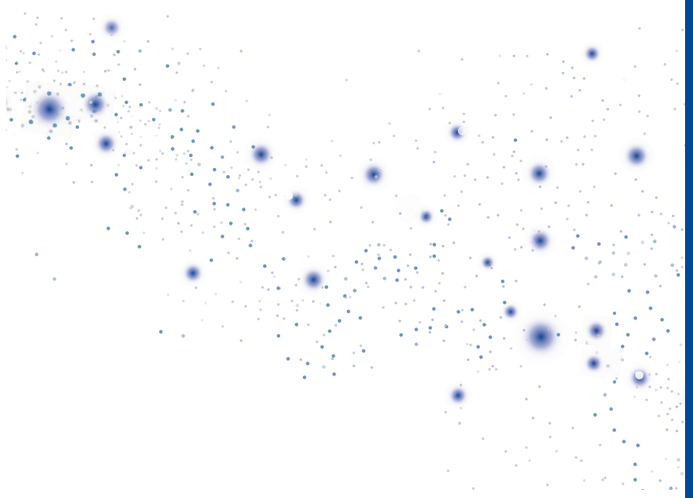
IBM Speicherkonzepte für Künstliche Intelligenz:

Machine Learning und Deep Learning im Fokus

MEGWARE®

Der Stellenwert von Daten in Unternehmen wächst rasant: Die Betrachtung als reiner Kostenfaktor mit Sicherheitsrisiken weicht vielerorts der Erkenntnis, dass Daten einen unternehmerischen Vermögenswert darstellen, der zu erheblichen Wettbewerbsvorteilen führen kann. Dies gilt besonders, wenn es m digitale Geschäftsmodelle im Internet of Things (IoT) geht. Um solche Potenziale auszureizen, ist eine Anpassung der unternehmerischen Storage-Strategie unverzichtbar. Besonders im Umfeld von Workloads, die auf Künstlicher Intelligenz (KI) und maschinellem Lernen (ML) basieren, gewinnen neue leistungsfähige Speichertechnologien wie All-Flash, Hyper Converged Storage oder Object Storage massiv an Bedeutung.



Inhalt

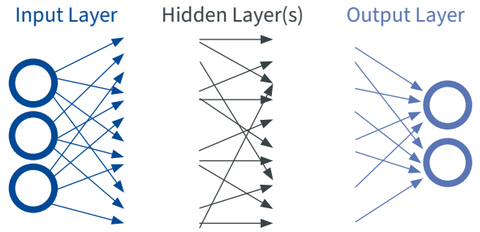
01 Künstliche Intelligenz erobert den Businessalltag

Schon in den vergangenen Jahrzehnten rückten Teilbereiche des maschinellen Lernens in den Mittelpunkt der IT-Entwicklung, scheiterten aber an technischen Hürden unterschiedlichster Art. Erst jetzt ist eine Zonstellation zwischen bezahlbaren GPUs, wirtschaftlichem Datenspeicher, dem Know-how der neuronalen Netze und entsprechenden Fachkräften vorhanden, die KI und Co. in vielen Branchen nutzbar macht.

AI & Deep Learning Komponenten

1. SERVER MIT GPUs

2. NEURONALE NETZE



3. DATEN UND STORAGE

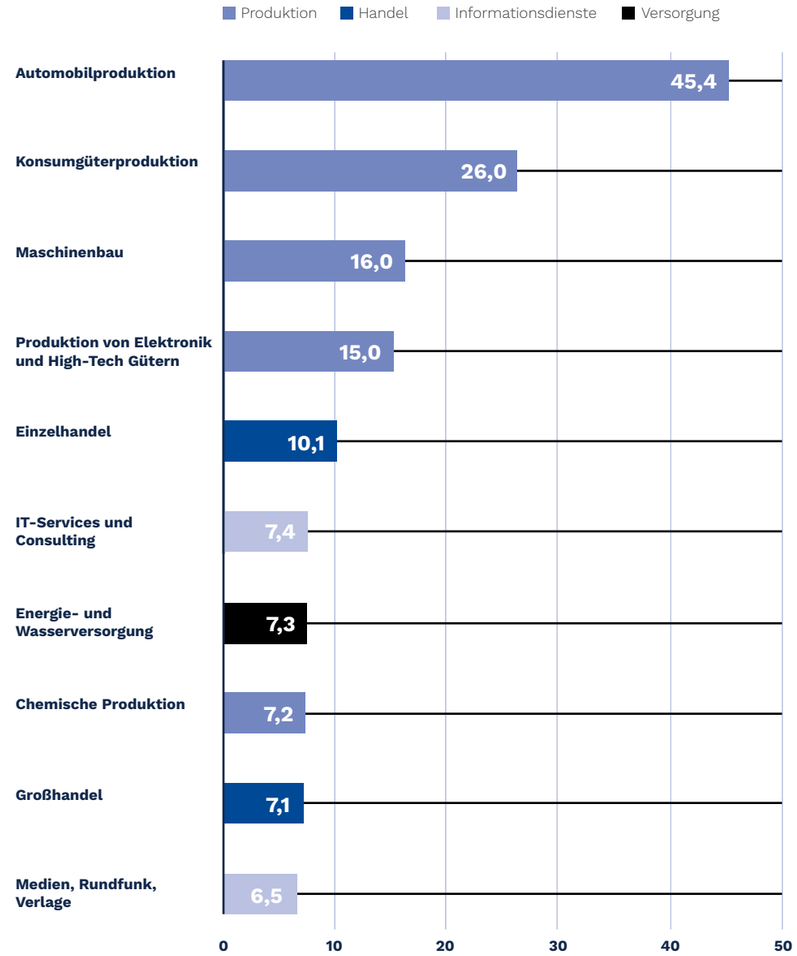
Was bis vor kurzem für Machine- und Deep-Learning-Projekte fehlte, waren neben preiswerten Speicherlösungen die nötigen Rechner-Ressourcen. Mit der hohen Leistungsfähigkeit von GPU-Computing (Geographical Processor Unit) können solche Projekte erfolgreich umgesetzt werden. Für erfolgreiche Use Cases im ML-Bereich spielen aber auch noch andere Faktoren eine maßgebliche Rolle: die CPU, die Kommunikationsbandbreite zwischen CPU und GPU, die Übertragungsgeschwindigkeit von den Rechnern auf die Speicherinfrastruktur oder die Parallelisierung von IOs zwischen Rechner und Storage. Hinzu kommen die Datensammlung und deren Aufbereitung für das Training, die Frameworks zur datenstromorientierten Programmierung, die Tools und Frameworks für die Algorithmen, die Wahl der neuronalen Netze und vieles mehr. Schon die Definition eines Use Cases und die damit verbundene Projektplanung können sehr komplex ausfallen.

02 Aktuelle Marktsituation und Perspektiven

Im Rahmen der derzeitigen Digitalisierungsoffensive führt kaum ein Weg an KI-Projekten und Algorithmen vorbei.

Diese Aussage bestätigen auch die Experten von AI Research: „Only one kind of enterprise will survive in the future: the data driven enterprise“ (nur eine Art von Unternehmen wird in Zukunft überleben: das datengetriebene).

KI- und Deep Learning-Verfahren werden in den nächsten fünf bis zehn Jahren alle Industriezweige erobern. In der Automobilbranche geht es dabei vor allem um intelligente Assistenzsysteme im Fahrzeug, Predictive-Analytics-Verfahren für die effiziente Wartung und die Minimierung von Unfallrisiken. Auch das autonome Fahren ist ohne Künstliche Intelligenz nicht umsetzbar. Weitere vielversprechende Einsatzfelder sind die Spracherkennung, Übersetzungen, Stimmungsanalysen, Bild-Tagging, Chatbots, Videoüberwachung, Gesichtserkennung, Drohnensteuerung, Verkehrsstrom-Optimierung, Qualitätsanalyse, Lagerautomatisierung, autonome Maschinen, Risiko- und Portfolio-Analysen, Betrugserkennung, medizinische Diagnostik und Genomanalysen bis hin zu Human-Brain-Projekten in der Forschung.



*Brutto-Umsatz; Prognose auf Basis der steuerbereinigten Gesamtumsätze aus Lieferungen und Leistungen von Unternehmen mit mehr als 17.500 Euro Jahresumsatz (lt. Statistischem Bundesamt)
Quelle: Appanion Labs © Statista.com

Insgesamt werden 2019 in Deutschland schätzungsweise

220,6 MRD. €

Umsatz durch KI-Anwendungen beeinflusst

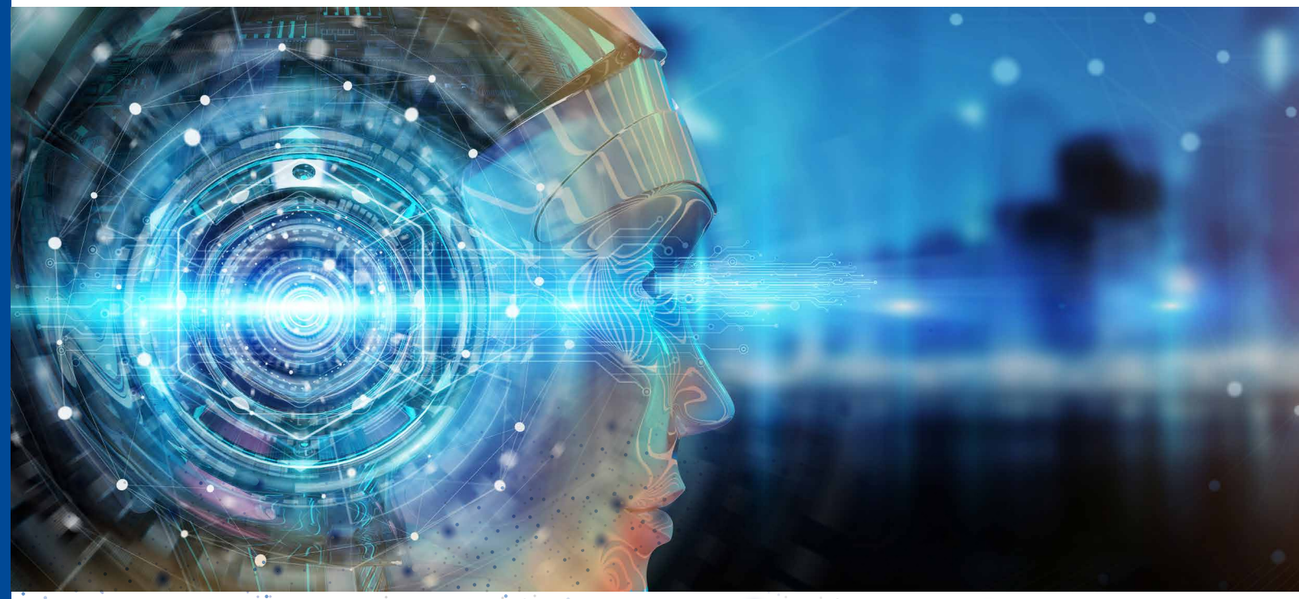
Experten schätzen, dass 2019 in Deutschland Umsätze in Höhe von 220,6 Milliarden Euro von Anwendungen mit Künstlicher Intelligenz beeinflusst werden. An erster Stelle steht mit großem Vorsprung die Automobilproduktion, gefolgt von der Konsumgüterfertigung. Dahinter rangieren der Maschinenbau sowie die Herstellung von Elektronik und High-Tech-Gütern.

(Quelle: Appanion Labs / Statista)

03 Aufbau und Funktion neuronaler Netze

In der Neurowissenschaft sind neuronale Netze eine beliebige Anzahl von Nervenzellen, die als Teil eines Nervensystems bestimmten Funktionen dienen und über Synapsen verknüpft sind.

Bei KI-Projekten kommen künstliche neuronale Netzwerke zum Einsatz. Dabei handelt es sich um eine Ansammlung von Einheiten zur Informationsverarbeitung, die schichtweise in einer Netzarchitektur angeordnet sind. Neben der Eingabe- und Ausgabeschicht liegen die versteckten Zwischenschichten, die sogenannten „Hidden Layers“. Je komplexer die Aufgabe, desto mehr Zwischenschichten werden notwendig.
Und damit auch höhere Rechenleistungen.



Die Aufnahmeschicht nimmt die Eingangssignale entgegen und gibt sie an die Zwischenschichten weiter. Jede Einheit führt eine Gewichtung durch – sowohl positiv als auch negativ – und gibt diese Info an die Einheiten der nächsten Schicht weiter, die für eine neue Gewichtung sorgt. Je höher das Gewicht der sendenden Unit, desto größer ist der Einfluss auf die empfangende Unit. Die Ausgabeschicht liefert schließlich das Ergebnis der Verarbeitung des gesamten neuronalen Netzwerks. Für die vektorbasierte Matrizenrechnung werden GPUs eingesetzt. Es gibt inzwischen eine ganze Menge künstlicher neuronaler Netze, die je nach Aufgabenstellung Verwendung finden können.

04 Anforderungen an die Speicherinfrastruktur

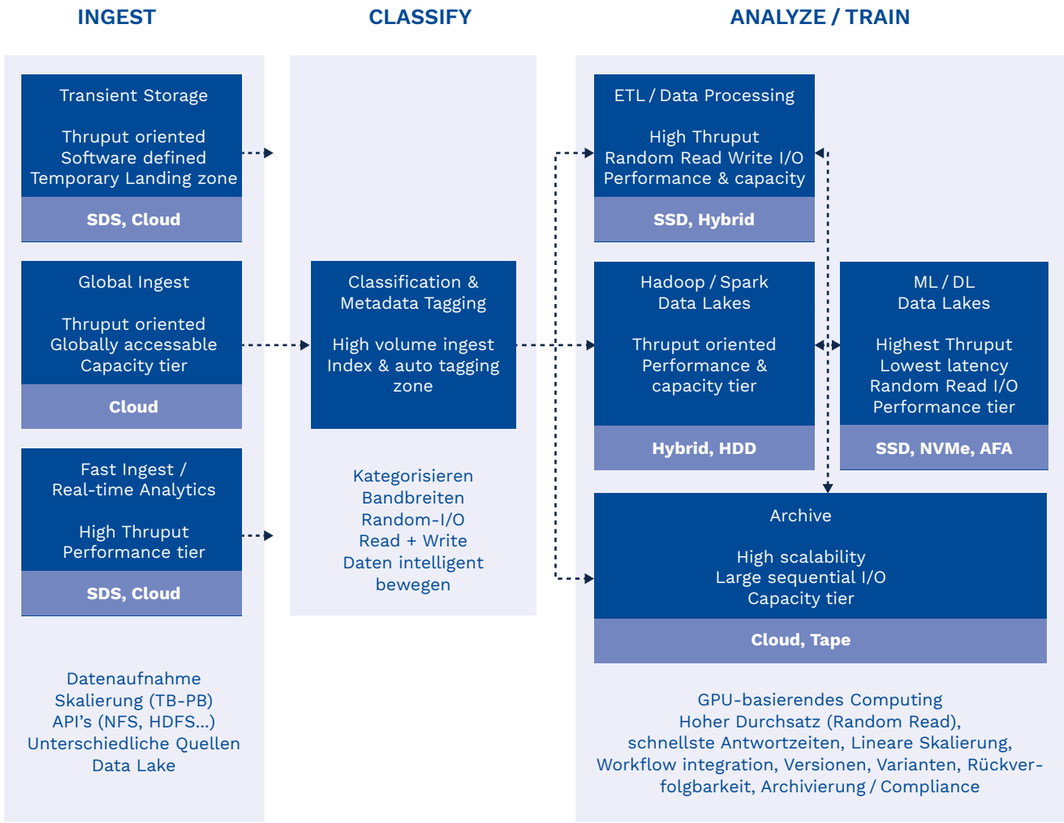
KI-Anwendungen in der Produktion lassen sich in der Regel mit vorhandenen Speicherinfrastrukturen abbilden.

Die größten Herausforderungen entstehen bei der Datensammlung, während der Aufbereitung der Daten und in der Trainingsphase der neuronalen Netze bei Machine- und Deep-Learning-Projekten.

Hier bedarf es Infrastrukturen der unterschiedlichsten Art. So können für das Sammeln der Daten („Ingest“) aus verschiedenen Quellen Cloudspeicher-Lösungen, Objektspeicher, Onlinearchive und Filesysteme eingesetzt werden. Auch die Skalierbarkeit spielt eine bedeutsame Rolle. Zur Kategorisierung und Aufbereitung der Daten („Classify“) sind ausreichende Bandbreiten, Random-IOs für das Lesen oder Schreiben und ein intelligentes Verschieben an die richtigen Speicherplätze erforderlich.

Für die Trainingsphase selbst („Analyze“ / „Train“) sind in den meisten Fällen ein hoher Datendurchsatz („Random Read“), schnellste Antwortzeiten, eine lineare Skalierung, eine Workflow-Integration sowie verschiedene Versionen und Varianten notwendig. Ist der Algorithmus einsatzbereit, muss für Compliance-Zwecke die Rückverfolgbarkeit der Trainingsdaten gewährleistet werden. Für die revisionssichere Archivierung empfiehlt sich das Medium Tape.

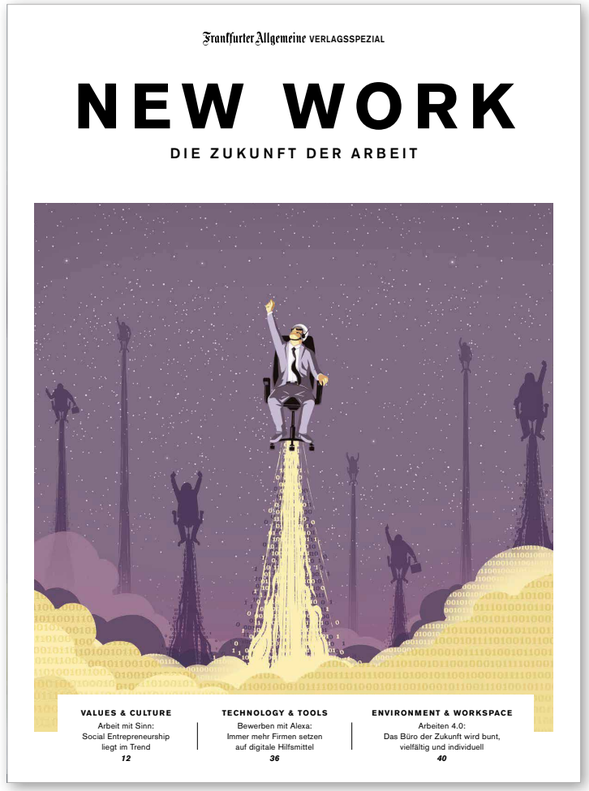
Die Daten-Pipeline vom Sammeln der Daten bis zum Training und danach spiegelt die unterschiedlichsten Anforderungen wider. Es gibt nicht viele Anbieter auf dem Markt, die eine End-to-End-Daten- und Speicherplattform für alle genannten Phasen zur Verfügung stellen können.



05 KI-getriebene Geschäftsfelder und Berufe

Eine Studie der Unternehmensberatung IDG von 2018 zu den Themenfeldern Machine- und Deep Learning zeigt, dass 90 Prozent aller KI-Projekte „inhouse“ im firmeneigenen Rechenzentrum stattfinden.

Dieses Vorgehen lässt sich mit dem Wunsch der Entscheider nach höchstmöglicher Sicherheit erklären, aber auch mit dem Wertschöpfungspotenzial der Algorithmen. Hier haben sich schon viele neue Geschäftsfelder entwickelt, etwa der Verkauf von Bilder- und Datensammlungen, neue Tools für die Aufbereitung von Daten, die Vermarktung von Frameworks mit entsprechenden Bibliotheken oder die Nutzung vorgefertigter neuronaler Netze. Die Vielfalt ist groß.



Parallel dazu sind völlig neue Berufsbilder für die einzelnen KI-Phasen entstanden. Hier gibt es neuerdings den Data Engineer oder Data Expert, den Data Scientist, den Content-Spezialisten, den KI-Plattform-Architekten und neben den Hardware- und Software-Entwicklern auch KI-Berater, KI-Integratoren, KI-Lifecycle-Manager, KI-Program-Manager und KI-Operations-Manager, um nur einige zu nennen. Zu diesem Thema hat die Frankfurter Allgemeine Zeitung im Juli 2019 eine Beilage publiziert („New Work – die Zukunft der Arbeit“), in der die neuen Arbeitsformen im Rahmen der Digitalisierung analysiert werden.

Die fortschrittliche Storage-Lösung für alle KI-Anwendungen: IBM Spectrum Storage

IBM kann Speicherinfrastrukturen für Machine- und Deep-Learning-Projekte als End-to-End-Data Pipeline in vielen flexiblen Varianten anbieten: als Cloud-Lösung (Public, Private und Hybrid), Software-defined, als Hardware im Rechenzentrum oder in allen erdenklichen Kombinationen. Neben den Power AI-Servern IBM AC922, den Servern von NVIDIA (wahlweise) und der Speicherinfrastruktur spielen folgende Produkte eine elementare Rolle:

IBM SPECTRUM Scale

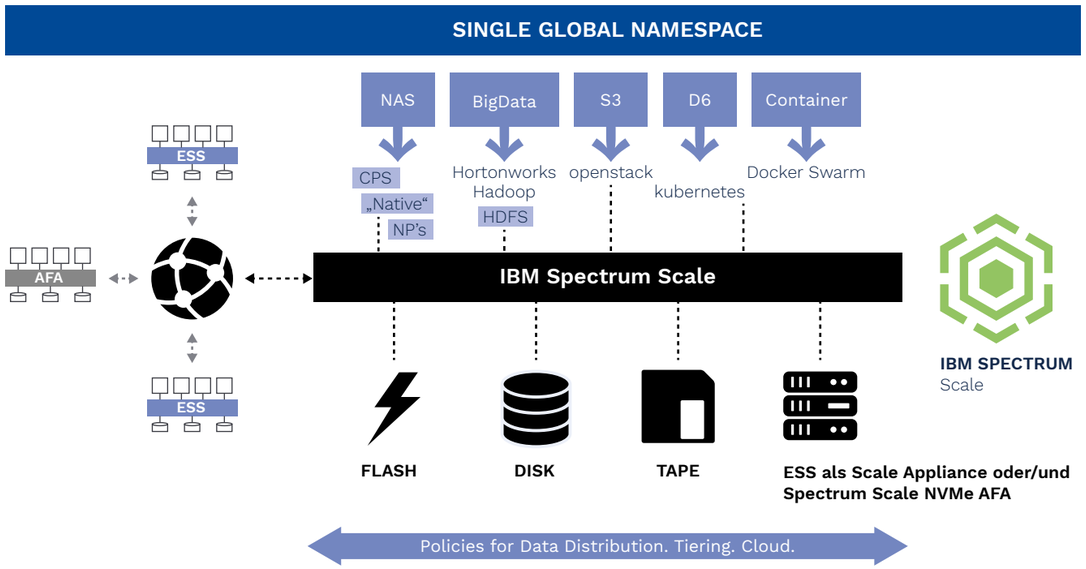
IBM Cloud Object Storage

IBM SPECTRUM Discover

06 IBM Spectrum Scale: Herzstück der KI-Speicherinfrastruktur

IBM Spectrum Scale ist eine skalierbare Hochleistungslösung für das Daten- und Dateimanagement, die auf dem General Parallel File System (GPFS) aufbaut.

Dieses kommt vor allem im HPC-Bereich zur Anwendung und wird weltweit in sämtlichen Industrien eingesetzt. Das extrem schnelle Filesystem unterstützt auch die schnellsten Supercomputer der Welt, Summit und Sierra und kann 2,6 Millionen File-I/O-Operationen pro Sekunde verarbeiten. Die IBM Lösung bietet Skalierung und Tiering über multiple physikalische Speicher und ist mit IBM Cloud Object Storage (COS), Spectrum Discover und Spectrum Archive erweiterbar.



Darüber hinaus verfügt IBM Spectrum Scale über einen Single Name Space und eine Single Data Plane. Dies ermöglicht jeder Quelle, über Standardschnittstellen wie NFS, SMB, Object, POSIX oder HDFS Daten in das Spectrum Scale Repository zu stellen. Hierbei erhalten die Aufbereitungstools direkten Zugriff auf die Daten; Kopien sind nicht notwendig. Auch die Trainings-Tools können über industrieübliche Interfaces angedockt werden. Das Filesystem führt ein hoch performantes und transparentes Tiering auf alle angeschlossenen Speicher- ressourcen durch – inklusive Disk, Flash, Objektspeicher, Cloud und Tape. Die hocheffiziente KI-Plattform stellt sicher, dass die Daten über die Erasure-CodeTechnolgie RAID abgesichert werden, inklusive der zugehörigen Error- Correction-Algorithmen. So garantiert IBM Spectrum Scale höchste Leistung, Verfügbarkeit und Zuverlässigkeit in jedem KI-Anwendungsszenario.

07 IBM Cloud Object Storage: Kapazitäts-Tier für Spectrum Scale

Bei dieser Objektspeicherlösung handelt es sich um ein kapazitätsoptimiertes Onlinearchiv, das nach der Akquisition der Firma Cleversafe, ehemals die erfolgreichste Objektspeicher-Lösung in den USA, in das IBM Storage Portfolio aufgenommen wurde.

IBM COS ist sehr flexibel einsetzbar, da das Tool als Software, in der Cloud, als Appliance oder als Hybrid-Lösung zur Verfügung steht. Einzigartig ist die nahezu grenzenlose Skalierbarkeit in Kapazität und Leistungsfähigkeit ohne Unterbrechungen. Auch die Verfügbarkeit und Sicherheit der Daten, die durch das von Cleversafe entwickelte Erase-Code-Verfahren mit Verschlüsselung gewährleistet wird, stellt ein Alleinstellungsmerkmal dar. Die Verfügbarkeit entspricht einer dreimal RAID-6-synchronen Spiegelung, benötigt jedoch nur einen Bruchteil der Ressourcen. Damit wird IBM COS zum idealen Online-Repository für Spectrum Scale und kann nicht nur im Frontend zum Einsatz kommen, sondern auch im Backend.

08 IBM Spectrum Discover: Top-Tool für die Datenaufbereitung

Dieses von IBM entwickelte Speichersoftware-Tool führt eine extrem schnelle Content-basierte Datenklassifizierung und -suche durch. Das Scannen und Katalogisieren von Milliarden Files oder Objekten wird mit einer extrem geringen Latenz bewältigt (mehr als 30.000 Files/Objekte pro Sekunde). Dabei kommen benutzerdefinierte Metadaten-Tags zur Anwendung. Diese basieren auf Stichwörtern, die im Inhalt der Files gefunden werden. Danach erfolgt eine automatisierte Datenaufnahme und Indexierung von Metadaten aus den unterschiedlichsten Quellen. Ergänzen lassen sich die Daten durch kundenspezifische Metadaten-Tags.

Die Version 2.0.1 von IBM Spectrum Discover ist über eine einfache Installationsroutine einsetzbar und bietet API-Support für Apache Tika sowie verschiedene Reportings. Neu ist auch die Unterstützung heterogener Speicher wie Amazon S3, Isilon, NetApp und Ceph. Visual Recognition SW oder NLP Engines könnnen via REST API integriert werden. Dazu kommt eine komfortablere Compliance-Einhaltung für schützenswerte Daten (beispielsweise GDPR) mit automatischer Erkennung und Labelung, etwa von Telefonnummern oder Kreditkarten.

IBM Spectrum Discover kann die Datenaufbereitung extrem verkürzen. Benötigt werden nicht mehr Wochen oder Monate, sondern häufig nur noch wenige Tage; je nachdem, ob die Data Scientists und Domain Experts die Vorschläge der intelligenten Software akzeptieren. Auf jeden Fall wird der Ablauf von Machine- und Deep-Learning Projekten stark beschleunigt.

09 Sicherstellung der Rückverfolgbarkeit

Ist der Learning-Prozess eines Projektes soweit abgeschlossen, dass der Algorithmus seine Zielsetzung erreicht und zum Einsatz kommen kann, gilt es, die verwendeten Daten für die Sicherstellung der Rückverfolgung zu archivieren. In dieser nachgelagerten Phase können die Trainingsdaten durchaus noch einige Zeit im Backend des IBM COS verbleiben. Sie könnten ja noch kurzfristig für Korrekturen oder Verbesserungen des Algorithmus benötigt werden. Auf lange Sicht empfiehlt sich allerdings eine Aufbewahrung auf dem Medium Tape, da es wesentlich günstiger als andere Medien zur Datensicherung ist und Schutz vor Cyberattacken sowie Ransomware-Angriffen bietet. IBM Spectrum Scale transferiert die Daten nach vordefinierten Regeln automatisch auf das Tape.

10 Training entscheidet: die Leistungsfähigkeit neuronaler Netze

Besonders in der Trainingsphase wird die Leistungsfähigkeit neuronaler Netze auf die Probe gestellt. Schließlich gilt es, die GPUs so zu bedienen, dass sie möglichst effektiv an ihrer Leistungsgrenze arbeiten, um die Trainingsphase kurz zu halten. Dabei ist Folgendes zu beachten: Je komplexer die Aufgabenstellung ist, desto mehr Hidden Layers werden für das neuronale Netz benötigt. Und je mehr Hidden Layers Verwendung finden, desto höher sind die Anforderungen an die Leistungsfähigkeit. Eine elementare Rolle spielt auch die Größe der Datensammlung. Je mehr qualifizierte Daten zur Verfügung stehen, desto besser sind die Ergebnisse, die der trainierte Algorithmus liefert. Die Learning-Phase ist ein iterativer Rechenprozess, der sich ständig wiederholt, um sich schrittweise der exakten Lösung anzunähern.

11 IBM und mehr: Die Hardware hinter den Storage-Konzepten

Die Server sollten mit der Storage-Infrastruktur über ein leistungsfähiges Netzwerk verbunden werden. Im Rahmen der IBM Spectrum Storage for AI Reference Architecture empfehlen wir den Mellanox SB7800 Infiniband Fabric. Auf der Storage-Seite bieten sich die IBM ESS (Elastic Storage Server) mit Flash oder die neuen NVMe basierenden ESS 3000 AFAs (All Flash Appliances) mit extrem schnellen Flash-Core-Modulen an. Als Server können NVDIA DGX- oder IBM Power AI Systeme (AC922) eingesetzt werden.

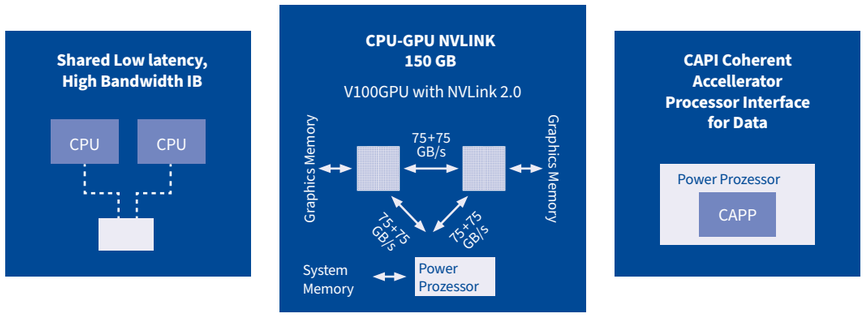
NVIDIA DGX oder IBM Power AI AC922 mit IBM Spectrum Scale

ESS Elastic Storage Server mit Flash und Spectrum Scale oder NVMe All Flash Appliance mit Spectrum Scale

Die Hardware-Referenz: IBM Power AI (AC922)

Wird eine besonders hohe Leistungsfähigkeit verlangt, sollte der Focus auf der IBM AC922 liegen. Dieser Server bietet im KI-Umfeld viele Vorteile:

- Die IBM AC922 nutzt kein NAS-Protokoll wie NFS zur Speicheranbindung, sondern Spectrum Scale, das als Filesystem in der AC922 installiert wird und mit Infiniband einen dreifach höheren Datendurchsatz erlaubt.
- Die IBM AC922 hat einen NVLink zwischen CPU und GPUs mit einer Bandbreite von 150 GB/s (bidirektional 300 GB/s), also 4,5-fach höher als PCIe-Verbindungen mit 32 GB/s. Damit kann die CPU mit den GPUs in derselben hohen Bandbreite kommunizieren wie die GPUs untereinander.
- Das CAPI (Coherent Accelerator Processor Interface) benötigt wesentlich weniger CPU-Zyklen für einen Storage I/O (statt 16.000 bis 20.000 Zyklen nur 300), weil der Device Driver Overhead nahezu wegfällt. Dies macht sich auch im Antwortzeitverhalten bemerkbar: Statt 13 µs werden lediglich 0,36 µs benötigt. Dank einer geringeren Zahl von Prozessorzyklen können I/Os in großer Zahl parallel über RDMA (Remote Direct Memory Access) umgesetzt werden. RDMA ermöglicht das Abarbeiten mehrerer Anwendungen durch erheblich geringer belastete Prozessoren bei wesentlich besserer Ausnutzung der vorhandenen Bandbreite.



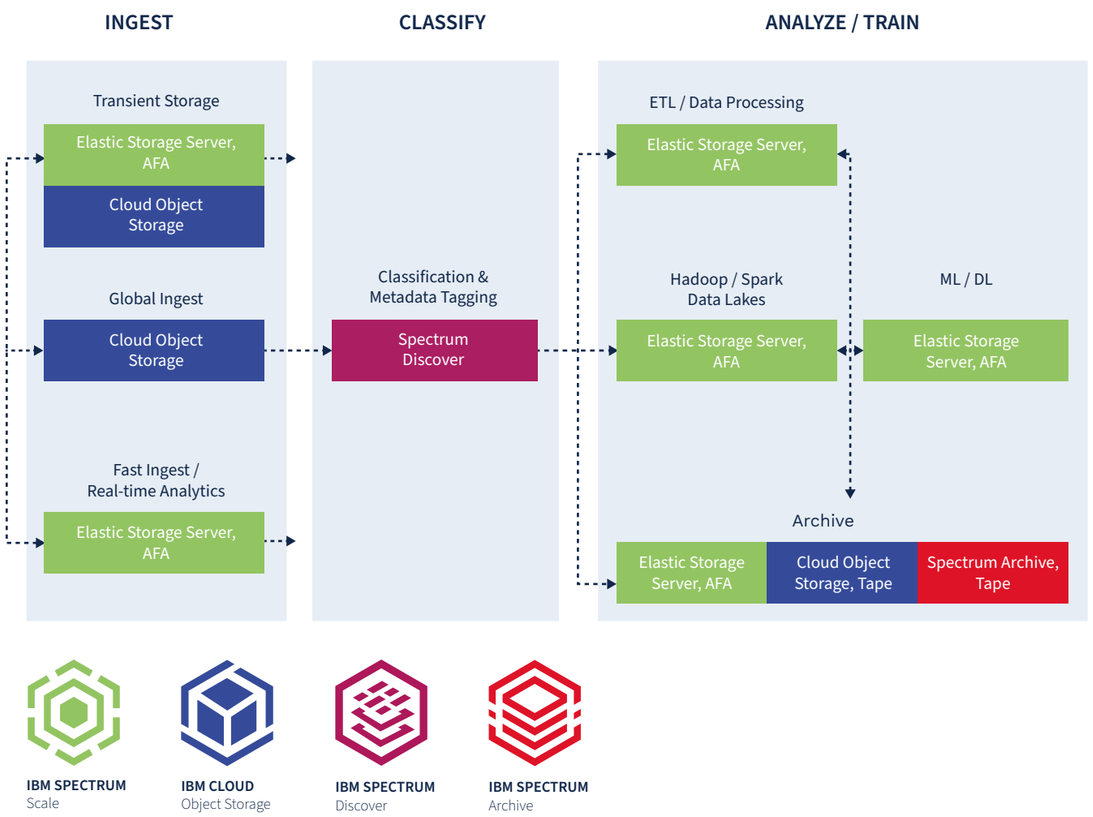
Zur Leistungsfähigkeit trägt selbstverständlich auch die passende Storage-Infrastruktur bei. Das neue IBM Spectrum Scale NVMe-basierte Elastic Storage System 3000 als All Flash Appliance, das seit November 2019 zur Verfügung steht, kommt mit einer durchschnittlichen Antwortzeit von etwa 100 µs aus. Möglich wird diese extreme Geschwindigkeit durch die integrierte IBM FlashCore-Technologie. Das Gerät mit nur 2U Höhe basiert auf dem IBM FlashSystem 9100, bietet acht Infiniband-Links mit 100 GB/s und liefert eine Bandbreite von 40 GB/s per ESS 3000 AFA. Das ESS 3000 skaliert linear und bietet High Performance Parallel Processing auf multiplen GPUs mit hoher Bandbreite bei extrem schnellen Antwortzeiten.

IBM Spectrum Scale ESS 3000 NVMe All Flash Appliance

IBM NVMe Flash Core Modules (FCM's)

IBM Spectrum Scale

- IBM Spectrum Scale ist das Herzstück der IBM Spectrum AI End-to-End-Lösung. Es verbindet alle Phasen von Machine- und Deep-Learning- Projekten und liefert eine hochperformante Datenmanagement-Plattform für sämtliche KI-Prozesse.



Unser Fazit

→ Die IBM Lösungen Spectrum Scale, IBM COS, Spectrum Discover und die konkurrenzlos schnelle Hardware der IBM Systeme AC922 und ESS 3000 All Flash Appliance erfüllen optimal die Anforderungen von Machine- oder Deep-Learning-Projekten – und das in jeder Projektphase. Mit seinen intelligenten Storage-Tools bietet IBM KI-interessierten Unternehmen ein Maximum an Performance, Kosteneffizienz und Zeitersparnis.

IHR KONTAKT FÜR IBM STORAGE LÖSUNGEN BEI MEGWARE

Tobias Pfennig
Sales Director HPC
E-Mail: tobias.pfennig@megware.com

MEGWARE Computer
Vertrieb und Service GmbH
Nordstraße 19
09247 Chemnitz-Röhrsdorf