White Paper

KI-Modernisierung durch Dateninfrastrukturen beschleunigen

Gesponsert von: IBM Corporation

Ashish Nadkarni Sriram Subramanian Matt Leib
February 2021

EXECUTIVE SUMMARY

Künstliche Intelligenz (KI), maschinelles Lernen (ML) und oft auch Deep Learning (DL) sind heute wesentliche Bestandteile von Projekten zur digitalen Transformation (DX). Die Geschäftsmöglichkeiten, die sich mithilfe von KI erschließen lassen, sind außerordentlich vielversprechend. Unternehmen sind sich zunehmend bewusst, dass ein Verzicht auf den Einsatz von KI eine wirtschaftliche Katastrophe bedeuten kann, da Konkurrenten eine Fülle von bisher nicht verfügbaren Erkenntnissen und Fähigkeiten erhalten, um ihren Kundenstamm zu erweitern und zu begeistern. Nur wenige Unternehmen behaupten heute: „KI ist nichts für uns“ oder „KI ist nur ein Hype.“ Vielmehr werden weltweit, branchenübergreifend und über alle Unternehmensgrößen hinweg, ernsthafte KI-Initiativen durchgeführt.

Unternehmen suchen nach einer KI-gesteuerten Transformations- und Modernisierungsinitative, bei der es darum geht, von der Experimentierphase zum geschäftlichen Nutzen ihrer KI-Investitionen überzugehen. Der Erfolg von Unternehmensinvestitionen in die digitale Transformation im Zusammenhang mit KI ist direkt mit dem Umfang des Fachwissens verbunden, das für die Entwicklung, Implementierung und Wartung von KI-Lösungen in großem Maßstab erforderlich ist. Viele Geschäftsbereiche (LOBs), IT-Mitarbeiter, Datenwissenschaftler und Entwickler in Unternehmen arbeiten daran, mehr über KI zu erfahren, die Anwendungsfälle zu verstehen, eine KI-Strategie für ihr Unternehmen zu definieren, erste KI-Initiativen zu starten und die daraus resultierenden KI-Anwendungen zu entwickeln und zu testen, die neuen Erkenntnisse und Fähigkeiten mithilfe von Algorithmen für maschinelles Lernen, insbesondere Deep Learning, liefern.

Wenn Unternehmen diese Initiativen ausbauen, tauchen neue Fragen auf. Sie wissen - und haben es vielleicht sogar schon selbst erlebt -, dass sie keine standardmäßige, universelle Datenverarbeitung und bestehende oder veraltete Speicherinfrastruktur verwenden können. Sie stellen auch fest, dass KI-Training (das Trainieren des KI-Modells) und KI-Inferenz ("Inferencing", d. h. die Verwendung des trainierten Modells zum Verstehen oder Vorhersagen eines Ereignisses) unterschiedliche Arten von skalierbarer Datenverarbeitung mit einer ebenso skalierbaren Speicherinfrastruktur erfordern.

IDC stellt fest, dass Unternehmen, die versuchen, ihre vorhandene Ausstattung zu verwenden, ohne einen Teil oder die gesamte Infrastruktur zu modernisieren, ein höheres Risiko des Scheiterns haben. Selbst innerhalb der Modernisierung variiert der Schwerpunkt von Unternehmen zu Unternehmen. Während Unternehmen die Rechenleistung besser im Blick haben, unterschätzen sie häufig den Wert der Datenspeicherung für KI.

IDC stellt fest, dass Unternehmen, die versuchen, ihre vorhandene Ausstattung zu verwenden, ohne einen Teil oder die gesamte Infrastruktur zu modernisieren, ein höheres Risiko des Scheiterns haben. Selbst innerhalb der Modernisierung variiert der Schwerpunkt von Unternehmen zu Unternehmen. Während Unternehmen die Rechenleistung besser im Blick haben,

unterschätzen sie häufig den Wert der Datenspeicherung für KI. Darüber hinaus sind KI-Anwendungen und insbesondere Deep-Learning-Systeme, die exponentiell wachsende Datenmengen auswerten, extrem anspruchsvoll und erfordern leistungsfähige parallele Verarbeitungskapazitäten auf der Grundlage einer großen Anzahl von Rechenkernen, und Standardspeichersysteme können die Ausführung dieser KI-Aufgaben nicht ausreichend unterstützen. Schließlich ist es wichtig, dass KI-bezogene Initiativen in die Bemühungen zur Anwendungsmodernisierung einfließen, die Kubernetes und/oder Container und die Integration mit einem oder mehreren Public-Cloud-Diensten über eine hybride Cloud-Architektur umfassen.

Der Research von IDC zeigt, dass in Bezug auf die Speicherinfrastruktur eine unsachgemäße oder unzureichende Beachtung von Details Initiativen zur KI-Transformation schnell zum Scheitern bringen kann. Um diese Lücke zu schließen, müssen Unternehmen, die mit der bestehenden Infrastruktur experimentiert haben und nun bereit sind, diese in die Produktion zu überführen, ihre Infrastruktur überarbeiten, um die erforderliche Parallelverarbeitung zu erreichen, und zwar durch Investitionen in modernere Speicherlösungen, die sich in großem Umfang skalieren lassen und in die Cloud, in Container und in leistungsintensive Berechnungen für die globale Bereitstellung und den Datenzugriff integriert sind. Hier liefern Lösungen wie IBM Spectrum Scale und IBM Elastic Storage System (ESS) die notwendigen Komponenten für eine KI-Informationsarchitektur. Die Lösung eignet sich für KI-Workloads, die Bereitstellung in Containern und die Bereitstellung in einer Hybrid-Cloud, die speziell auf KI-Workloads ausgerichtet ist

SITUATIONSÜBERBLICK

AI/ML ist hier und jetzt

Unternehmen auf der ganzen Welt reagieren mit Nachdruck auf die neuen Möglichkeiten, die sich durch Investitionen in KI ergeben, um ihre Initiativen zur digitalen Transformation voranzutreiben. Künstliche Intelligenz ist ein Bündel von Technologien, die natürliche Sprachverarbeitung (NLP), Bild-/Videoanalyse, maschinelles Lernen, Wissensgraphen und andere Technologien nutzen, um Fragen zu beantworten, Erkenntnisse zu gewinnen und Empfehlungen zu geben. Diese Systeme stellen Hypothesen auf und formulieren mögliche Antworten auf der Grundlage verfügbarer Beweise. Sie können durch die Aufnahme großer Mengen von Inhalten trainiert werden, passen sich an und lernen aus ihren Fehlern und Misserfolgen durch Nachschulung oder menschliche Aufsicht. IDC erwartet, dass bis zum Jahr 2022 mindestens 60 % dieser KI-zentrierten Anwendungsfälle in mindestens 65 % der Global-2000-Organisationen eingesetzt werden, was einem Zuwachs von 34 % gegenüber 2019 entspricht.

IDC erwartet, dass bis zum Jahr 2022 mindestens 60 % dieser KI-zentrierten Anwendungsfälle in mindestens 65 % der Global-2000-Organisationen eingesetzt werden, was einem Zuwachs von 34 % gegenüber 2019 entspricht.

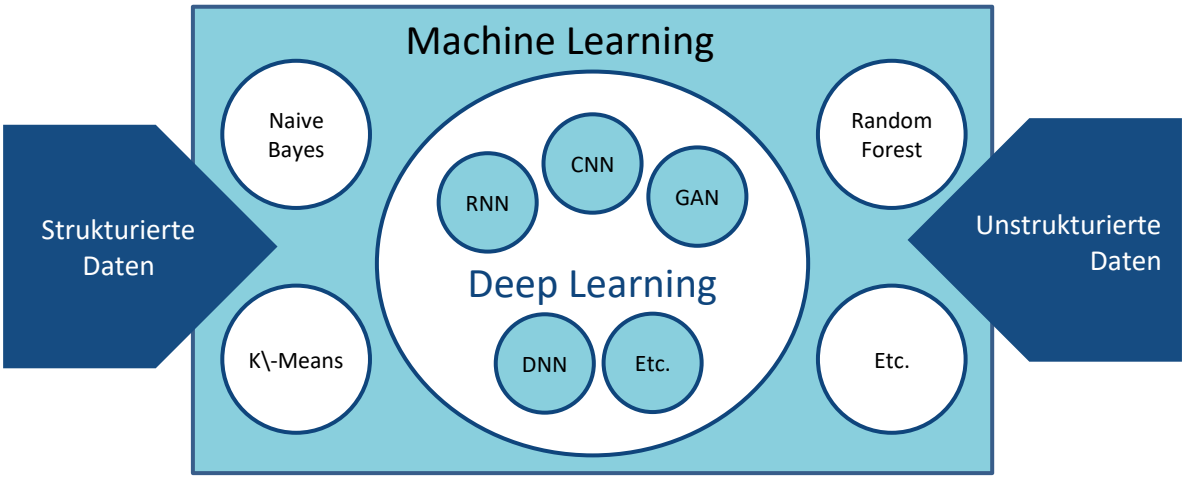
Bei der Automatisierung von Prozessen und Arbeitsabläufen setzt sich künstliche Intelligenz schnell in allen Unternehmen durch. Im Jahr 2019 untersuchte IDC 176 Anwendungsfälle für die digitale Transformation in acht Funktionsbereichen, darunter Kundenerfahrung (CX), Recht und Unternehmensstrategie, Facilities und Beschaffung, und schätzte, dass etwa 26 % dieser Anwendungsfälle sowohl von KI abhängig sind als auch derzeit in den meisten Unternehmen eingesetzt werden.

Das bedeutet, dass die meisten führenden Unternehmen bald KI-Technologien wie natürliche Sprachverarbeitung, maschinelles Lernen, Deep Learning und Speech-to-Text unternehmensweit einsetzen werden, um Abläufe zu skalieren, unstrukturierte Daten sinnvoll zu nutzen und intelligente Geschäftseinblicke zu liefern. In der Zwischenzeit werden Unternehmen, die immer noch nicht herausgefunden haben, wie sie KI-basierte Anwendungsfälle vom Proof of Concept (POC) in die Produktion überführen können, weiter zurückfallen und die digitale Kluft vertiefen.

Maschinelles Lernen ist ein Teilbereich der KI-Techniken, der es Computersystemen ermöglicht, zu lernen und ihr Verhalten für eine bestimmte Aufgabe zu verbessern, ohne dass sie von einem Menschen programmiert werden müssen. Modelle des maschinellen Lernens sind Algorithmen, die sich durch wiederholte Tests mit großen Mengen strukturierter und/oder unstrukturierter Daten stetig optimieren lassen, bis sie eine Aufgabe als „gelernt" betrachten (z. B. das Erkennen eines menschlichen Gesichts). Abbildung 1 zeigt, dass Deep Learning eine Untergruppe von ML ist. Typische DL-Architekturen sind Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs) und viele mehr.

ABBILDUNG 1

Maschinelles Lernen und Deep Learning-Anwendungen



Quelle: IDC, 2021

Zu den KI-Softwareplattformen gehören unter anderem dialogorientierte KI-Software (z. B. digitale Assistenten), Vorhersageanalysen zur Entdeckung verborgener Beziehungen zwischen Daten und zur Erstellung von Vorhersagen, Textanalysen und natürliche Sprache zur Erkennung, zum Verständnis und zur Extraktion von Werten aus Texten, Sprachanalysen zur Erkennung, Identifizierung und Extraktion von Informationen aus Audio-, Sprach- und Sprechdaten sowie Bild- und Videoanalysen zur Erkennung, Identifizierung und Extraktion von Informationen aus Bildern und Videos, einschließlich Mustererkennung, Objekten, Farben und anderen Attributen wie Menschen, Gesichtern, Emotionen, Autos und Landschaften, um nur einige zu nennen.

Viele Unternehmen sind mit ihren KI-Initiativen auf einem guten Weg und haben ein Stadium erreicht, in dem sie bereit sind, KI im großen Maßstab einzusetzen. Andere wiederum experimentieren noch mit KI, und eine dritte Gruppe befindet sich in der Phase, in der sie evaluiert, was KI-Anwendungen für ihr Unternehmen bedeuten können.

Was die erste Gruppe (bereit für den Einsatz) betrifft, so sieht IDC eine Reihe von KI-Anwendungsfällen, die Unternehmen, Behörden und andere Organisationen bereits implementiert

haben. Die fünf häufigsten Anwendungsfälle sind heutzutage, geordnet nach dem Betrag, den Unternehmen für sie in Form von Hardware, Software und Dienstleistungen ausgeben, die folgenden:

- **Automatisierte Kundenbetreuer.** Im Bankensektor beispielsweise bieten diese KI-Anwendungen einen Kundenservice über ein lernendes Programm, das die Bedürfnisse und Probleme der Kunden versteht und der Bank hilft, den Zeit- und Ressourcenaufwand für die Lösung von Kundenproblemen zu reduzieren. Diese Assistenten werden branchenübergreifend immer häufiger eingesetzt.
- **Empfehlung und Automation von Verkaufsprozessen.** Dabei handelt es sich um KI-Anwendungen, die mit Customer Relationship Management (CRM)-Systemen zusammenarbeiten, um den Kundenkontext in Echtzeit zu verstehen und den Vertriebsmitarbeitern entsprechende Aktionen zu empfehlen.
- **Automatisierte Systeme zur Erkennung und Abwehr von Bedrohungen.** Diese KI-Anwendungen, die sich zu einem wichtigen Bestandteil der Bedrohungsprävention in Regierungen und Unternehmen entwickelt haben, verarbeiten nachrichtendienstliche Berichte, extrahieren Informationen aus ihnen, stellen Beziehungen zwischen verschiedenen Informationen her und identifizieren dann Bedrohungen für Datenbanken, Systeme, Websites usw.
- **Analyse und Untersuchung von Betrugsfällen.** In der Versicherungsbranche, aber auch in anderen Bereichen weit verbreitet, nutzen diese KI-Anwendungen das Prinzip des regelbasierten Lernens, um betrügerische Transaktionen zu erkennen, und sie lernen automatisch, verschiedene versicherungsbezogene Betrugsmethoden zu identifizieren.
- **Automatisierte Predictive Maintenance.** In der Fertigungsindustrie basieren diese KI-Anwendungen auf Algorithmen des maschinellen Lernens, die ein genaues Prognosemodell für potenzielle Anlagen- und Maschinenausfälle erstellen und so Ausfallzeiten und Wartungskosten reduzieren.

Weitere KI-Anwendungsfälle, die sich in Unternehmen durchgesetzt haben (in der Reihenfolge der Ausgaben für Hardware, Software und Dienstleistungen), sind:

- Programmbetreuer und Empfehlungssysteme
- Diagnose- und Behandlungssysteme
- Intelligente Automatisierung der Verarbeitung
- Qualitätsmanagement Untersuchungs- und Empfehlungssysteme
- IT-Automatisierung und digitale Assistenten für Fachkräfte in Unternehmen
- Kompetente Einkaufsberater und Produktempfehlungen
- Beschaffung und Logistik sowie rechtliche Informationen
- Anlagen-/Flottenmanagement und automatisierte Schadenbearbeitung
- Digitaler Zwilling/intelligente digitale Simulation
- Öffentliche Sicherheit und Notfallmaßnahmen
- Anpassungsfähiges Lernen
- Intelligente Netzwerke
- Fracht-, Anlagen- und Flottenmanagement
- Pharmazeutische Forschung und Entwicklung

KI-Transformation erfordert eine solide Datengrundlage

ML- und DL-Initiativen stützen sich in hohem Maße auf eine Kombination verschiedener strukturierter und unstrukturierter Dateneingaben (siehe Abbildung 1). KI-Projekte umfassen mehrere Schritte, darunter Datenerfassung, Konfiguration, Bereinigung, Überprüfung, Modellerstellung, Training, Tests, Schlussfolgerungen und Datenabgabe. Da jeder Schritt unterschiedliche Anforderungen mit sich bringt - von schnellem Zugriff über

Speicher mit geringer Latenz bis hin zu kostengünstigem Archivierungsspeicher - müssen Unternehmen für diese Schritte eine geeignete und kostengünstige Speicherinfrastruktur auswählen. Außerdem müssen sie für die Verwaltung von KI-Daten über den gesamten Lebenszyklus hinweg eine Reihe unterschiedlicher Tools verwenden.

Mit der Verbreitung der 5G-Technologie und der Zunahme von IoT-basierten Sensoren werden mehr Daten am Rande des Netzes generiert und verbraucht. Kameras und intelligente Assistenten werden am Rande des Geräts eingesetzt, um eine schnellere Reaktion und ein besseres Nutzererlebnis zu ermöglichen. Der Bedarf an lokaler KI, die in der Netzperipherie oder auf den Endpunkten verarbeitet wird, wächst. Latenzabhängige KI-Anwendungen, die auf Edge-Geräten mit begrenzter Konnektivität ausgeführt werden, erfordern eine hohe Skalierbarkeit. KI-gestützte Edge-Computing-Anwendungsfälle werden übernommen.

Eine KI-Dateninfrastruktur beschleunigt die KI-Transformation

Wenn Unternehmen in künstliche Intelligenz investieren, werden mehrere Faktoren für die Gestaltung und Bereitstellung der Infrastruktur immer wichtiger. In Übereinstimmung mit den Hyperscale- und Cloud-Service-Providern gehen die Unternehmen die Infrastrukturanforderungen über die Schaffung einer einheitlichen Dateninfrastruktur an. Eine Dateninfrastruktur ist eine gemeinsame Grundlage für KI-Initiativen, einschließlich einer hocheffizienten und skalierbaren Rechen- und Speicherebene. Anstatt KI-Workloads als homogen zu betrachten, behandelt eine Dateninfrastruktur sie auf eine zusammengesetzte Weise und verbindet einen Teil dieses Workloads mit der richtigen Rechenschicht, die von einer geeigneten Speicherebene abhängig von einer Mischung aus strukturierten und unstrukturierten Datensätzen unterstützt wird. Dieser Verbundansatz beschleunigt die KI-Transformation in drei Dimensionen:

Herausforderungen bei KI-Initiativen

IDC-Untersuchungen zeigen, dass Unternehmen bei ihren KI-Initiativen häufig die folgenden Herausforderungen im Hinblick auf die Datenverwaltung anführen:

- Datenerfassung und -aufbereitung zu zeitaufwändig
- Silos von Infrastrukturen für verschiedene analytische Anwendungsfälle
- Mehrere Kopien der gleichen Daten ohne eine einzige Wissensquelle
- Notwendigkeit der sicheren Verwaltung und des Schutzes der Datenherkunft für die Wiederholgenauigkeit
- Notwendigkeit der globalen Zugänglichkeit (Hybrid Cloud) und Zusammenarbeit
- Datenintegrität nach der Erfassung und Aufbereitung der Daten

- **Skalierung.** Die Skalierungsdimension beschreibt den Maßstab, auf dem der Workload arbeitet. Die wesentlichen Unterdimensionen - Datenverarbeitung, Vernetzung und Datenpersistenz (Speicherung) - sind alle hardwarebezogen. Entscheidend ist, dass softwarebezogene Unterdimensionen wie die Koordination mit der zunehmenden Größe und Komplexität des Datenstapels an Bedeutung gewinnen, um das Gleichgewicht zu wahren.
- **Portierbarkeit.** Dies ist die Fähigkeit des Workloads, über Kern-, Rand- und Endpunktbereitstellungen hinweg verschoben zu werden. Heute sind viele dieser Workloads statischer Natur (d. h. sie sind für die Ausführung in einer einzigen Bereitstellung konzipiert). während Unternehmen zunehmend die Entwicklung von Arbeitslasten in einer Bereitstellung (z. B. Public Cloud) und deren Installation (in der Produktion) in einer anderen (z. B. Edge) in Betracht ziehen. Dies ist analog zum laufenden Modell der Entwicklung und Bereitstellung von mobilen Anwendungen.
- **Zeit.** Diese Dimension bezieht sich auf die zeitliche Kontinuität des Workloads selbst. Viele KI-Workloads lehnen sich in ihrem Design an das High-Performance-Computing oder die Big-Data- und Analyse-Bereitstellung an - sie sind für den Batch-Betrieb konzipiert. Dank der zunehmenden Verbreitung von leistungsfähigen Beschleunigungssystemen können KI-Workloads zunehmend Streaming-Daten in Echtzeit oder nahezu in Echtzeit analysieren.

Die Mythen und Must-Haves der KI-Dateninfrastruktur

Eine wichtige, wenn auch oft übersehene Grundlage ist die Speicherinfrastruktur. Der Einsatz von KI in großem Maßstab stellt häufig höhere Anforderungen an die Speicherinfrastruktur in Bezug auf Kapazität (Wachstum) und Leistung (IOPS und Bandbreite). Häufig gehen Unternehmen davon aus, dass interner serverbasierter Speicher oder Unternehmensspeicher, der für andere Workloads verwendet wird, für die Ausführung von KI-Anwendungen ausreicht. Und sobald die Infrastruktur aufgebaut ist, stellen sie fest, dass die Speicherung das schwächste Glied in der Kette ist. Jede dieser KI-Anwendungen stellt andere Anforderungen und damit auch Herausforderungen an die IT-Organisation. IT-Einkäufer und -Anbieter sollten daher die sprichwörtliche „Ich habe einen Hammer, also ist alles ein Nagel"-Situation vermeiden.

Ein umsichtigerer Ansatz ist es, die Dateninfrastruktur ganzheitlich zu betrachten. Während Skalierungs- und Zugriffsmechanismen für die Datenpersistenz zum Standard gehören, müssen Unternehmen ihren Blickwinkel auf die Vernetzung und Integration zwischen den Ebenen Computing, Speichersoftware und Systeme erweitern. Unternehmen müssen zu einer konsistenten, durchgängigen Dateninfrastruktur übergehen und nicht nur ein weiteres Speichersystem „anschließen". IDC ist der Ansicht, dass sich die Anforderungen an die Dateninfrastruktur speziell im Hinblick auf die Speicherung auf die in den folgenden Abschnitten erörterten Schlüsselbereiche herunterbrechen lassen.

Integration der Datenverarbeitung

Es wird oft angenommen – und das ist eine fehlerhafte Annahme - dass alle KI-Workloads in Containern untergebracht sind. Im Gegenteil, viele KI-Workloads werden auf Bare Metal oder sogar virtualisiert ausgeführt. Vor allem KI-gestützte Anwendungen werden häufig auf Bare Metal oder virtualisierten Rechnern ausgeführt. Viele KI-Workloads werden für den Einsatz von Beschleunigern optimiert. Das bedeutet nicht, dass alle KI-Workloads am besten auf beschleunigten Rechnern laufen - beschleunigte Rechner stellen für Workloads im Allgemeinen eine Reihe anderer Herausforderungen dar.

Datenpersistenz und Zugang

Wenn die Rechenanforderungen für KI-Workloads variieren, dann auch die Anforderungen an die Datenpersistenz. Ein unterrepräsentierter und missverstandener Aspekt des KI-Workloads-Stapels ist die Datenpersistenz-Ebene. Häufig wird angenommen - auch diese Annahme ist unzutreffend -, dass alle KI-Workloads eine große Menge an Hochleistungsspeicher benötigen. Tatsache ist, dass nicht alle KI-Workloads „große Datensätze" sind - sie können viele kleine Datensätze gleichzeitig für einen kurzen Zeitraum abfragen. Auch Bare-Metal-Workloads, die auf Open-Systems-Computing-Plattformen ausgeführt werden, nutzen häufig den Scale-Out-Block- oder Dateizugriff. Es ist nicht ungewöhnlich, dass virtualisierte Workloads auf einer hyperkonvergenten Infrastruktur (HCI) ausgeführt werden.

Bei einer Mischung aus strukturierten und unstrukturierten Daten, die in die Speicherinfrastruktur eingespeist werden, ist der Multiprotokoll-Zugriff eine Selbstverständlichkeit. Viele IoT- und Edge-Geräte kommunizieren über SMB oder NFS, und einige wenige nutzen S3. In einigen Fällen ist ein Streaming-Datenzugriff erforderlich. Und in einigen Fällen kann auch ein nativer paralleler Dateisystem-Client verwendet werden.

Skalierung und Tiering

Die Unterstützung von KI- und ML-Anwendungen bedeutet, dass Speichersysteme Leistung in großem Umfang bieten müssen. Bei unstrukturierten Datenbeständen handelt es sich um Speichersysteme, die parallele Dateisysteme mit Netzzugriff nutzen. Für strukturierte Daten werden Flash-basierte Speichersysteme verwendet. Unter Skalierung versteht man im Wesentlichen das Erhöhen oder

Verringern von Leistung und Kapazität unabhängig voneinander, um die Anforderungen von KI- und ML-Anwendungen zu erfüllen.

Für eine zukunftssichere Infrastruktur muss das System auch in der Lage sein, ältere Daten einfach und kostengünstig auf einem Objektspeicher mit einer bekannten Objektspeicherschnittstelle wie S3 zu verarbeiten.

Software-Defined Storage

KI und ML wirken als Katalysatoren für softwaredefinierte Speicherung. Sie ermöglichen Infrastruktur als Code und Automatisierung über eine heterogene Software-Steuerungsschicht über der Hardware. Dies trägt zu einer besseren Integration in die KI/ML-Workflows bei und gewährleistet, dass der Speicher nahtlos mit den Anforderungen der Anwendung skaliert.

Agilität und Flexibilität bei der Bereitstellung

Die Anwendungen, die die Einsatzfälle abdecken, können von einem Unternehmen individuell entwickelt werden, sie können auf kommerzieller KI-Software basieren oder als KI-SaaS bereitgestellt werden. Als Einsatzmöglichkeiten für die kundenspezifisch entwickelte und kommerzielle Software bieten sich der Einsatz vor Ort, in der Cloud auf IaaS oder als Hybrid-Cloud an, wobei die lokale Umgebung über eine gemeinsame Automatisierungs- und Orchestrierungsschicht mit einer Public-Cloud-Umgebung interagiert.

Angesichts der verteilten Natur der KI kann man davon ausgehen, dass es am besten ist, die Datenverarbeitung näher an den Ort zu verlegen, an dem die Daten beschafft oder erzeugt werden, als umgekehrt. In letzter Zeit hat sich das Core-Edge-Endpoint-Modell als De-facto-Beschreibung für KI durchgesetzt (wobei Core die Cloud und Endpoints die eingebettete Intelligenz umfassen). Es ist wichtig zu beachten, dass die Workload-Profile für jeden Standort unterschiedlich sind und somit auch die zugrunde liegenden Infrastrukturanforderungen.

Für die verschiedenen Einsatzszenarien müssen Lösungen in Betracht gezogen werden:

- **Sichere Verarbeitung der Datenmengen, die für das Training von KI-Modellen mit extrem hoher Leistung erforderlich sind.** Die Leistungsanforderungen für Deep-Learning-Training umfassen die Fähigkeit zur massiv parallelen Verarbeitung mit GPUs in Kombination mit einer Dateneingabe mit hoher Bandbreite.
- **Sichere Verarbeitung großer Datenmengen, die das KI-Modell mit extrem hoher Leistung ableiten wird.** Leistung in Bezug auf Inferenz bedeutet die Fähigkeit, eingehende Daten durch das trainierte KI-Modell zu verarbeiten und KI-Einsichten oder Entscheidungen nahezu in Echtzeit zu liefern.

Für Datenwissenschaftler und -entwickler kann es manchmal einfacher sein, eine KI-Initiative in der Cloud zu starten, da sie sich so die Einrichtung von lokalem Rechenzentrum ersparen, das für Deep Learning normalerweise beschleunigt werden muss. Beschleunigte KI-Cloud-Instanzen sind in den meisten Public Clouds verfügbar, in der Regel mit Open-Source-KI-Stapeln. Bei beschleunigten Cloud-Instanzen für das KI-Training diktiert natürlich der Cloud-Service-Provider (SP), was dem Endnutzer in Bezug auf Prozessoren, Coprozessoren, Verbindungen, Speichergröße, E/A-Bandbreite usw. zur Verfügung steht. Nicht alle Cloud-SPs bieten die bestmöglichen Kombinationen dieser Komponenten, die letztlich die Geschwindigkeit und Qualität bestimmen, mit der Datenwissenschaftler Trainingsmodelle entwickeln können. Aus diesem Grund entscheiden sich viele Unternehmen für eine Bereitstellung vor Ort.

Bei ihren KI-Experimenten in den letzten Jahren stießen viele Unternehmen mit ihrer Standardinfrastruktur oder mit den einfachen Cloud-Instanzen an ihre Grenzen. Das Trainieren der Modelle dauerte zu lange, und das Inferenzverfahren war zu langsam. Die Studie von IDC zeigt, dass 77,1 % der Befragten angaben, dass sie mit ihrer KI-Infrastruktur vor Ort auf eine oder mehrere Einschränkungen gestoßen sind, und 90,3 % der Befragten gaben an, dass sie in der Cloud auf Rechenbeschränkungen gestoßen sind.

ERWEITERBARE GLOBALE KI-INFORMATIONSARCHITEKTUR MIT IBM STORAGE

IBM Storage für Daten- und KI-Lösungen ermöglicht Kunden die nahtlose Einführung von KI-Initiativen im Produktionsmaßstab in hybriden Cloud-Umgebungen. IBM ist nach wie vor Marktführer bei skalierbaren Hochleistungs-Workloads sowie bei effizientem, sicherem und skalierbarem Speicher mit hoher Kapazität für leistungsstarke KI- und Big Data-Lösungen. Das Speicherportfolio von IBM bietet integriertes Speicher- und Datenmanagement am Edge, im zentralen Rechenzentrum und in der Public Cloud und beschleunigt so die KI-Modernisierung. Es bietet umfassende Unterstützung für und Integration mit Kubernetes-Containern und der Red Hat OpenShift-Plattform und kann in der Public Cloud oder für Rechenzentrums-Workloads bereitgestellt und genutzt werden. IBM Speichersysteme für Daten und KI sollen Komplexität und Kosten reduzieren, indem sie eine tiefgreifende, verbesserte Integration mit einer KI-Informationsarchitektur bieten, die in großem Umfang für das gesamte Unternehmen eingesetzt werden kann.

IBM Spectrum Scale

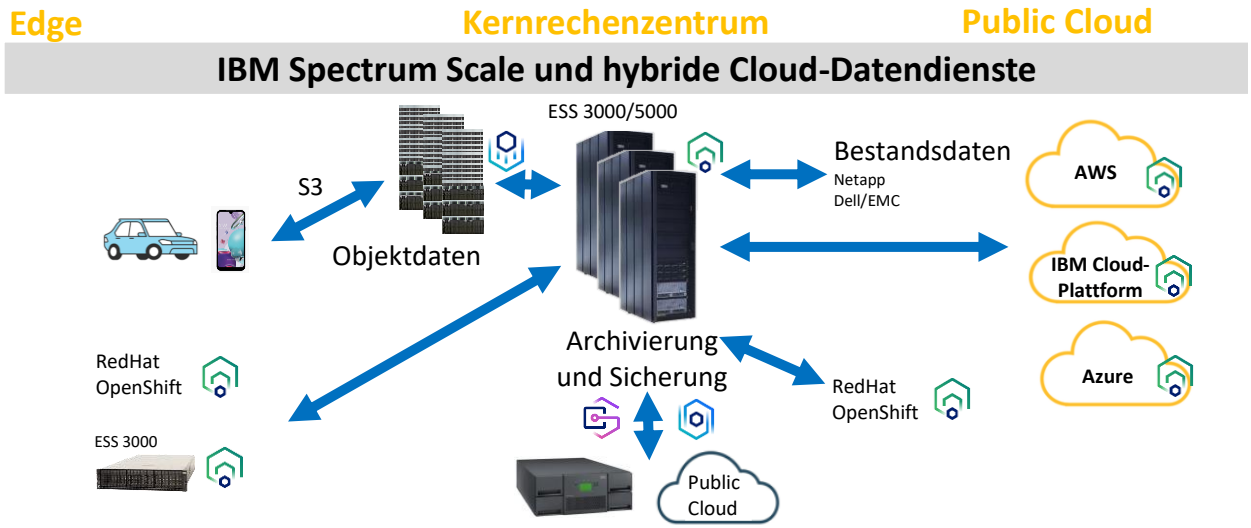
IBM Spectrum Scale basiert auf einer verteilten Computing-Architektur, die für leistungsintensive Workloads wie KI/ML, Modellierung und Simulation sowie Analytik konzipiert ist. Es handelt sich um ein gruppiertes, paralleles Dateisystem, das die Betriebskosten durch einfache Verwaltung und Skalierbarkeit sowie die Investitionskosten durch richtlinienbasierte Datenoptimierung und transparentes Datenlebenszyklusmanagement reduziert. IBM Spectrum Scale ermöglicht den gleichzeitigen Zugriff auf ein einzelnes Dateisystem oder eine Reihe von Dateisystemen von mehreren Knoten aus. Die Knoten können entweder direkt oder über das Netzwerk angeschlossen sein, eine Kombination aus direktem Anschluss und Netzwerkanschluss darstellen oder in einer gemeinsamen Nicht-Clusterkonfiguration betrieben werden. Diese skalierbare Lösung und hochverfügbare Plattform ermöglicht einen leistungsstarken gemeinsamen Zugriff auf gemeinsame Datensätze.

IBM Spectrum Scale unterstützt Datenreplikation, richtlinienbasiertes Speichermanagement und Multisite-Betrieb. IT-Betriebsteams können einen Cluster aus Kubernetes-Container-Knoten, IBM AIX-Knoten, IBM Z- oder LinuxONE-Knoten, Linux-Knoten, Microsoft Windows Server-Knoten oder einer Mischung aus allen fünf erstellen. IBM Spectrum Scale kann auf virtualisierten oder containerisierten Instanzen ausgeführt werden und bietet gemeinsamen Datenzugriff in Umgebungen, die logische Partitionierung oder andere Hypervisoren nutzen. Mehrere IBM Spectrum Scale-Cluster können Daten innerhalb eines Standorts oder über Wide Area Network (WAN)-Verbindungen für die globale Datenzusammenarbeit und den Datenzugriff gemeinsam nutzen. IBM Spectrum Scale bietet eine solide Grundlage für KI-Infrastrukturen, die auf jahrelanger Erfahrung in der High-Performance-Computing-Branche beruht (siehe Abbildung 2).

IBM Spectrum Scale bietet eine solide Grundlage für KI-Infrastrukturen, die auf jahrelanger Erfahrung in der High-Performance-Computing-Branche beruht.

ABBILDUNG 2

IBM Spectrum Scale und Dienste für hybride Cloud-Daten



Quelle: IDC, 2021

IBM Spectrum Scale bietet einen globalen Namensraum, gemeinsamen Dateisystemzugriff zwischen IBM Spectrum Scale Clustern, gleichzeitigen Dateizugriff von mehreren Knoten aus, hohe Wiederherstellbarkeit und Datenverfügbarkeit durch Replikation, die Möglichkeit, Änderungen vorzunehmen, während ein Dateisystem gemountet ist, und eine vereinfachte Verwaltung selbst in großen Umgebungen. Die wichtigsten Unterscheidungsmerkmale von IBM Spectrum Scale sind:

- Gemeinsamer Dateisystemzugriff zwischen IBM Spectrum Scale-Clustern ermöglicht die gemeinsame Nutzung von Daten zwischen separaten Clustern an einem Standort oder über ein WAN
- Verbesserte Systemleistung durch ein patentiertes paralleles Dateisystem, welches die Systemleistung verbessert
- Dateikonsistenz durch gleichzeitigen und detaillierten Zugriff auf Clients im gesamten Cluster mit Hilfe von Token-Management
- Erhöhte Datenverfügbarkeit und -zuverlässigkeit durch Funktionen wie die Protokollierung von Dateisystemprüfungen und konfigurierbare Funktionen wie intelligente Mounts, die große Entfernungen überbrücken können
- Verbesserte Systemflexibilität, die das Hinzufügen oder Löschen von Festplatten- oder Serverressourcen ermöglicht, während das Dateisystem eingebunden ist
- Vereinfachtes Speichermanagement für das Information Lifecycle Management (ILM) durch leistungsfähiges, richtliniengesteuertes, automatisiertes Tiered Storage Management von Flash zu HDDs zu Cloud zu Tape und sogar mit richtliniengesteuerter Datenreduzierung
- Vereinfachte Verwaltung über viele Standard-Dateisystemschnittstellen, die direkt aus den meisten Anwendungen ausgeführt werden können
- Bereitstellung einer hybriden Cloud, die Datenverfügbarkeit, -integrität und -sicherheit sowie optimierte Container-native Speicherung und Integration mit Red Hat OpenShift bietet.

IBM Elastic Storage System

IBM Elastic Storage System ist eine moderne Implementierung von softwaredefiniertem Speicher, die eine einfachere Konfiguration und Verwaltung von Bausteinen ermöglicht. Dies erleichtert IT-Organisationen die Bereitstellung von schnellem, hoch skalierbarem Speicher für leistungsintensive Datenverarbeitungsanwendungen, einschließlich KI-, Big Data- und Analyseanwendungen (siehe Abbildung 3). IBM ESS:

- Wurde mit dem NVMe-Flash-Speicher entwickelt, um Skalierbarkeit auf Exabyte-Ebene und konsistente Servicequalität in der gesamten Infrastruktur zu bieten
- Kann in die Dateiverwaltungs- und Datenservicefunktionen von IBM Spectrum Scale integriert werden, um ein föderiertes globales Speichersystem bereitzustellen
- Ermöglicht Unternehmen die Konsolidierung der Speicheranforderungen vom Rand bis zu den Kerndaten und die Integration mit der Public Cloud, wodurch die Ineffizienz verringert, die Anschaffungskosten gesenkt, die Speicherverwaltung vereinfacht und Datensilos beseitigt werden
- Bietet konsistenten Hochleistungszugriff für mehrere anspruchsvolle Anwendungen, die als Bare Metal oder in virtualisierten Umgebungen bereitgestellt werden (es unterstützt die Integration von Kubernetes und Red Hat OpenShift und erleichtert so die Bereitstellung von Cloud-nativen Anwendungen.)

ABBILDUNG 3

IBM Elastic Storage System

Quelle: IDC, 2021

IBM ESS ist in zwei Formfaktoren erhältlich - 3000 und 5000:

- IBM Elastic Storage System 3000 (ESS 3000) wurde entwickelt, um die Anforderungen an die Verwaltung von Daten für die Analyse zu erfüllen. Der in einem kompakten 2U-Gehäuse untergebrachte ESS 3000 verkürzt die Zeit bis zur Wertschöpfung für Anwendungen im Bereich künstliche Intelligenz/Deep Learning und leistungsintensive Rechenanwendungen - dank seines reinen NVMe-Speichers und der einfachen und schnellen containerisierten Softwareinstallation und -aufrüstung. Das Hardware- und Softwaredesign des ESS 3000 bietet dem Unternehmen

Zugang zu einer branchenführenden Leistung, die erforderlich ist, um die datenintensiven Rechenressourcen voll auszulasten.

- IBM ESS 5000 bietet skalierbare Petabyte-Kapazitätsknoten mit hohem Durchsatz, die den softwaredefinierten IBM Spectrum-Scale-Speicher mit IBM POWER9-Prozessor-basierten E/A-intensiven Servern kombinieren. Durch die Konsolidierung der Speicheranforderungen im gesamten Unternehmen auf dem IBM ESS 5000 und dem NVMe-basierten ESS 3000 können IT-Teams Ineffizienzen reduzieren, die Anschaffungskosten senken und die anspruchsvollen KI-, HPC-, Analyse- und/oder Hochleistungsspeicheranforderungen unterstützen, die in den Bereichen Gesundheitswesen, Medien, Behörden und Finanzdienstleistungen üblich sind. Der ESS 5000 kann mit Terabytes beginnen und auf Hunderte von Petabytes oder sogar Exabytes anwachsen, mit einem einzigen einheitlichen Namensraum, der kostspielige Datensilos eliminiert. IBM Spectrum Scale ist das parallele Dateisystem im Herzen von IBM ESS 5000, das den Durchsatz mit dem Wachstum des Systems steigert. Es ermöglicht die Integration mit früheren ESS-Modellen zum Investitionsschutz und bietet kostengünstigere Optionen wie Cloud-Speicher und IBM Tape. Mit IBM Spectrum Scale kann die IT-Abteilung Datensilos und Engpässe beseitigen, die Speicherverwaltung vereinfachen und einen schnelleren Zugriff auf Daten erhalten.

IBM Cloud Object Storage

IBM Cloud Object Storage (COS) ist eine branchenführende, softwaredefinierte, hoch skalierbare und kosteneffiziente Speicherlösung für die Speicherung unstrukturierter Daten am Rande, im zentralen Rechenzentrum und in privaten oder öffentlichen Clouds. IBM Cloud Object Storage ist ideal für die Bereitstellung oder Modernisierung von leistungsintensiven Infrastrukturen für KI, Analysen, IoT, Video und Bildspeicher. Darüber hinaus bietet es einen beispiellosen Wert, der es Unternehmen ermöglicht, die Speicherkosten um bis zu 12 % zu senken, indem sie neue 18-TB-SMR-Laufwerke (Shingled Magnetic Resonance) verwenden und gleichzeitig den Durchsatz in einem 12-Knoten-Cluster auf bis zu 55 GB/s erhöhen. Unternehmen können ihre Daten mit der lokalen oder geodispersen Datensicherung von IBM schützen, die für anspruchsvolle Data Lakes und große Kapazitätsanforderungen angepasst werden kann (siehe Abbildung 4).

ABBILDUNG 4

IBM Cloud Object Storage

IBM Cloud Object Storage

Quelle: IDC, 2021

IBM Cloud Object Storage ist ein grundlegendes System für KI-Infrastrukturen mit wichtigen Vorteilen wie:

- **Skalierbarkeit:** COS unterstützt exponentielles Datenwachstum und skaliert Leistung und Kapazität von Terabytes bis Exabytes.
- **Sicherheit:** COS verfügt über integrierte Verschlüsselung und richtlinienaktivierten, abschließbaren WORM-Speicher (Write Once, Ready Many).
- **Einfachheit:** COS kann von jedem Standort aus gleichzeitig auf Daten zugreifen und bietet automatisches Failover, Datenwiederherstellung, automatische Erweiterung und Rebalancing.
- **Effiziente Einsparungen:** COS liefert geografisch geschützte Daten mit der Effizienz des Information Dispersal Algorithm (IDA) und ist als reine Software- oder vollständig unterstützte Appliance-Lösung erhältlich.
- **Suchfunktionen:** COS bietet benutzerdefinierte Einblicke und Suchfunktionen, um Zeit zu sparen. Unternehmen können benutzerdefinierte Metadaten erstellen, um den Wert zu steigern
- **Verbesserter Dateizugriff:** Ein neues Software-Gateway für den Dateizugriff kann nahtlos an jedes Windows- oder Linux-Dateisystem mit SMB- oder NFS-Zugriff angeschlossen werden, so dass dateibasierte Anwendungen problemlos mit Objektspeichern verbunden werden können.
- **Hochgeschwindigkeitsübertragung:** Die IBM Aspera-Hochgeschwindigkeitsdatenübertragungsoption erleichtert die Datenübertragung, und flexible Speicherklassen helfen bei der Kostenkontrolle und erfüllen gleichzeitig die Anforderungen an den Datenzugriff.

AUSBLICK

Im Jahr 2022 werden insgesamt 65 % des globalen BIP digitalisiert sein, was zu IT-Ausgaben in Höhe von 6,8 Billionen US-Dollar im Zeitraum 2020-2023 führen wird (siehe *IDC FutureScape: Weltweite Prognosen zur digitalen Transformation 2021* IDC #US46880818, Oktober 2020). Die digitale Infrastruktur ist weder auf traditionelle zentrale Unternehmensdienste noch auf einzelne Cloud-Rechenzentren beschränkt. Sie umfasst alle Anlagen und Ressourcen, die die Verlagerung von Anwendungen und Code zur Umwandlung ermöglichen. Es wird die Gründung für ein verbessertes Kundenerlebnis sein. Außerdem ermöglicht sie die Einbettung von Intelligenz/Automatisierung in die Geschäftsabläufe und unterstützt fortlaufende Innovationen bis hin zu den digitalen Grenzen des Unternehmens und der Branche. Eine erfolgreiche digitale Strategie muss die digitale Infrastruktur umgestalten, um Silos zu beseitigen, technologiebedingte Barrieren abzubauen und über die Unterstützung herkömmlicher Tools und Anwendungen hinauszugehen.

IDC ist der Ansicht, dass KI die Grundlage der digitalen Infrastruktur sein wird. Die Customer Insights and Analysis Group von IDC hat kürzlich eine Umfrage durchgeführt, um die aktuellen und zukünftigen IT-Ausgaben und die Pläne für die Einführung von IT in Unternehmen aller Größen und aus verschiedenen Branchen zu ermitteln. Die Studie ergab, dass fast 76 % von 3.600 wichtigen IT-Entscheidungsträgern in IT-Organisationen aus verschiedenen Branchen weltweit angaben, dass künstliche Intelligenz ein wichtiger Bestandteil ihrer DX-Strategie ist oder in den nächsten ein bis zwei Jahren erwartet wird. Nur 22 % der Umfrageteilnehmer gaben an, dass KI in den nächsten drei bis fünf Jahren der wichtigste Bestandteil ihrer DX-Strategie sein wird. Von denjenigen, die erwarten, dass KI in den nächsten ein bis zwei Jahren eine Schlüsselkomponente ihrer DX-Strategie sein wird, werden Telekommunikations-, Versorgungs-, Bildungs- und Dienstleistungsunternehmen am ehesten auf KI bei ihren Bemühungen um digitale Transformation setzen.

Bei der KI geht es um „time to value“ - die Wertschöpfung, die Unternehmen in der schnellstmöglichen Zeit aus Daten erzielen können. Der FutureScape von IDC über künstliche Intelligenz schätzt, dass „künstliche Intelligenz die größte Innovation unseres Lebens ist.“ KI ist nicht mehr nur etwas, das man „haben sollte“. Die globale Pandemie hat den Einsatz von KI beschleunigt, und sie wird in allen Geschäftsprozessen allgegenwärtig. KI-Lösungen, die auf maschinellem Lernen, Konversations-KI und Computer-Vision basieren, stehen an vorderster Front, wenn es darum geht, die Widerstandsfähigkeit von Unternehmen zu erhöhen, Innovationen zu beschleunigen und neue Kunden- und Mitarbeitererfahrungen zu schaffen. Etwa 51 % der Etwa 51 % der Befragten in der oben genannten Umfrage gaben an, dass sie derzeit künstliche Intelligenz evaluieren oder bereits in Produktion sind; dies ist ein Anstieg gegenüber 34 % der Befragten im Jahr 2019. Die größte Auswirkung hat KI, wenn es darum geht, die Mitarbeiter bei ihrer Arbeit zu unterstützen. Die Verbreitung von KI in Unternehmen wird weiter zunehmen, da die Vorteile einer vollständigen Implementierung immer spürbarer werden.

Für 2020 wird ein Wachstum von 11,4 % für KI-Speicher prognostiziert. Für den gesamten Hardwaremarkt wird für 2021 eine starke Erholung mit einem Wachstum von 35,5 % gegenüber dem Vorjahr erwartet, angeführt von der KI-Speicherung, die im Vergleich zum Vorjahr um 43,1 % wachsen soll.

IDC schätzt, dass die Investitionen in die KI-Infrastruktur in den nächsten Jahren weiterhin stark sein werden. IDC geht davon aus, dass der Umsatz mit KI-Hardware (Server und Speicher zusammen) im Jahr 2020 13,4 Milliarden US-Dollar erreichen wird, was einem Wachstum von 10,3 % im Vergleich zum Vorjahr entspricht. Innerhalb des Hardware-Marktes wird für den Bereich KI-Speicher bis 2020 ein Wachstum von 11,4 % prognostiziert. Für den gesamten Hardwaremarkt wird für 2021 eine starke Erholung mit einem Wachstum von 35,5 % gegenüber dem Vorjahr erwartet, angeführt von der KI-Speicherung, die im Vergleich zum Vorjahr um 43,1 % wachsen soll.

Ein Großteil dieses Speichers wird in hybriden Cloud-Umgebungen mit einer nahtlosen Mobilitätsschicht eingesetzt werden, die es ermöglicht, die Arbeitslast vom Kern in die Cloud oder zum Edge und zurück zu verschieben. Die Speicherung wird eine entscheidende Grundlage für die Hybrid-Cloud sein, da sie die Verlagerung von Rechenleistung zu Daten unterstützt und gleichzeitig eine gemeinsame Zugriffs- und Steuerungsebene bietet.

Speicherung - und insbesondere Speicherung in hybriden Cloud-Umgebungen - wird auch weiterhin das Fundament sein, auf dem KI-Initiativen jetzt und in Zukunft skalieren können. Investitionen in KI und Datenmodernisierungsinitiativen aufgrund von KI werden Investitionen in Scale-out-Dateispeicher und unstrukturierte Daten vorantreiben. IDC befragte kürzlich 624 IT-Fachleute und Betriebsteams weltweit, um Trends bei der Einführung von IT-Infrastrukturen zu ermitteln. In dieser Studie fand IDC heraus, dass über 65 % der Befragten Scale-Out-Dateisysteme bevorzugen, auf die lokal oder über NFS für ihre Hochleistungs-Workloads wie KI zugegriffen wird. Speziell für KI-Workloads, zu denen auch Trainings- und Inferencing-Workloads gehören, war die Leistung die wichtigste Anforderung an den Speicher. Es folgten die einfache Bereitstellung in einer Hybrid-Cloud und die Servicequalität.

LEITFADEN FÜR IT-EINKÄUFER

Technologiekäufer sind zu Recht verwirrt über den Prozess des Aufbaus ihrer eigenen KI-Infrastruktur. Sie haben Anwendungsfälle definiert, KI-Initiativen gestartet und Datenwissenschaftler und Anwendungsentwickler eingestellt oder geschult, sehen sich aber plötzlich durch die Infrastruktur eingeschränkt, auf der sie KI-Modelle entwickeln können. Vielfach wird kurzzeitig auf die bestehende Infrastruktur zurückgegriffen, gefolgt von Investitionen in eine beschleunigte Infrastruktur.

Datenwissenschaftler stellen selbst Datenstapel auf der beschleunigten Infrastruktur zusammen und versuchen, sie zum Laufen zu bringen, was eigentlich nicht zu ihrer Aufgabenbeschreibung gehört. Die IT-Infrastrukturteams sind mit den Stacks, die Datenwissenschaftler benötigen, nicht vertraut und können sie nicht zusammenstellen und optimieren. Dies hat zu einer großen Qualifikationslücke geführt, die Serveranbieter und Cloud-SPs mit ihren eigenen Stacks zu füllen versuchen, jeder auf seine Weise. Heute gibt es so viele verschiedene Stapel, wie es Anbieter gibt, die sich oft überschneiden, wenn sie von mehreren Mitgliedern derselben Wertschöpfungskette entwickelt werden.

Beginnen Sie mit den Geschäftsergebnissen

Unternehmen müssen damit beginnen, Serviceeinschränkungen und Anwendungsfälle miteinander zu verknüpfen, um Geschäftsergebnisse zu ermitteln, die von einer Investition in eine KI-Infrastruktur profitieren. Sie müssen versuchen, den Nutzen solcher Investitionen zu quantifizieren und zu messen. Wenn man zum Beispiel eine Wettbewerbsdifferenzierung anstrebt, stellt sich die Frage, um wie viel und bis wann? Diese Kriterien sollten dann zur Auswahl einer Anwendungsarchitektur führen. Unternehmen müssen KI in Erwägung ziehen, wenn es darum geht, ihre Marke durch ein besseres Verständnis und eine gezielte Reaktion auf die Stimmung, Wünsche und Bedürfnisse der Kunden zu verbessern, was wiederum zu höheren Umsätzen und Gewinnen führt.

Verfolgen Sie einen ganzheitlichen Ansatz

Bei der Umsetzung einer KI-Initiative ist es wichtig, das Gesamtbild zu betrachten (d. h. eine umfassende Sichtweise einzunehmen). Die alleinige Betrachtung eines Problems kann dazu führen, dass (noch) ein weiteres Silo entsteht oder, schlimmer noch, die Komplexität in der Umgebung aufgrund mangelnder Integration und Interoperabilität zwischen verschiedenen Architekturen und Lösungen zunimmt. Eine Dateninfrastruktur muss als globale Lösung vom Edge über den Core bis zur Cloud und über Anwendungsfälle wie KI und andere unternehmenskritische Anwendungen hinweg betrachtet werden. Sie muss als globale Lösung dienen, die Anwendungsfälle wie KI und andere geschäftskritische Anwendungen im Core, Edge und in der Cloud bedient.

Entwickeln Sie die richtige Anwendungs- und Datenarchitektur

Die Entwicklung einer KI-Anwendung und einer Datenarchitektur ist eine komplexe Aufgabe. Dies beinhaltet die Umwandlung von Geschäftsanforderungen und -ergebnissen in einen deterministischen KI-gestützten Workflow. Der Workflow muss die Art und Weise beschreiben, wie KI-Funktionen das Verhalten dieser Anwendung verbessern, wie Daten aufgenommen und analysiert werden und wie die Anwendung mit anderen Geschäftsanwendungen und mit Benutzern interagiert. Der Fokus muss dabei auf der Art und Weise liegen, wie Daten von Anwendungen konsumiert, produziert und analysiert werden und welche Auswirkungen dies auf die Hardware hat. Bei der Erstellung eines Greenfield-Plans sollte der Schwerpunkt auf der Mischung zwischen benutzerdefinierten (Open-Source- oder proprietären) und handelsüblichen Softwarekomponenten liegen.

Wählen Sie den richtigen Referenzstapel

Mehrere Anbieter und Dienstleister haben Referenzstapel (Stacks) für die Implementierung einer KI-Infrastruktur herausgebracht. Viele davon sind "offen" und ermöglichen eine modulare "Plug-and-Play"-Erfahrung und können als Pay-as-you-go-Service für eine investitionsfreundliche Implementierung genutzt werden. Dies ist eine wichtige Überlegung, da KI-Infrastrukturinvestitionen schnell teuer werden können. IDC plant, seine Sichtweise auf die Referenzstapel der gängigen Anbieter in einem kommenden Dokument zu veröffentlichen.

Zu den IT-Vorteilen, die bei der Prüfung von Referenzstapeln zu berücksichtigen sind, gehören geringere Kosten, Daten- und Anwendungsverfügbarkeit, effektive Infrastrukturnutzung und -konsolidierung sowie, wenn möglich, eine einzige interoperable Plattform zur Anwendungsbereitstellung.

Informationsarchitektur für KI aufbauen

Unternehmen benötigen eine Datenverwaltungsstrategie, die einen flexiblen, organisierten Zugriff auf alle Daten jeder Art ermöglicht, unabhängig davon, wo sie gespeichert sind, und die die Probleme der Referenzarchitektur angeht. Im Rahmen einer Modernisierung würde eine Informationsarchitektur definiert und implementiert, die eine offene, erweiterbare Grundlage mit Wahlmöglichkeiten und Flexibilität bietet, die mit anderen Cloud-Plattformen kommunizieren kann. Die hybride Datenmanagement-Strategie von IBM zur Beschleunigung der Entwicklung von KI ist ein präskriptiver Ansatz, der durch eine vierstufige KI-Leiter definiert ist: Sammeln, Organisieren, Analysieren und Integrieren.

- **Sammeln:** Daten einfach und an der richtigen Stelle zugänglich machen, von jeder Datenbank oder Speichereinrichtung aus.
- **Organisieren:** Sicherstellen, dass die Daten in allen Phasen des Informationslebenszyklus vertrauenswürdig, vollständig und konsistent sind: Profilierung, Bereinigung und Katalogisierung der Daten; Schutz und Einhaltung von Vorschriften; richtliniengesteuerte Transparenz, Erkennung und Berichterstattung.
- **Analysieren:** Erstellen, bereitstellen und verwalten Sie KI-Modelle mithilfe integrierter Tools, um sowohl strukturierte als auch unstrukturierte Daten zu untersuchen und zu analysieren und sie sicher bereitzustellen.
- **Integrieren:** Vertrauen und Transparenz in die vom Modell empfohlenen Entscheidungen schaffen, Entscheidungen erklären, Voreingenommenheit aufdecken usw., unter Verwendung der angebotenen Lösungen und Dienste.

Auf dem Weg zur KI geht es darum, Daten von der Erfassung zu Erkenntnissen zu führen, und zwar mit einer Informationsarchitektur, die problemlos im gesamten Unternehmen eingesetzt werden kann. Es ist wichtig, dass jeder Teil der KI-Leiter eine Verbindung zum gesamten Weg herstellt. Die Speicherung wurde in der Regel auf taktische Art und Weise mit spezifischen Speicherlösungen implementiert, die zu Datensilos und Lösungen führen, die nicht miteinander oder mit einem umfassenden Satz von Infrastrukturlösungen integriert sind. Kunden können Daten auf einem großen Datei- oder Objektspeichersystem speichern, verfügen dann aber nicht über Details zu diesen Daten oder nutzen diese Daten nicht für zusätzliche Erkenntnisse. Kunden können immer noch ein Projekt starten oder sich auf einen Teil der Strecke konzentrieren, aber jedes Projekt sollte eine übergreifende KI-Informationsarchitektur berücksichtigen, um Ressourcen zu optimieren und Ihre Infrastruktur für wachsende KI-Workloads zu modernisieren.

Unternehmen, die KI als hochleistungsfähige Composite-Applikation (bestehend aus mehreren miteinander verbundenen Anwendungen) behandeln und dabei auf wichtige Elemente wie Scale-Out-Dateisysteme, heterogenes Computing und Hochgeschwindigkeitsverbindungen für verteilten Rechen- und Speicherzugriff zurückgreifen, sind diejenigen, die ihre KI-Infrastruktur letztendlich skalieren können.

Nutzen Sie die richtigen Partnerschaften

Für IT-Einkäufer ist die Partnerschaft mit einem Anbieter von End-to-End-Lösungen entscheidend für den langfristigen Erfolg. IDC ist jedoch der Meinung, dass derzeit noch kein Anbieter auf dem Markt eine solche End-to-End-Umgebung bietet, auch wenn die Anbieter hart daran arbeiten, dies zu erreichen. Dennoch können Unternehmen durch die Zusammenarbeit mit einem vertrauenswürdigen

Partner KI-Ansätze besser nutzen, um ihr Geschäft auszubauen. Sie können schnell und flexibel reagieren und interne Synergien nutzen, um die Rentabilität zu steigern. Schließlich können sie den disruptiven Kräften in ihrer Branche voraus sein, indem sie sich neu erfinden. Ein idealer Partner würde Folgendes bieten:

- Bewährte Lösungen, die von kleinen Laborinstallationen bis hin zu großen globalen Installationen skalierbar sind
- Vertikale Segmentexpertise, die sich auf Segmente konzentriert, die dem Geschäftsschwerpunkt des Unternehmens entsprechen
- Integrierten Zugang zu einem vielfältigen Ökosystem von ISVs und Infrastrukturanbietern
- Eine datenorientierte Sichtweise, die sicherstellt, dass Sie den langfristigen Wert Ihrer Investitionen in neue Datenquellen sichern und maximieren
- Nachgewiesene Erfolge bei der Vereinfachung der Hardware-, Software- und Sicherheitsaspekte von Großprojekten

HERAUSFORDERUNGEN/MÖGLICHKEITEN

Für Unternehmen

In diesem White Paper wurde eine Reihe von Herausforderungen erörtert, mit denen Unternehmen konfrontiert sind, wenn sie ihre KI-Anwendungen für die Produktion skalieren wollen. Von der Datenaufbereitung über die Modellentwicklung bis hin zu Laufzeitumgebungen für das Training, die Bereitstellung und die Verwaltung von KI-Modellen - die Anforderungen an die zugrunde liegende Infrastruktur widersprechen den alten Modellen der Allzweckhardware. Investitionen in eine Infrastruktur, die für datenintensive Workloads ausgelegt ist, über hervorragende Leistung, Skalierbarkeit, Datenzugriff und Integration verfügt und sich in eine hybride Cloud-Umgebung einfügen lässt, bieten langfristigen Wert und Servicequalität. Unternehmen müssen entscheiden, ob sie bestehende Allzweck-Speicherplattformen durch Speichersysteme ersetzen oder ergänzen, die auf KI-spezifische Verarbeitungsaufgaben ausgerichtet sind. Dies wird als Grundlage für die Entwicklung und den Betrieb modernster KI-Anwendungen dienen.

Für IBM

Die Herausforderung für IBM besteht immer in der Anerkennung auf dem Markt. IBM bietet außergewöhnliche und umfassende KI-Infrastrukturlösungen, die in Infrastruktur-Software-Stapel (z. B. Red Hat OpenShift) und die Public Cloud integriert sind, die aber von potenziellen Kunden - fälschlicherweise - als komplex oder kostspielig angesehen werden. Die anschließende reflexartige Reaktion, für extrem datenintensive Workloads einen der großen Anbieter von Standardspeichersystemen zu wählen, beraubt diese Unternehmen der KI-Infrastrukturlösungen, von denen sie wirklich profitieren könnten. Die IBM Storage for Data and AI-Lösungen mit IBM Spectrum Scale und IBM Elastic Storage System sind beispielsweise ein Supercomputing-Baustein für viele der größten Rechenzentren. Jetzt, da neue KI-Workloads beginnen, Supercomputing-Implementierungen mit HPC-Anforderungen ernsthaft zu imitieren und die On-Premises- und Cloud-Infrastruktur, auf der sie laufen, herauszufordern, ist dies der Moment für IBM, die Bühne zu betreten und neue Kunden zu gewinnen.

FAZIT

In den letzten Jahren hat IDC die KI-Transformation in vielen Unternehmen beobachtet, die damit begonnen haben, ein breites Spektrum an KI-Funktionen zu entwickeln. Diese Initiativen, die zunächst als Experimente von relativ unerfahrenen Mitarbeitern gestartet und mit der verfügbaren Infrastruktur durchgeführt wurden, haben inzwischen eine entscheidende Masse erreicht. Viele Unternehmen haben umfassende KI-Expertise aufgebaut und erleben aus erster Hand, wie schnell ihre KI-Fähigkeiten zu einem entscheidenden Aspekt ihres Geschäfts werden.

Heute besteht die wichtigste Entscheidung, die die IT-Abteilung treffen muss, darin, wie sie ihre KI-Anwendungen für die Dateninfrastruktur am besten konzipiert und einsetzt und wie sie diese miteinander verbindet und optimiert.

Gleichzeitig hat sich auch die IT durch eine Lernkurve im Hinblick auf die Infrastruktur, auf der KI laufen soll, gewandelt. Inzwischen gibt es viel mehr Klarheit über die Infrastrukturanforderungen für Deep Learning Training oder Inferencing und darüber, wie diese Umgebungen für die Produktion skaliert werden können. Dass Deep Learning-Training eine andere Infrastruktur erfordert als andere Anwendungen, ist so gut wie sicher. Deep Learning-Training erfordert Cluster-Knoten mit starken Prozessoren, leistungsstarken Coprozessoren, skalierbaren, speicherfähigen, schnellen Verbindungen, großer E/A-Bandbreite und viel Speicher.

Heute besteht die wichtigste Entscheidung, die die IT-Abteilung treffen muss, darin, wie sie ihre KI-Anwendungen für die Dateninfrastruktur am besten konzipiert und einsetzt und wie sie diese miteinander verbindet und optimiert.

IDC ist der Ansicht, dass die IBM Speicherlösungen für Daten und KI, zu denen Spectrum Scale, Elastic Storage System und Cloud Object Storage gehören, einen beispiellosen Wert und eine beispiellose Leistung bieten.

Über IDC

Die International Data Corporation (IDC) ist der führende weltweite Anbieter von Marktinformationen, Beratungsdienstleistungen und Veranstaltungen für den IT-, Telekommunikations- und Technologiesektor. IDC unterstützt IT-Spezialisten, Führungskräfte und die Investment-Community bei faktenbasierten Entscheidungen zu Technologiekäufen und der Entwicklung von Geschäftsstrategien. Über 1 100 Analysten von IDC stellen globales, regionales und lokales Know-how zu Technologie und branchenspezifischen Geschäftschancen und Trends in über 110 Ländern weltweit zur Verfügung. IDC liefert seit über 50 Jahren strategische Erkenntnisse, um Kunden dabei zu unterstützen, die wichtigsten geschäftlichen Ziele zu erreichen. IDC ist eine Tochtergesellschaft von IDG, einem der weltweit führenden Unternehmen aus dem Bereich Technologiemedien, -forschung und -events.

Hauptgeschäftsstelle

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com