MASCHINELLES LERNEN UND DEEP LEARNING DER ENTERPRISE-KLASSE MIT INTELLIGENTEM SPEICHER

ZUSAMMENFASSUNG

Befeuert durch Daten, Fortschritte bei der Infrastruktur und die Allgegenwärtigkeit von Tools für maschinelles Lernen und Deep Learning (ML/DL) haben sich KI-Lösungen (künstliche Intelligenz) schnell zu einer wichtigen Stütze im Enterprise-Rechenzentrum entwickelt. KI verwandelt Daten in Erkenntnisse – in einer Fülle verschiedenster vertikaler Unternehmensbranchen von der Automobilindustrie über das Gesundheitswesen, Life Sciences und Finanzen bis hin zu Technologie, Einzelhandel und mehr. Daten sind jetzt ein Wettbewerbsvorteil in Branchen wie Versicherungen, wo vorausschauende KI-Anwendungen Versicherungsrisiken beseitigen, Finanzen, wo Deep Learning in Echtzeit Betrugsfälle erkennt, sobald diese stattfinden, und sogar Rechenzentrumsmanagement, wo Muster analysiert werden, um Ausfälle und Skalierbarkeitsprobleme vorherzusehen.

Künstliche Intelligenz und vor allem Deep Learning bringen neue Anforderungen mit sich, wenn es darum geht, wie Daten für die Compute Engines zur Verfügung gestellt werden, die diese nutzen. Die neuen Gegebenheiten bei der Bereitstellung von künstlicher Intelligenz im Rechenzentrum ändern die Anforderungen an Dichte, Durchsatz, Parallelität und sogar die Scale-out-Datenarchitektur. Die IT muss hinsichtlich der Kombination aus Speicher und Compute umdenken, um das KI-Versprechen für das Unternehmen einzulösen.

In diesem Whitepaper wird beschrieben, welche neuen Workflows und Herausforderungen für die Rechenzentrumsarchitektur Deep Learning und künstliche Intelligenz im Unternehmen mit sich bringen. Außerdem wird erläutert, wie Lösungen aus Infrastrukturarchitekturen erstellt werden können, die speziell darauf ausgelegt sind, Scale-out Compute und -Speicher näher zusammenzubringen.

Bei Deep Learning müssen große Datenmengen in den Prozessor eingespeist werden, ohne dass Prozessoren auf diese Daten warten müssen. Durch die ordnungsgemäße Kombination von Compute mit der richtigen Speichertechnologie wie der Dell EMC Isilon-Serie können Daten mit der Geschwindigkeit des Prozessors

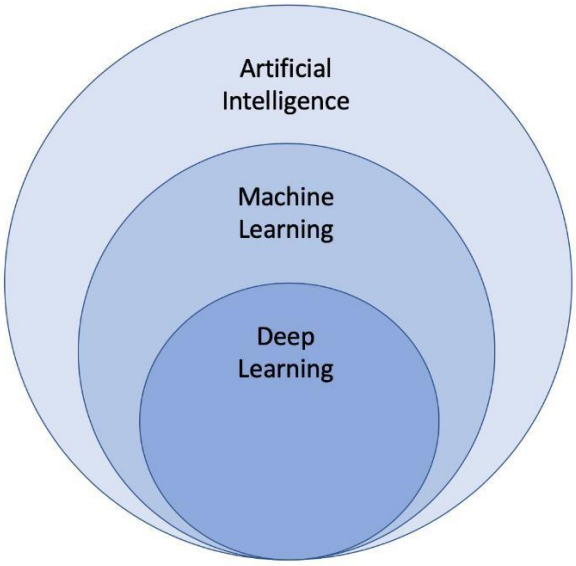
in die Pipeline für maschinelles Lernen eingespeist werden. Korrekt ausbalancierte Systeme beschleunigen Innovationen und bieten Flexibilität und Agilität sowohl für IT-Abteilungen als auch die Data Scientists, die diese Systeme nutzen.

DEEP LEARNING ÄNDERT UNTERNEHMEN

Wahrscheinlich hat mittlerweile jeder das berühmte Titelbild der Zeitschrift „The Economist“ gesehen, auf dem es vor dem Hintergrundbild einer Bohrinsel heißt, dass Daten „die wertvollste

Ressource der Welt“ sind.[1] Fortschritte bei maschinellem Lernen (ML) und Deep Learning (DL) haben in der Tat dazu geführt, dass die Daten jedes Unternehmens neue Bedeutung gewonnen haben – *Daten sind zu einem Wettbewerbsvorteil geworden*.

ABBILDUNG 1: DIE BEZIEHUNG ZWISCHEN KI, ML UND DL



Quelle: Moor Insights & Strategy

Künstliche Intelligenz beschreibt eine allgemeine Klasse von Technologien, bei denen Computer Entscheidungen treffen oder Erkenntnisse bereitstellen, die normalerweise mit menschlicher Intelligenz verknüpft sind. Ein einfaches Beispiel für KI ist eine Empfehlungs-Engine im Einzelhandel, die auf der Website eines Einzelhändlers zu finden ist und Produktempfehlungen bereitstellt, die auf Ihren bisherigen Einkäufen und Ihrer aktuellen Suche basieren.

[1] The Economist, The World’s Most Valuable Resource is No Longer Oil, But Data. 6. Mai 2017.

Maschinelles Lernen ist eine Art künstlicher Intelligenz, bei der Daten von Algorithmen analysiert werden. Das System lernt aus diesen Daten und wendet das Gelernte dann in realen Umgebungen an, um Entscheidungen zu treffen. Ein auf maschinellem Lernen basierendes System zur Erkennung von Spam-E-Mails kann beispielsweise mit Mustern aus den Millionen von E-Mails trainiert werden, die Nutzer täglich in ihre Spamordner verschieben.

Deep Learning, einer der interessanteren und aktiveren Bereiche künstlicher Intelligenz, ist eine Untergruppe des maschinellen Lernens. Deep Learning verwendet als *neuronale Netze* bezeichnete Algorithmen, mit denen

Vorhersagen kontinuierlich anhand der aufgenommenen Daten optimiert werden. Deep Learning steht im Zentrum von autonomen Fahrzeugen, Stimmungsanalysen, die menschliche Stimmungen erkennen, und den meisten anderen KI-Techniken, die reale Daten verarbeiten, um dynamische Entscheidungen zu treffen oder Empfehlungen abzugeben. Auch wenn es in diesem Whitepaper hauptsächlich um DL geht, bestehen architekturbezogene Ähnlichkeiten bei der Implementierung jeder Art von datenintensivem KI-System.

Es gibt zahlreiche und vielfältige KI-Anwendungsbeispiele, die von KI-gesteuerten Chatbots und Sprachdialogsystemen über die Prognostizierung des Kundenverhaltens bis hin zu einem optimierten Lieferkettenmanagement reichen. Wir sind im Zeitalter des intelligenten Unternehmens angekommen und viele Unternehmen fühlen sich überwältigt, wenn es darum geht zu ermitteln, wie sie Technologien für maschinelles Lernen am besten nutzen können, um einen Wettbewerbsvorteil in ihrer Branche zu erzielen.

Laut einer kürzlich durchgeführten Umfrage[2] unter mehr als 1.300 IT-Fachkräften arbeiten mehr als 60 % der Befragten für Unternehmen, die mindestens 5 % ihres IT-Budgets für künstliche Intelligenz ausgeben möchten. Ein Fünftel dieser Befragten arbeiten für Unternehmen, die planen, mehr als 20 % ihres IT-Budgets für KI auszugeben – eine überwältigende Zahl.

Künstliche Intelligenz mit Deep-Learning-Techniken wirkt sich auf jedes Unternehmen aus, oft auf unerwartete Weise. Im Folgenden sind nur einige Beispiele dafür aufgeführt, welche Auswirkungen maschinelles Lernen und sein speziellerer Ableger Deep Learning auf das moderne Unternehmen haben:

[2] O'Reilly Media, The State of Machine Learning Adoption in the Enterprise, 2019

- Die **Medien- und Unterhaltungsbranche (M&U)** nutzt maschinelles Lernen, um eine Reihe von Aufgaben durch Intelligenz zu unterstützen. Stimmungsanalysen werden verwendet, um Reaktionen des Publikums auf Filme und Fernsehsendungen zu klassifizieren. Die M&U-Branche nutzt außerdem eine durch Deep-Learning-Algorithmen trainierte Bilderkennung für die automatisierte Metadatenerzeugung für enorme Mengen an Videoinhalten.
- Moderne **Fertigungsvorgänge** in verschiedenen Branchen setzen bei vielen Aspekten ihres Betriebs künstliche Intelligenz und maschinelles Lernen ein. Bilderkennungssysteme analysieren Produkte in Fertigungslinien, um Fehler zu erkennen. Auf maschinellem Lernen basierende Systeme unterstützen zudem die vorausschauende Fehleranalyse, indem Sensordaten aus dem gesamten Werk analysiert werden, um Muster zu erkennen und zu identifizieren, die zu Ausfällen führen können, wenn nicht darauf reagiert wird. Maschinelles Lernen wird außerdem verwendet, um Lieferkettenentscheidungen zu unterstützen, die einen Just-in-Time-Betrieb mit intelligenter Beschaffung und Logistik optimieren.
- Die **Automobil- und Transportbranchen** nutzen Deep Learning, um unsere bisherige Wahrnehmung rund um das Auto zu revolutionieren. Deep-Learning-Techniken unterstützen den Wettlauf, dessen ultimatives Ziel das autonome Fahren ist.

Auf dem Weg dorthin sehen wir immer mehr reale Anwendungen in Form von intelligenten und adaptiven Geschwindigkeitsreglern, halbautomatischen Fahrzeugen, vorausschauender Fehleranalyse und sogar Fahrerüberwachung, um sicherzustellen, dass einem Fahrer bewusst ist, was im Fahrzeug geschieht. Nichts davon wäre ohne die kontinuierlichen Fortschritte rund um maschinelles Lernen und Deep Learning möglich.

Auch wenn die Bereitstellung von KI-Technologien im Unternehmen sehr wirkungsvoll ist, ist all das für die meisten Unternehmen noch Neuland. Deshalb ist es vor Beginn eines KI-Projekts wichtig, die gemeinsamen Bausteine anzusehen und zu vereinfachen. Für nahezu alle DL-Lösungen – ob unterstützende Bilderkennung, Bildklassifizierung, Segmentierung, Verarbeitung natürlicher Sprache und/oder vorausschauende Analysen – wird ein gemeinsamer Satz an Kerntechnologien verwendet. Diese Techniken werden auf Plattformen bereitgestellt, die eine native Unterstützung für gängige, bei der Implementierung dieser Anwendungsbeispiele allgegenwärtige Softwarepakete wie TensorFlow, PyTorch und Caffe2 bieten und für diese optimiert sind.

Die meisten herkömmlichen IT-Experten verfügen nicht über die erforderlichen Fähigkeiten, um KI-Lösungen auf effiziente Weise zu konzipieren und für sehr unterschiedliche Anwendungsbeispiele bereitzustellen. Maschinelles Lernen und Deep Learning sind Technologien, die neue Herausforderungen mit sich bringen und innovative Denkweisen rund um Daten erfordern.

Die oben schon erwähnte Umfrage zeigt, dass fehlendes Verständnis rund um die Bereitstellung von Deep Learning, kombiniert mit einer Infrastruktur, die für diese Workloads nicht bereit ist, die größte Hürde für eine erfolgreiche Einführung ist.

Die wichtigste Aufgabe für jede wettbewerbsorientierte IT-Abteilung besteht darin, diese Lücke zu schließen und die für die Bereitstellung von Deep Learning erforderlichen Kompetenzen zu entwickeln – unterstützt durch flexible und zukunftssichere Analysearchitekturen.

Architektur für Deep Learning im Rechenzentrum

Um die mehrdimensionalen Auswirkungen von Deep Learning auf die Speicherarchitektur verstehen zu können, ist ein allgemeines Verständnis für einen typischen Lernworkflow erforderlich. Jede Phase in der Lernpipeline stellt unterschiedliche Anforderungen an die zugrunde liegende Infrastruktur. Dies ist in Abbildung 2 dargestellt.

ABBILDUNG 2: TYPISCHE PIPELINE IM BEREICH MASCHINELLES LERNEN/DEEP LEARNING

	INGEST IOT, Logs, Sensors, Users, Etc	DATA PREP CPU-intensive Servers	REFINE GPU-enabled Server & Workstations	TRAIN High Performance GPU-based Servers	DEPLOY CPU or Inference Accelerated Edge, Client, or Server	RETENTION Long-term Storage
Access Pattern	Sequential	Sequential or Random	Random	Random	Random	Sequential
Access Type	Write	Read & Write	Read	Read	Read	Write
Concurrency	Variable	Low	Moderate	High	Low	Low
Performance	High	High	High	High	Moderate	Low
Storage	Block, File, or Object	Block or File	File or Object	File or Object	Block or Memory	Block, File, or Object
Scale	MB-GB	MB-TB	TB-PB	TB-PB	KB-MB	TB-PB

Quelle: Moor Insights & Strategy

Diese Schritte können wie folgt zusammengefasst werden:

- **Datenaufnahme:** Daten kommen aus einer externen Quelle (oder mehreren Quellen) wie Edge-Geräten, Protokolldateien, Sprach- oder Videostreams oder Customer-Relationship-Management-Systemen an. Die Daten gehen ein und werden gespeichert. Die Speicherlösung muss nur so leistungsfähig sein, wie es für die eingehenden Daten erforderlich ist.
- **Datenvorbereitung:** Die Daten werden für das Training bereinigt und transformiert. Mit diesem wichtigen Schritt wird sichergestellt, dass die Daten konsistent sind. Außerdem werden Ausreißer identifiziert und die Daten werden für die Trainingsalgorithmen optimiert. Für einige Arten von maschinellem Lernen, z. B. überwachtes Lernen, müssen die Daten in dieser Phase beschriftet werden.
- **Datenerkennung und -visualisierung:** Data Scientists arbeiten mit den Daten, um die Trainingsalgorithmen und -parameter zu optimieren. Dabei handelt es sich um einen sehr iterativen Prozess, für den jedoch nur eine mäßige Speicher- und Compute-Menge benötigt wird.

- **Modelltraining und -entwicklung:** Der Großteil der Arbeit findet in dieser Phase statt. Die bereinigten Daten werden in ein Cluster mit GPUs oder andere leistungsfähige Compute Engines eingespeist und dort oft über sehr lange Zeiträume iteriert. Das Training erfordert Speicher mit hohem Durchsatz, der für zufällige Lesevorgänge mit hoher Parallelität optimiert ist.
- **Modellbereitstellung** oder **Ableitung in der Produktion:** Die Modelle, die in der Trainingsphase erzeugt wurden, werden mit Daten aus der realen Welt bereitgestellt. Die Merkmale dieser Phase hängen stark von der Art des bereitgestellten Deep-Learning-Systems ab. Eine Bilderkennung kann beispielsweise auf einem Clientgerät wie einer intelligenten Kamera mit wenig Interaktion mit Speichersystemen stattfinden, während fortschrittlichere Anwendungen möglicherweise in einem Enterprise-Rechenzentrum ausgeführt werden.
- **Datenaufbewahrung:** Die beim Training des Modells oder der wiederholten Ableitung für Deep Learning verwendeten Daten werden zum Zweck der Archivierung oder erneuten Verwendung aufbewahrt. Das ist ein wichtiger Schritt. Durch die Archivierung der Daten wird sichergestellt, dass Modelle erneut erstellt und die Daten für zukünftige Erkenntnisse genutzt werden können.

Diese Schritte werden durch eine Reihe von Designprinzipien unterstützt, die bei der Implementierung einer Infrastruktur zur Unterstützung von Deep Learning berücksichtigt werden müssen:

- **Performance und Skalierung:** Die Performance darf sich bei Skalierung nicht verschlechtern. Jede Komponente – ob Compute, Speicher oder Netzwerk – sollte linear und unabhängig skaliert werden können, damit das System nahtlos mit der Workload wachsen kann und Compute-, I/O- und Netzwerkengpässe vermieden werden.
- **Flexibilität:** KI-Systeme werden rund um Daten aufgebaut. Die Realität dieser Dynamik besteht darin, dass sich Software, Analysetechniken und Anwendungsbeispiele unweigerlich ändern werden, wenn sich die KI-Umgebung weiterentwickelt, die Daten eines Unternehmens jedoch relativ konstant bleiben. Systeme sollten einen langlebigen Datenspeicher unterstützen, aber gleichzeitig die Flexibilität bieten, mit den Änderungen der Geschäftsanforderungen Schritt zu halten.
- **Enterprise-Datenmanagement:** Von Deep Learning genutzte Daten sind trotz der nicht herkömmlichen Verwendung Unternehmensdaten und sollten als solche gemanagt werden. Sicherheit, Data Protection, Compliance und andere herkömmliche Datenmanagementbedenken gelten auch für Deep-Learning-Daten. In diesen Umgebungen bereitgestellte Speicherlösungen müssen gut in die vorhandenen Policies und Verfahren für das Datenmanagement im Unternehmen integriert werden können.

Auch wenn es in Diskussionen rund um maschinelles Lernen und Deep Learning oft vor allem um Compute geht, ist offensichtlich, dass bei diesen Lösungen unweigerlich neu über Daten nachgedacht werden muss. Bei Deep Learning müssen wir überdenken, wie Daten gemanagt, analysiert und gespeichert werden.

DATEN IN EINER DEEP-LEARNING-UMGEBUNG

Daten in einem Deep-Learning-Workflow weisen andere Merkmale als die Daten in den meisten anderen IT-Anwendungen auf:

- **Daten sind hauptsächlich unstrukturiert** und bestehen aus Bildern, Audio, Freitext oder sogar Streams mit Zeitreihendaten. Die Speicherarchitektur für eine Deep-Learning-Umgebung muss für unstrukturierte Daten optimiert sein. Der Speicher sollte außerdem mehrere Datenzugriffsprotokolle wie SMB, NFS, HDFS, S3 und HTTP unterstützen, um größtmögliche betriebliche Flexibilität bereitzustellen.
- Bei Video- und Edge-Sensoren nimmt die **Skalierung der Daten** erheblich zu, insbesondere bei Inhalten mit höherer Auflösung, die viele Terabyte an Daten für Analysen über kurze Zeiträume erzeugen. Eine Aufbewahrung dieser Daten für eine spätere Analyse oder ein erneutes Training kann Petabyte an Speicher erfordern. Um zuverlässige Erkenntnisse aus DL ziehen zu können, ist ein umfassender Verlaufsdatensatz für die Analyse erforderlich. Speicherlösungen in dieser Umgebung sollten die Möglichkeit bieten, ein einfaches und unterbrechungsfreies Scale-out durchzuführen.
- **Die Datennutzung variiert erheblich**, da für jede Phase der Lernpipeline unterschiedliche Anforderungen bestehen. Für das Bereinigen oder Beschriften von Daten bestehen beispielsweise ganz andere Performanceanforderungen als für die Prozesse, die dieselben Daten für das Training oder das Ableiten in Echtzeit in ein Cluster einspeisen. An einem Ende dieser Pipeline kann herkömmlicher lokaler Speicher, DAS oder Mid-Tier-Speicher ausreichen. Am anderen Ende der Pipeline sind ein hoher Durchsatz und Enterprise-Funktionen erforderlich, die mit moderner Verarbeitungstechnologie Schritt halten können.
- **Daten kommen von überall.** Die Daten für Deep-Learning-Anwendungen stammen aus sehr verschiedenen Quellen. Daten für Analysen oder die Erzeugung von Modellen können von der Edge, aus Cloud-nativen Anwendungen, von Sprachservices und sogar aus Serverprotokoll-Aggregationsanwendungen stammen. Der Speicher muss so entworfen werden, dass Daten aus verschiedenen Quellen aufgenommen werden können.
- **Der Lebenszyklus von Datenmodellen.** KI erfordert eine konsistente Sammlung von Managementtools, die die gesamte Bandbreite von hoher Performance bis zu umfangreichem Archivspeicher abdecken, sodass die Daten in einer Speicherarchitektur gespeichert werden, die auf den allgemeinen KI-Workflow eines Unternehmens abgestimmt ist. Um vorhandene Daten in

Eingaben für neue KI-Funktionen umzuwandeln, sind auf ähnliche Weise Datenmanagementtools erforderlich, die es einer IT-Abteilung ermöglichen, neue Lösungen mit vorhandenem Speicher bereitzustellen.

Diese übergeordneten Merkmale werden bei der Auswahl einer Datenmanagementlösung für Deep Learning zu echten Überlegungen. Es ist wichtig zu betonen, dass diese Daten nach wie vor „Unternehmensdaten“ sind, die vor Hardware- und Softwareausfällen geschützt, gegen Sicherheitsverletzungen abgesichert und effizient gemanagt werden müssen.

Welche Art von Deep Learning ein Unternehmen bereitstellt, wirkt sich auch auf die Speicherarchitektur aus, die diese Workflows unterstützt. Die Bilderkennung beispielsweise, die in Branchen wie Medien und Unterhaltung, Fertigung und Automobil stark genutzt wird, basiert auf der Anwendung von CNN (Convolutional Neural Networks) und DNN (Deep Neural Networks).

CNN ist eine Art neuronales Netz, das mithilfe einer Reihe von sehr repetitiven Schritten lernt, Bilder zu klassifizieren und zu erkennen. Die Datenzugriffsmuster für CNNs benötigen sowohl während des Trainings als auch der Erkennung eine Speicherarchitektur, die für eine sehr große Anzahl von Lesezugriffen mit kleinen Blöcken auf das zugrunde liegende Speicherarray optimiert ist.

Ein Beispiel aus der realen Welt: Bei einem Benchmarking von Dell EMC und NVIDIA wurde ein Dell EMC Isilon F800-Speichersystem mit NVIDIA DGX-1-Servern kombiniert, die aus mehreren NVIDIA Tesla V100-GPUs bestanden. Jede GPU führte mehr als 5.000 parallele Threads aus, was einem Durchschnitt von 703 gleichzeitigen Dateilesevorgängen pro GPU entspricht[3]. Es ist wichtig, dass ein Speichersystem, das mit einem Deep-Learning-System kombiniert wird, Daten in großem Umfang und mit extremer Parallelität bereitstellen kann, ohne Unterbrechungen bei den Verarbeitungselementen zu verursachen, weil diese auf Daten warten müssen.

Das ist nur ein Beispiel. Bei anderen Deep-Learning-Systemen bestehen auch andere Anforderungen. Intelligente Systeme, die beispielsweise eine Mustererkennung in Echtzeit für die Erkennung von Finanzbetrug bereitstellen, benötigen möglicherweise einen sehr leistungsfähigen Blockspeicher. Anwendungen mit diesen Einschränkungen sollten möglicherweise eher von Blockspeicherarrays mit hohem Durchsatz und niedriger Latenz wie der Dell EMC PowerMax-Serie unterstützt werden.

[3] Whitepaper: Dell EMC Isilon and NVIDIA DGX-1 Servers for Deep Learning, https://www.dellemc.com/de-de/collaterals/unauth/white-papers/products/storage/Dell_EMC_Isilon_and_NVIDIA_DGX_1_servers_for_deep_learning.pdf.

Ähnliche Überlegungen bestehen rund um Blockgrößen, Datei-I/O-Muster und Skalierung. Der entscheidende Aspekt ist, dass die Bereitstellung von Daten für maschinelles Lernen und Deep Learning sich sehr von anderen Enterprise-Workloads unterscheidet. Das Managen von Daten für Deep Learning erfordert die Bereitstellung von Lösungen, die für hohe Parallelität und mehrdimensionale Performance im großen Maßstab entwickelt wurden und ein Tiering über einen einzigen Namespace sowie ein einfaches Management durch eine konsistente Sammlung von Tools bereitstellen.

DELL EMC: BEREITSTELLUNG VON SPEICHER FÜR DEEP LEARNING

Die Vorteile von KI können nur durch eine effiziente und leistungsfähige Bereitstellung von Daten realisiert werden. Deshalb müssen bei der Entwicklung von Speicherlösungen für Anwendungen mit maschinellem Lernen und Deep Learning, bei denen in den verschiedenen Phasen der Lernpipeline unterschiedliche Anforderungen an Performance, Skalierung und Parallelität bestehen, mehrere Faktoren berücksichtigt werden.

Gleichzeitig ist es sinnvoll, Speicherarchitekturen bereitzustellen, die nahtloses Tiering und Skalieren ermöglichen, um die Anforderungen aller Phasen einer Deep Learning Workload zu erfüllen.

Die Dell EMC Isilon-Produktreihe bietet eine solide Basis für die Bereitstellung von Speicherfunktionen zur Unterstützung des gesamten Lebenszyklus von Enterprise Deep Learning. Diese folgt dem Workflow vom Training über das Lernen und die Bereitstellung bis hin zu langfristigen Archivierungsanforderungen.

DELL EMC ISILON ONEFS

Die Leistungsstärke jedes Speichersystems basiert auf der zugrunde liegenden Betriebssystemsoftware. Das Dell EMC Isilon OneFS-Betriebssystem bietet die Intelligenz, die den Dell EMC Isilon-Scale-out-NAS-Speicherlösungen zugrunde liegt.

Die leistungsstarken Funktionen und Merkmale von OneFS optimieren und vereinfachen den Datenspeicher am Core jedes auf künstlicher Intelligenz basierenden Workflows. Die Software bietet ein nahtloses Tiering und stellt gleichzeitig einen einzigen Namespace bereit. Mit der Software können Sie die Datenplatzierung managen, die Performance jedes Arrays basierend auf erkannten Datenverkehrsmustern optimieren und tunen sowie von einer unterbrechungsfreien und linearen Speicherskalierung profitieren. Das Dell EMC Isilon OneFS-Betriebssystem bietet jede dieser Funktionen.

Durch die Einfachheit des Speichermanagements können sich Data Scientists verstärkt auf das Management des maschinellen Lernprozesses konzentrieren, ohne sich über die Details der zugrunde liegenden Speicherinfrastruktur Gedanken machen zu müssen. Diese Einfachheit ermöglicht es IT-Administratoren außerdem, die richtige Mischung aus flexiblen und effizienten Speicherlösungen bereitzustellen, die die Anforderungen von maschinellem Lernen und Deep Learning abdecken.

- **Konsolidierter Data Lake:** Konsolidiert Daten für den gesamten Analyseworkflow an einem Ort, um Data-Analytics-Pipelines zu vereinfachen.
- **Multiprotokollunterstützung:** Ermöglicht Analysen am Standort der Daten, um eine Methodik nach dem Motto „Einmal speichern, mehrfach verwenden“ zur Verbesserung der Flexibilität zu unterstützen.
- **Enterprise Data Governance:** Schützt Daten durch native Ausfallsicherheits- und Sicherheitsfunktionen.
- **Nahtloses Tiering:** Ermöglicht ein Tiering von Speicher zwischen All-Flash-, Hybrid- und Archiv-Nodes im selben Cluster, um eine wirtschaftliche Petabyte-Skalierung und einen Zugriff auf größere Datenvolumen zu ermöglichen.
- **Intelligentes Caching:** Bietet die Möglichkeit, die Caching-Eigenschaften des Speichersystems basierend auf den Workloads, die Daten nutzen, dynamisch zu optimieren. Das Isilon OneFS-Caching zielt auf eine parallele Leseperformance ab, die ein entscheidendes Performancemerkmal bei Deep-Learning-Workflows ist.
- **Lineare Skalierbarkeit:** Ermöglicht Isilon-Systemen die Aufrechterhaltung einer konsistenten Performance bei gleichzeitiger Bereitstellung der hochgradig parallelen Workloads, die für Deep-Learning-Implementierungen charakteristisch sind.
- **DevOps- und As-a-Service-Support ohne Konfigurationsaufwand:** Ermöglicht es Unternehmen, Entwicklungs-, Test- und Produktionsdatenumgebungen aufzubauen oder mehrere Produktionsdatenumgebungen mit einer eindeutigen Mandantentrennung über mehrere Zugriffszonen innerhalb desselben Isilon-Clusters bereitzustellen.

Die Software managt die allgemeine Erfahrung und Intelligenz, die in die Dell EMC Isilon-Serie integriert sind. Die Kombination aus einfacher Verwaltbarkeit mit der soliden Performance und den Skalierbarkeitsmerkmalen des Arrays macht Isilon zu einer attraktiven Plattform für Deep Learning Workloads.

DELL EMC ISILON: EINE PLATTFORM, ENTWICKELT FÜR MASCHINELLES LERNEN UND DEEP LEARNING

Das Spitzensystem der Dell EMC Isilon-Speicherproduktreihe ist Isilon F800-All-Flash-Scale-out-NAS. Laut Dell[4]bietet das F800-Array eine Performance und Kapazität, die zu den führenden in der Branche zählen. Das F800-Array bietet bis zu 250.000 IOPS mit einem aggregierten Durchsatz von 15 GB/Sekunde in einem einzigen 4-HE-Gehäuse und bis zu 15,75 Mio. IOPS und 945 GB/Sekunde in einem vollständigen Cluster mit 252 Nodes.

Bezüglich der Kapazität beginnt das Isilon F800-Array mit Dutzenden von Terabyte an Speicher und ermöglicht ein unterbrechungsfreies Scale-out auf Dutzende Petabyte in einem einzigen Namespace. Isilon bietet eine Speichereffizienz von bis zu 85 % sowie eine Deduplizierungs- und Komprimierungstechnologie, die

die Anforderungen an die Datenspeicherkapazität um ein Verhältnis von bis zu 3:1 reduziert und so die effektive Kapazität der Lösung erhöht.

Das Isilon F800-Array kann dafür sorgen, Compute Nodes für Deep Learning stets gut versorgt zu halten. Diese Systeme sind mit 60 leistungsfähigen SSDs und 8 Ethernetverbindungen mit 40 Gbit/s ausgestattet und bieten eine konsistente Performance für die hohen Parallelitätsanforderungen, die für Deep Learning erforderlich sind. Abgesehen von der Bereitstellung einer konsistenten Performance kann das Isilon F800-Array sowohl mit Isilon-Hybrid- als auch mit Isilon-Archiv-Nodes kombiniert werden, um eine einfach zu managende Skalierbarkeit im Petabyte-Bereich bereitzustellen.

[4] Technisches Datenblatt –Dell EMC Isilon F800: https://www.dellemc.com/de-de/collaterals/unauth/data-sheets/products/storage/h15963-ss-isilon-all-flash.pdf

Diese Performance lässt sich nirgendwo besser nachweisen als in den gemeinsam entwickelten Dell EMC Referenzarchitekturen, die die Funktionen von Isilon F800 mit den durch NVIDIA Tesla V100-GPUs beschleunigten Servern wie dem PowerEdge C4140, DSS 8440 und NVIDIA DGX-1 kombinieren. Benchmarks dieser Lösungen zeigten die Performance der
ResNet-50-Benchmark mit bis zu 72 GPUs, die eine lineare Bilder-pro-Sekunde-Performance von 8 bis 72 GPUs mit einer GPU-Auslastung von 97 % erzielen[5].

Diese Benchmarkzahlen zeigen, dass der Prozessor bei einem der leistungsstärksten heute erhältlichen Deep-Learning-Computern der Engpass ist, während Dell EMC Isilon F800 dafür sorgt, dass stets Daten eingespeist werden.

Dell EMC PowerMax: Leistungsfähiger Blockspeicher

Es gibt einige Schritte im KI-Workflow und bestimmte ML- und DL-Algorithmen, die einen Blockspeicher mit sehr geringer Latenz für eine Echtzeit-Reaktionsgeschwindigkeit während der Datenaufnahme, Datenvorbereitung und Ableitung in der Produktion benötigen.

Die Dell EMC PowerMax-Serie mit Blockspeicherlösungen ist als eine der leistungsstärksten derzeit erhältlichen Speicherarchitekturen gut auf die Unterstützung dieser Szenarien ausgelegt. PowerMax basiert auf End-to-End-NVMe und bietet Latenzen von weniger als 300 ms bei 1,7 bis 10 Mio. IOPS (bei PowerMax 2000 bzw. PowerMax 8000) sowie bis zu 13 TB Speicherkapazität pro Baustein[6].

[5] Dell EMC Whitepaper: Dell EMC Isilon and NVIDIA DGX-1 Servers for Deep Learning. November 2018. https://www.dellemc.com/de-de/collaterals/unauth/white-papers/products/storage/Dell_EMC_Isilon_and_NVIDIA_DGX_1_servers_for_deep_learning.pdf

[6] Technisches Datenblatt – Dell EMC PowerMax: https://germany.emc.com/collateral/data-sheet/h16739-powermax-2000-8000-ss.pdf

Dell hat PowerMax positioniert, um die anspruchsvollsten Echtzeit-KI-Workloads zu unterstützen, die heute in Unternehmen bereitgestellt werden.

DELL EMC: DEEP LEARNING IM GESAMTEN STACK

Speicher und Compute sind in Deep-Learning-Umgebungen miteinander verflochten. Eine gut entwickelte Infrastruktur für Deep Learning mit all den damit verbundenen Komplexitäten des Datenmanagements läuft letztendlich auf Ausgewogenheit, Interoperabilität, Performance und Flexibilität hinaus. Trotz der großen Ähnlichkeit bei den Implementierungen gibt es keinen einzigen richtigen Weg. Jede Bereitstellung und jede Umgebung unterscheiden sich leicht.

Es gibt eine Fülle an Optionen für die Bereitstellung von auf maschinellem Lernen und Deep Learning basierenden Workloads. Die verschiedenen Phasen erfordern nicht nur einen unterschiedlichen Datenzugriff, sondern auch unterschiedliche Compute-Lösungen. KI-Fachkräfte können Workloads auf Bare-Metal-Servern, auf virtuellen Maschinen oder sogar in Docker-ähnlichen Containern ausführen.

Neben der einfachen Bereitstellung einzelner Elemente in einer Deep-Learning-Infrastruktur möchte Dell EMC Lösungen ermöglichen, die schnell von IT-Fachkräften bereitgestellt werden können. Dell EMC vereinfacht Architekturentscheidungen und verkürzt Bereitstellungszeiten mit Ready Solutions und Referenzarchitekturen (RA), die Elemente kombinieren, um das jeweils vorliegende Problem zu lösen. Dell EMC bietet Richtlinien für die Konfiguration von Lösungen, die Unternehmen dabei unterstützen, ihre Data-Analytics- und KI-Lösungen in Übereinstimmung mit ihren spezifischen Workload-Anforderungen zu dimensionieren und zu skalieren.

Die Ready Solutions und RAs kombinieren dieoptimal dimensionierten Dell PowerEdge-Server mit Dell EMC Netzwerkschaltern, Isilon-Speicher und einem für die Lösung optimierten Software-Stack. Die Ready Solutions sind validierte und bestellbare Hardware- und Software-Stacks, die für die Beschleunigung von KI-Initiativen optimiert sind und die Zeit für die Entwicklung neuer Lösungen um 6–12 Monate verkürzen. Beratungs-, Support-, Finanzierungs- und Bereitstellungsservices von Dell Technologies sorgen dafür, dass Sie noch mehr von der Leistung und den Vorteilen von Dell EMC Ready Solutions profitieren.
All diese Services arbeiten zusammen, um eine reibungslose Lösungsbereitstellung sicherzustellen.

Referenzarchitekturen sind getestete und validierte Stacks, die auf Dell Kunden und Lösungspartner ausgelegt sind. Während Ready Solutions direkt bei Dell bestellt werden können, sind RAs darauf ausgelegt, IT-Experten dabei zu unterstützen, ihre eigenen Best-of-Breed-Lösungen basierend auf bewährten Produkten von Dell Technologies zu entwickeln.

TABELLE 1: BEISPIELE FÜR EINIGE DER DER VERFÜGBAREN READY SOLUTIONS UND REFERENZARCHITEKTUREN

Typ	Lösung	Wesentliche Elemente	Wichtigste
Ready Solutions für KI	Deep Learning mit Intel	Isilon H600 PowerEdge R740xd PowerEdge C6420	Intel
	Deep Learning mit NVIDIA	Isilon F800 PowerEdge R740xd PowerEdge C4140	NVIDIA
	maschinelles Lernen mit Hadoop	Isilon H500/H600 PowerEdge R640	Hortonworks
Referenzarchitekturen für KI	Dell EMC Isilon und NVIDIA DGX-1 für Deep Learning	Isilon F800 NVIDIA DGX-1	NVIDIA
	Dell EMC Isilon und PowerEdge C4140 für Deep Learning	Isilon F800 PowerEdge C4140	NVIDIA
	Dell EMC Isilon und DSS 8440 für Deep Learning	Isilon F800 DSS 8440	NVIDIA
	Dell EMC Isilon und PowerEdge R940 für algorithmischen Handel	Isilon F800 PowerEdge R940	Intel

Quelle: Moor Insights & Strategy

FAZIT

Daten sind in vielen Unternehmen zur wichtigsten strategischen und differenzierenden Ressource geworden. KI-Techniken revolutionieren die Art und Weise, in der Daten interpretiert und genutzt werden. Unternehmen investieren in hohem Maße in den Aufbau von Wissen und die Bereitstellung einer Infrastruktur zur Unterstützung dieser Realität.

Gleichzeitig verlangt künstliche Intelligenz – ob maschinelles Lernen oder Deep Learning – von IT-Abteilungen eine andere Denkweise rund um Daten und Speicherarchitekturen als sie bei diejenigen vorliegt, die eher herkömmliche Enterprise Workloads unterstützen. Die Merkmale der Daten unterscheiden sich. Die Komplexität der Analysen unterscheidet sich. Die Anforderungen der Nutzer dieser Daten unterscheiden sich. Dabei ist vor allem die Fähigkeit, beschleunigte Compute Nodes mit Daten zu versorgen, von größter Bedeutung. Die Dell EMC Isilon-basierten KI-Lösungen sind auf genau diese Anforderungen ausgelegt.

Die Bereitstellung von Deep-Learning-Lösungen muss gut durchdacht sein. Sie erfordert eine Partnerschaft mit Technologieanbietern, die die Anforderungen dieser neuen Welt verstehen und umfassende gezielte Lösungen bereitstellen, die erforderlich sind, um die Problembereiche von IT-Fachkräften anzugehen, die in dieser Welt leben.

Dell EMC ist ein großartiges Beispiel für einen solchen Partner. Deep Learning setzt Daten an erste Stelle und Dell EMC gehört zu den weltweit führendenUnternehmen, wenn es um das Management von Daten aus Rechenzentren, Private und Public Clouds sowie Edge-Netzwerken geht. Das umfassende KI-Portfolio von Dell EMC bringt das Unternehmen in die einzigartige Position, die Entwicklung der bestmöglichen Umgebung zur Erfüllung von Kundenanforderungen zu unterstützen. Dell EMC verfügt über ein umfassendes Speicherportfolio für das Management und den Schutz von Kundendaten sowie Services und Lösungen, die für erfolgreiche KI-Bereitstellungen optimiert sind.

Weitere Informationen finden Sie auf der dedizierten Website von Dell EMC unter: https://www.dellemc.com/de-de/solutions/artificial-intelligence/index.htm

WICHTIGE INFORMATIONEN ÜBER DIESES WHITEPAPER

MITWIRKENDER
Steve McDowell, Senior Analyst bei Moor Insights & Strategy

HERAUSGEBER
Patrick Moorhead, Gründer, President und Principal Analyst bei Moor Insights & Strategy

ANFRAGEN
Kontaktieren Sie uns, wenn Sie über diesen Bericht sprechen möchten. Moor Insights & Strategy wird sich umgehend bei Ihnen melden.

ZITIERUNG
Dieses Dokument kann durch akkreditierte Pressemitarbeiter und Analysten zitiert werden. Zitate müssen jedoch im Kontext genannt und der Name des Autors, der Titel des Autors und „Moor Insights & Strategy“ erwähnt werden. Andere Personen als Pressemitarbeiter und Analysten müssen vor jeglichen Zitaten die schriftliche Genehmigung von Moor Insights & Strategy einholen.

LIZENZIERUNG

OFFENLEGUNGEN
Dieser Artikel wurde von Dell in Auftrag gegeben. Moor Insights & Strategy stellt vielen der in diesem Whitepaper erwähnten High-Tech-Unternehmen Studien, Analysen, Empfehlungen und Beratung bereit. Kein Mitarbeiter des Unternehmens hat ein Eigenkapitalinteresse an den in diesem Dokument zitierten Unternehmen.

HAFTUNGSAUSSCHLUSS
Die im vorliegenden Dokument präsentierten Informationen dienen nur zu Informationszwecken und enthalten eventuell technische Ungenauigkeiten, Auslassungen und Rechtschreibfehler. Moor Insights & Strategy lehnt jede Gewährleistung hinsichtlich der Richtigkeit, Vollständigkeit oder Angemessenheit dieser Information ab und übernimmt keinerlei Haftung für Fehler, Auslassungen oder Ungenauigkeiten in diesen Informationen. Dieses Dokument stellt die Meinung von Moor Insights & Strategy dar und darf nicht als Tatsachenbehauptung ausgelegt werden. Die hierin geäußerten Meinungen können ohne vorherige Ankündigung geändert werden.

Moor Insights & Strategy stellt Prognosen und zukunftsgerichtete Aussagen als Richtungsweiser und nicht als präzise Vorhersagen zukünftiger Ereignisse bereit. Unsere Prognosen und zukunftsgerichteten Aussagen geben zwar unsere aktuelle Beurteilung zur zukünftigen Entwicklung wieder, unterliegen aber Risiken und Unsicherheiten, die dazu führen können, dass die tatsächlichen Ergebnisse wesentlich abweichen. Sie werden darauf hingewiesen, kein übermäßiges Vertrauen in diese Prognosen und zukunftsgerichteten Aussagen zu setzen, die unsere Meinungen ausschließlich zum Datum der Veröffentlichung dieses Dokuments darstellen. Bitte denken Sie daran, dass wir uns nicht dazu verpflichten, diese Prognosen und zukunftsgerichteten Aussagen zu überarbeiten oder die Ergebnisse einer Überarbeitung angesichts neuer Informationen oder zukünftiger Ereignisse zu veröffentlichen.