie der angebotenen Dell On-Premise-Lösung für die vier zuvor beschriebenen Aufgaben so genau wie möglich entspricht: Notebooks, Verarbeitung, Modellfeinabstimmung und Inferenz. Der AWS-Preisrechner für SageMaker listet jede Aufgabe als separates Preismodul auf, das durch Umschalten zur Schätzung hinzugefügt werden kann. Wir haben SageMaker Studio Notebooks, SageMaker Processing, SageMaker Training und SageMaker Real-Time Inference hinzugefügt. Für jedes Modul haben wir die erforderlichen Felder ausgefüllt, um die Stundenkosten für jede Instanz zu bestimmen, die wir für jede Aufgabe ausgewählt haben. Anhand dieser Stundenkosten haben wir dann ermittelt, wie viel NutzerInnen für die Ausführung jeder Instanz für die vorab berechnete Anzahl von Stunden aufwenden würden, die wir basierend auf den Dell Systemen ermittelt haben. Wir haben doppelt so viele Verarbeitungsinstanzen und damit Verarbeitungsstunden berechnet, um sicherzustellen, dass die Verarbeitungskapazität mit den anderen Lösungen übereinstimmt. Wir haben auch EBS-Storage zu den Verarbeitungsinstanzen hinzugefügt, da sie nicht mit Storage außerhalb des Betriebssystemvolumens hochgefahren werden. Nach der Anwendung des Sparplans mit einer Laufzeit von drei Jahren beliefen sich unsere Gesamtkosten auf 2.357.549 USD für drei Jahre. Die vollständigen Instanzdetails finden Sie in Tabelle 14.

Tabelle 14: Kosten für SageMaker-Lösungsinstanzen über drei Jahre.

Service	Instanztyp	Instanz USD/Std.	Laufzeit (Stunden/Mo.)	Kosten für drei Jahre
Notebooks mit SageMaker Studio	ml.t3.medium	0,02244 $	3520	2.843,60$
Verarbeitung mit SageMaker	ml.r5.16xlarge*	2,2908$	2112	174.174,11$
EBS-Storage für die Verarbeitung mit SageMaker (7 TB pro Monat)				25.804,80$
Training mit SageMaker	ml.p5.48xlarge**	56,5340$	792	1.611.897,41$
Echtzeitinferenz mit SageMaker	ml.p5.48xlarge**	56,5340$	264	537.299,14$
S3-Datenübertragung (1 TB für die eingehende und 15 TB für die ausgehende Datenübertragung)				5.529,60$
Gesamt (aufgerundet)				2.357.549$

* Wir beziehen zwei Prozessorinstanzen ein, um 1 TB Arbeitsspeicher für die Verarbeitungsaufgaben bereitzustellen.
** Die Preise wurden basierend auf dem Prozentsatz der öffentlichen P5-Verpflichtung (nicht ml.p5) geschätzt, da der Rechner diese Instanz nicht berücksichtigt.

Azure Machine Learning-Lösung

Wir haben den Azure-Preisrechner verwendet, um jeden Instanztyp einzubinden, um die stündlichen Kosten für jeden Instanztyp im Machine Learning-Service zu ermitteln. Da die beste GPU-Instanz, die Azure zum Zeitpunkt dieser Recherche angeboten hat, acht A100-GPUs umfasste, haben wir die Kosten für die Ausführung der vierfachen Anzahl dieser GPU-Instanzen pro Aufgabe berechnet, um eine genauere Annäherung an die Leistung der acht H100-GPUs zu erhalten, die die AWS-Instanzen und Dell PowerEdge XE9-Server bereitstellen. Anhand dieser Stundenkosten haben wir dann ermittelt, wie viel NutzerInnen für die Ausführung jeder Instanz für die vorab berechnete Anzahl von Stunden aufwenden würden, die wir basierend auf den Dell Systemen ermittelt haben. Nach der Anwendung der reservierten Laufzeit von drei Jahren für Azure beliefen sich unsere Gesamtkosten für drei Jahre auf 2.231.805 USD. Die vollständigen Instanzdetails finden Sie in Tabelle 15.

Tabelle 15: Kosten für die Azure Machine Learning-Lösungsinstanzen über drei Jahre.

Service	Instanztyp	Instanz USD/Std.	Laufzeit (Stunden/Mo.)	Kosten für drei Jahre
Notebooks mit Azure ML	D2 v2	0,0476$	3.520	6.027,72$
Verarbeitung mit Azure ML	M64	1,8258 $	1.056	69.407,81$
Training mit Azure ML	ND96amsr A100 v4	14,1475$	3.168	1.613.491,01$
Echtzeitinferenz mit Azure ML	ND96amsr A100 v4	14,1475$	1.056	537.830,34$
Azure Block Blob Storage-Datenübertragungsvorgänge (10.000.000)				5.047,20$
Gesamt (aufgerundet)				2.231.805$

1. AWS, „Get Started with P5 instances“, abgerufen am 29. April 2024, https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/p5-instances-started.html.
2. Microsoft Build, „NDm A100 v4-series“, abgerufen am 29. April 2024, https://learn.microsoft.com/en-us/azure/virtual-machines/ndm-a100-v4-series
3. Amazon, „Total Cost of Ownership of Amazon SageMaker“, abgerufen am 29. April 2024, https://pages.awscloud.com/rs/112-TZM-766/images/Amazon_SageMaker_TCO_uf.pdf.
4. StackOverflow, „Why should preprocessing be done on CPU rather than GPU?“, abgerufen am 20. April 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu
und Hugging Face, „Model Memory Requirements“, abgerufen am 29. April 2024, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
5. AWS, „New – Amazon EC2 P5 Instances Powered by NVIDIA H100 Tensor Core GPUs for Accelerating Generative AI and HPC Applications“, abgerufen am 29. April 2024, https://aws.amazon.com/blogs/aws/new-amazon-ec2-p5-instances-powered-by-nvidia-h100-tensor-core-gpus-for-accelerating-generative-ai-and-hpc-applications/.
6. Comet, „Comparison of NVIDIA A100, H100 + H200 GPUs“, abgerufen am 29. April 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.
7. The Jerusalem Post, „Maximizing Efficiency: Your 2023 Guide to GPU Servers“, abgerufen am 8. April 2024, https://www.jpost.com/insights/article-770858.
8. Systems Administrator II – Gesamtvergütung (Gehalt und Leistungen) in Höhe von 130.616 USD pro Jahr. Quelle: Salary.com, „Systems Administrator II“, abgerufen am 25. März 2024, https://www.salary.com/tools/salary-calculator/systems-administrator-ii-benefits.
9. Dell, „The market's most complete deployment offer“, abgerufen am 28. März 2024, https://www.delltechnologies.com/asset/en-us/services/deployment/briefs-summaries/prodeploy_plus_deployment_unification_ds.pdf.
10. Principled Technologies, „Using Dell ProDeploy Plus for Infrastructure can improve deployment times for Dell technology“, abgerufen am 28. März 2024, https://www.delltechnologies.com/asset/en-us/products/cross-company/industry-market/principled-technologies-prodeploy-plus-for-infrastructure-services-whitepaper.pdf.
11. Dell, „Dell Enterprise Infrastructure Planning Tool“, abgerufen am 24. März 2024, https://dell-ui-eipt.azurewebsites.net/#/
12. Uptime Institute, „Large data centers are mostly more efficient, analysis confirms“, abgerufen am 29. März 2024, https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/.
13. EIA, „Electricity Data Browser“, abgerufen am 29. März 2024, https://www.eia.gov/electricity/data/browser/#/topic/7?agg=0,1&geo=g&endsec=vg&linechart=ELEC.PRICE.US-ALL.A~ELEC.PRICE.US-RES.A~ELEC.PRICE.US-COM.A~ELEC.PRICE.US-IND.A&columnchart=ELEC.PRICE.US-ALL.A~ELEC.PRICE.US-RES.A~ELEC.PRICE.US-COM.A~ELEC.PRICE.US-IND.A&map=ELEC.PRICE.US-ALL.A&freq=A&ctype=linechart&ltype=pin&rtype=s&maptype=0&rse=0&pin=.
14. Dell, „Dell Enterprise Infrastructure Planning Tool“, abgerufen am 25. März 2024, https://dell-ui-eipt.azurewebsites.net/#/.
15. Softchoice, ein Partner von VMware by Broadcom, verwendet diese Rack-Kosten in einer TCO-Analyse zum Vergleich der Kosten für den Betrieb von VMware vor Ort oder in der Cloud. Quelle: Softchoice, „VMware Cloud on AWS“, abgerufen am 24. März 2024, https://www.softchoice.com/technology-partners/vmware/cloud-on-aws-tco-calculator.

Lesen Sie die Originalversion dieses Berichts in englischer Sprache unter https://facts.pt/9PHKEUe

Dieses Projekt wurde in Auftrag gegeben von Dell Technologies.

Principled Technologies®

Facts matter.®

Principled Technologies ist eine eingetragene Marke von Principled Technologies, Inc.
Alle anderen Produktnamen sind Marken der jeweiligen Inhaber.

GEWÄHRLEISTUNGSAUSSCHLUSS, HAFTUNGSEINSCHRÄNKUNG:
Principled Technologies, Inc. hat angemessene Anstrengungen unternommen, die Genauigkeit und Richtigkeit der Tests sicherzustellen. Principled Technologies, Inc. schließt jedoch jegliche ausdrückliche und implizite Gewährleistung aus, die sich auf die Testergebnisse und Analysen, deren Genauigkeit, Vollständigkeit oder Qualität bezieht, einschließlich jeglichen impliziten Eignungsversprechens für einen bestimmten Zweck. Alle natürlichen oder juristischen Personen, die sich auf die Ergebnisse der Tests verlassen, tun dies auf eigenes Risiko und stimmen zu, dass Principled Technologies, Inc., seine Mitarbeiter und Auftragsnehmer keinerlei Haftung für Verlust oder Beschädigung jeglicher Art aufgrund vermeintlicher Fehler oder Mängel in einem Testverfahren oder Testergebnis übernehmen.

In keinem Fall haftet Principled Technologies, Inc. für indirekte, spezielle, zufällige oder Folgeschäden in Verbindung mit den Tests, auch wenn das Unternehmen auf die Möglichkeit solcher Schäden hingewiesen wurde. In keinem Fall geht die Haftung von Principled Technologies, Inc. über den in Verbindung mit den Tests von Principled Technologies, Inc. gezahlten Betrag hinaus. Dies gilt auch für direkte Schäden. Die alleinigen und ausschließlichen Rechtsmittel, die dem Kunden zur Verfügung stehen, sind die in diesem Dokument beschriebenen Rechtsmittel.