Generative KI und NetApp -Wert

NetApp artificial intelligence solutions

NetApp
December 04, 2025

Inhalt

Generative KI und NetApp -Wert

Die Nachfrage nach generativer künstlicher Intelligenz (KI) führt zu Umbrüchen in allen Branchen und steigert die Kreativität der Unternehmen sowie die Produktinnovation.

Abstrakt

Viele Organisationen nutzen generative KI, um neue Produktfunktionen zu entwickeln, die Produktivität der Ingenieure zu steigern und Prototypen KI-gestützter Anwendungen zu erstellen, die bessere Ergebnisse und ein besseres Kundenerlebnis bieten. Generative KI wie Generative Pre-trained Transformers (GPT) verwenden neuronale Netzwerke, um neue Inhalte zu erstellen, die so unterschiedlich sind wie Text, Audio und Video. Angesichts des extremen Umfangs und der enormen Datensätze, die mit großen Sprachmodellen (LLMs) verbunden sind, ist es von entscheidender Bedeutung, eine robuste KI-Infrastruktur zu entwickeln, die die überzeugenden Datenspeicherfunktionen von On-Premises-, Hybrid- und Multi-Cloud-Bereitstellungsoptionen nutzt und die mit Datenmobilität, Datenschutz und Governance verbundenen Risiken reduziert, bevor Unternehmen KI-Lösungen entwickeln können. In diesem Dokument werden diese Überlegungen und die entsprechenden NetApp KI-Funktionen beschrieben, die eine nahtlose Datenverwaltung und Datenbewegung über die KI-Datenpipeline zum Trainieren, Umtrainieren, Feinabstimmen und Inferenzieren generativer KI-Modelle ermöglichen.

Zusammenfassung

Zuletzt haben nach der Einführung von ChatGPT, einem Spin-off von GPT-3 im November 2022, neue KI-Tools, die zur Generierung von Text, Code, Bildern oder sogar therapeutischen Proteinen als Reaktion auf Benutzeraufforderungen verwendet werden, erhebliche Bekanntheit erlangt. Dies bedeutet, dass Benutzer eine Anfrage in natürlicher Sprache stellen können und die KI Text interpretiert und generiert, beispielsweise Nachrichtenartikel oder Produktbeschreibungen, die die Benutzeranfrage widerspiegeln, oder Code, Musik, Sprache, visuelle Effekte und 3D-Assets mithilfe von Algorithmen erstellt, die auf bereits vorhandenen Daten trainiert wurden. Infolgedessen tauchen bei der Entwicklung von KI-Systemen immer häufiger Begriffe wie „Stabile Diffusion“, „Halluzinationen“, „Prompt Engineering“ und „Wertausrichtung“ auf. Diese selbstüberwachten oder halbüberwachten Modelle des maschinellen Lernens (ML) werden als vortrainierte Basismodelle (FM) über Cloud-Service-Provider und andere Anbieter von KI-Unternehmen immer häufiger verfügbar und von verschiedenen Unternehmen branchenübergreifend für eine breite Palette nachgelagerter NLP-Aufgaben (Natural Language Processing) übernommen. Wie Marktforschungsunternehmen wie McKinsey behaupten: „Die Auswirkungen der generativen KI auf die Produktivität könnten der Weltwirtschaft einen Mehrwert von Billionen Dollar verleihen.“ Während Unternehmen KI als Denkpartner des Menschen neu konzipieren und FMs gleichzeitig ihre Möglichkeiten für Unternehmen und Institutionen mit generativer KI erweitern, werden die Möglichkeiten zur Verwaltung riesiger Datenmengen weiter zunehmen. Dieses Dokument enthält einführende Informationen zur generativen KI und den Designkonzepten in Bezug auf NetApp -Funktionen, die NetApp Kunden sowohl vor Ort als auch in Hybrid- oder Multicloud-Umgebungen einen Mehrwert bieten.

Welchen Vorteil bietet es den Kunden, NetApp in ihren KI-Umgebungen einzusetzen?

NetApp unterstützt Unternehmen dabei, die Komplexität zu bewältigen, die durch schnelles Daten- und Cloud-Wachstum, Multi-Cloud-Management und die Einführung von Technologien der nächsten Generation wie KI entsteht. NetApp hat verschiedene Funktionen in intelligenter Datenverwaltungssoftware und Speicherinfrastruktur kombiniert, die gut ausbalanciert und mit hoher Leistung für KI-Workloads optimiert sind. Generative KI-Lösungen wie LLMs müssen ihre Quelldatensätze mehrmals aus dem Speicher in den Arbeitsspeicher lesen und verarbeiten, um die Intelligenz zu fördern.

NetApp ist ein führender Anbieter von Technologien für Datenmobilität, Datenverwaltung und Datensicherheit im gesamten Edge-to-Core-to-Cloud-Ökosystem und unterstützt Unternehmenskunden beim Aufbau maßstabsgetreuer KI-Lösungen. NetApp unterstützt mit einem starken Partnernetzwerk Chief Data Officers, KI-Ingenieure, Unternehmensarchitekten und Datenwissenschaftler bei der Entwicklung einer frei fließenden Datenpipeline für die Datenaufbereitung, den Datenschutz und die strategischen Datenverwaltungsaufgaben beim Training und der Inferenz von KI-Modellen und optimiert so die Leistung und Skalierbarkeit des KI/ML-Lebenszyklus. NetApp Datentechnologien und -Funktionen wie NetApp ONTAP AI für Deep-Learning-Datenpipelines, NetApp SnapMirror für den nahtlosen und effizienten Datentransport zwischen Speicherendpunkten und NetApp FlexCache für Echtzeit-Rendering, wenn der Datenfluss von Batch auf Echtzeit umgestellt wird und die Datenentwicklung zeitnah erfolgt, sind für die Bereitstellung von Echtzeit-Modellen der generativen KI von Nutzen. Da Unternehmen aller Art neue KI-Tools einsetzen, stehen sie vor Datenherausforderungen vom Rand über das Rechenzentrum bis hin zur Cloud, die skalierbare, verantwortungsvolle und erklärbare KI-Lösungen erfordern.

Als Datenexperte für Hybrid- und Multi-Cloud engagiert sich NetApp für den Aufbau eines Netzwerks aus Partnern und gemeinsamen Lösungen, die bei allen Aspekten der Erstellung einer Datenpipeline und von Datenseen für das Training (Vortraining) generativer KI-Modelle, die Feinabstimmung, kontextbasierte Inferenz und die Überwachung des Modellverfalls von LLMs helfen können.

Was ist generative KI?

Generative KI verändert die Art und Weise, wie wir Inhalte erstellen, neue Designkonzepte entwickeln und neuartige Kompositionen erkunden. Es veranschaulicht neuronale Netzwerk-Frameworks wie Generative Adversarial Network (GAN), Variational Autoencoders (VAE) und Generative Pre-Trained Transformers (GPT), die neue Inhalte wie Text, Code, Bilder, Audio, Video und synthetische Daten generieren können. Transformer-basierte Modelle wie Chat-GPT von OpenAI, Bard von Google, BLOOM von Hugging Face und LLaMA von Meta haben sich als grundlegende Technologie herausgestellt, die vielen Fortschritten bei großen Sprachmodellen zugrunde liegt. Ebenso sind Dall-E von OpenAI, CM3leon von Meta und Imagen von Google Beispiele für Text-zu-Bild-Diffusionsmodelle, die den Kunden einen beispiellosen Grad an Fotorealismus bieten, um neue, komplexe Bilder von Grund auf neu zu erstellen oder vorhandene Bilder zu bearbeiten, um mithilfe von Datensatzerweiterung und Text-zu-Bild-Synthese, die textuelle und visuelle Semantik verknüpft, qualitativ hochwertige kontextsensitive Bilder zu erzeugen. Digitale Künstler beginnen, eine Kombination aus Rendering-Technologien wie NeRF (Neural Radiance Field) mit generativer KI anzuwenden, um statische 2D-Bilder in immersive 3D-Szenen umzuwandeln. Im Allgemeinen werden LLMs grob durch vier Parameter charakterisiert: (1) Größe des Modells (normalerweise in Milliarden von Parametern); (2) Größe des Trainingsdatensatzes; (3) Trainingskosten und (4) Modellleistung nach dem Training. LLMs lassen sich auch hauptsächlich in drei Transformatorarchitekturen unterteilen. (i) Nur-Encoder-Modelle. Z. B. BERT (Google, 2018); (ii) Encoder-Decoder, z. B. BART (Meta, 2020) und (iii) Nur-Decoder-Modelle. Z. B. LLaMA (Meta, 2023), PaLM-E (Google, 2023). Abhängig von den Geschäftsanforderungen und unabhängig davon, welche Architektur ein Unternehmen wählt, bestimmen im Allgemeinen die Anzahl der Modellparameter (N) und die Anzahl der Token (D) im Trainingsdatensatz die Grundkosten für das Training (Vortraining) oder die Feinabstimmung eines LLM.

Unternehmensanwendungsfälle und nachgelagerte NLP-Aufgaben

Unternehmen aller Branchen entdecken immer mehr Potenzial für KI, um aus vorhandenen Daten neue Wertschöpfungsformen für Geschäftsabläufe, Vertrieb, Marketing und Rechtsdienstleistungen zu extrahieren und zu produzieren. Laut Marktinformationen von IDC (International Data Corporation) zu globalen Anwendungsfällen und Investitionen in generative KI wird das Wissensmanagement in der Softwareentwicklung und im Produktdesign am stärksten betroffen sein, gefolgt von der Storyline-Erstellung für das Marketing und der Codegenerierung für Entwickler. Im Gesundheitswesen betreten klinische Forschungsorganisationen Neuland in der Medizin. Vortrainierte Modelle wie ProteinBERT enthalten Gene Ontology (GO)-Anmerkungen, um schnell Proteinstrukturen für Medikamente zu entwerfen, was einen bedeutenden Meilenstein in der Arzneimittelforschung, Bioinformatik und Molekularbiologie darstellt. Biotech-

Unternehmen haben Versuche am Menschen mit einem mithilfe generativer KI entdeckten Medikament eingeleitet, das auf die Behandlung von Krankheiten wie der Lungenfibrose (IPF) abzielt, einer Lungenerkrankung, die zu einer irreversiblen Vernarbung des Lungengewebes führt.

Abbildung 1: Anwendungsfälle für generative KI

[Abbildung 1: Anwendungsfälle für generative KI]

Die zunehmende Automatisierung durch generative KI verändert auch Angebot und Nachfrage bei Arbeitstätigkeiten in vielen Berufen. Laut McKinsey hat der US-Arbeitsmarkt (Diagramm unten) einen rasanten Wandel durchlaufen, der sich nur fortsetzen könnte, wenn man die Auswirkungen der KI berücksichtigt.

Quelle: McKinsey & Company

[Abbildung 2: Quelle: McKinsey Company]

Rolle des Speichers in der generativen KI

LLMs basieren weitgehend auf Deep Learning, GPUs und Computern. Wenn der GPU-Puffer jedoch voll ist, müssen die Daten schnell in den Speicher geschrieben werden. Während einige KI-Modelle klein genug sind, um im Speicher ausgeführt zu werden, erfordern LLMs hohe IOPS und einen Speicher mit hohem Durchsatz, um einen schnellen Zugriff auf große Datensätze zu ermöglichen, insbesondere wenn es sich um Milliarden von Token oder Millionen von Bildern handelt. Bei einem typischen GPU-Speicherbedarf eines LLM könnte der zum Trainieren eines Modells mit 1 Milliarde Parametern benötigte Speicher bis zu 80 GB bei voller 32-Bit-Präzision betragen. In diesem Fall benötigt Metas LLaMA 2, eine Familie von LLMs mit einer Größenordnung von 7 bis 70 Milliarden Parametern, möglicherweise 70 x 80, also ca. 5600 GB oder 5,6 TB GPU-RAM. Darüber hinaus ist die benötigte Speichermenge direkt proportional zur maximalen Anzahl der Token, die Sie generieren möchten. Wenn Sie beispielsweise Ausgaben von bis zu 512 Token (ca. 380 Wörter) generieren möchten, benötigen Sie"512 MB" . Dies mag unbedeutend erscheinen, aber wenn Sie größere Chargen verarbeiten möchten, summiert es sich. Daher ist es für Unternehmen sehr kostspielig, LLMs im Speicher zu trainieren oder zu optimieren, sodass die Speicherung zu einem Eckpfeiler der generativen KI wird.

Drei Hauptansätze für LLMs

Basierend auf den aktuellen Trends lässt sich der Ansatz zur Bereitstellung von LLMs für die meisten Unternehmen auf drei grundlegende Szenarien reduzieren. Wie in einer kürzlich erschienenen"Harvard Business Review" Artikel: (1) Training (Vortraining) eines LLM von Grund auf – kostspielig und erfordert Expertenkenntnisse im Bereich KI/ML; (2) Feinabstimmung eines Basismodells mit Unternehmensdaten – komplex, aber machbar; (3) Verwendung von Retrieval-Augmented Generation (RAG) zum Abfragen von Dokument-Repositories, APIs und Vektordatenbanken, die Unternehmensdaten enthalten. Bei jeder dieser Methoden gibt es Kompromisse zwischen Aufwand, Iterationsgeschwindigkeit, Kosteneffizienz und Modellgenauigkeit in ihren Implementierungen, die zur Lösung unterschiedlicher Arten von Problemen verwendet werden (Diagramm unten).

Abbildung 3: Problemtypen

[Abbildung 3: Problemtypen]

Stiftungsmodelle

Ein Foundation-Modell (FM), auch Basismodell genannt, ist ein großes KI-Modell (LLM), das anhand großer Mengen nicht gekennzeichneter Daten trainiert wird, Selbstüberwachung im großen Maßstab nutzt und im Allgemeinen für eine breite Palette nachgelagerter NLP-Aufgaben angepasst ist. Da die Trainingsdaten nicht von Menschen beschriftet werden, entsteht das Modell, anstatt explizit kodiert zu werden. Dies bedeutet, dass

das Modell eigene Geschichten oder Erzählungen generieren kann, ohne explizit dazu programmiert zu sein. Daher ist die Homogenisierung ein wichtiges Merkmal von FM, was bedeutet, dass in vielen Bereichen dieselbe Methode verwendet wird. Dank Personalisierungs- und Feinabstimmungstechniken eignen sich FMs, die heutzutage in neue Produkte integriert werden, jedoch nicht nur gut zum Generieren von Text, zur Umwandlung von Text in Bilder und von Text in Code, sondern auch zum Erklären domänenspezifischer Aufgaben oder zum Debuggen von Code. Beispielsweise können FMs wie Codex von OpenAI oder Code Llama von Meta Code in mehreren Programmiersprachen basierend auf natürlichsprachlichen Beschreibungen einer Programmieraufgabe generieren. Diese Modelle beherrschen über ein Dutzend Programmiersprachen, darunter Python, C#, JavaScript, Perl, Ruby und SQL. Sie verstehen die Absicht des Benutzers und generieren spezifischen Code, der die gewünschte Aufgabe erfüllt und für die Softwareentwicklung, Codeoptimierung und Automatisierung von Programmieraufgaben nützlich ist.

Feinabstimmung, Domänenspezifität und Umschulung

Eine der gängigen Vorgehensweisen bei der LLM-Bereitstellung nach der Datenvorbereitung und -vorverarbeitung besteht darin, ein vortrainiertes Modell auszuwählen, das anhand eines großen und vielfältigen Datensatzes trainiert wurde. Im Kontext der Feinabstimmung kann dies ein Open-Source-Sprachmodell für große Systeme sein, wie zum Beispiel"Metas Lama 2" trainiert mit 70 Milliarden Parametern und 2 Billionen Token. Sobald das vortrainierte Modell ausgewählt ist, besteht der nächste Schritt darin, es anhand der domänenspezifischen Daten zu optimieren. Dabei werden die Parameter des Modells angepasst und es anhand der neuen Daten trainiert, um es an eine bestimmte Domäne und Aufgabe anzupassen. Beispielsweise BloombergGPT, ein proprietärer LLM, der auf eine breite Palette von Finanzdaten für die Finanzbranche spezialisiert ist.

Domänenspezifische Modelle, die für eine bestimmte Aufgabe entwickelt und trainiert wurden, weisen in ihrem Anwendungsbereich im Allgemeinen eine höhere Genauigkeit und Leistung auf, sind jedoch nur schwer auf andere Aufgaben oder Domänen übertragbar. Wenn sich das Geschäftsumfeld und die Daten über einen bestimmten Zeitraum ändern, kann die Vorhersagegenauigkeit des FM im Vergleich zur Leistung während des Tests nachlassen. In diesem Fall ist eine erneute Schulung oder Feinabstimmung des Modells von entscheidender Bedeutung.

Unter Modellneutraining in herkömmlicher KI/ML versteht man die Aktualisierung eines bereitgestellten ML-Modells mit neuen Daten. Dies wird im Allgemeinen durchgeführt, um zwei Arten von auftretenden Abweichungen zu beseitigen. (1) Konzeptdrift – wenn sich die Verbindung zwischen den Eingabevariablen und den Zielvariablen im Laufe der Zeit ändert, kann das Modell ungenaue Vorhersagen liefern, da sich die Beschreibung dessen ändert, was wir vorhersagen möchten. (2) Datendrift – tritt auf, wenn sich die Eigenschaften der Eingabedaten ändern, beispielsweise Änderungen der Gewohnheiten oder des Verhaltens der Kunden im Laufe der Zeit, und das Modell daher nicht in der Lage ist, auf solche Änderungen zu reagieren.

In ähnlicher Weise gilt die Umschulung für FMs/LLMs, sie kann jedoch wesentlich teurer sein (in Millionenhöhe) und ist daher für die meisten Organisationen nichts, was sie in Betracht ziehen würden. Es wird derzeit aktiv erforscht und befindet sich im Bereich LLMOps noch in der Entwicklung. Wenn es bei fein abgestimmten FMs zu einem Modellverfall kommt, können sich Unternehmen statt für ein erneutes Training für eine erneute Feinabstimmung (die wesentlich günstiger ist) mit einem neueren Datensatz entscheiden. Zur Kostenperspektive ist unten ein Beispiel einer Modellpreistabelle von Azure-OpenAI Services aufgeführt. Für jede Aufgabenkategorie können Kunden Modelle anhand bestimmter Datensätze optimieren und auswerten.

Quelle: Microsoft Azure

[Quelle: Microsoft Azure]

Prompte Entwicklung und Inferencing

Prompt Engineering bezieht sich auf die effektiven Methoden zur Kommunikation mit LLMs, um gewünschte

Aufgaben auszuführen, ohne die Modellgewichte zu aktualisieren. So wichtig das Training und die Feinabstimmung von KI-Modellen für NLP-Anwendungen ist, so wichtig ist auch das Inferenzieren, bei dem die trainierten Modelle auf Benutzereingaben reagieren. Die Systemanforderungen für die Inferenz hängen im Allgemeinen viel stärker von der Leseleistung des KI-Speichersystems ab, das Daten von LLMs an die GPUs überträgt, da es in der Lage sein muss, Milliarden gespeicherter Modellparameter anzuwenden, um die beste Antwort zu erzielen.

LLMOps, Modellüberwachung und Vectorstores

Wie herkömmliche Machine Learning Ops (MLOps) erfordern auch Large Language Model Operations (LLMOps) die Zusammenarbeit von Datenwissenschaftlern und DevOps-Ingenieuren mit Tools und Best Practices für die Verwaltung von LLMs in Produktionsumgebungen. Allerdings können der Arbeitsablauf und der Technologie-Stack für LLMs in gewisser Weise variieren. Beispielsweise reihen LLM-Pipelines, die mit Frameworks wie LangChain erstellt wurden, mehrere LLM-API-Aufrufe an externe Einbettungsendpunkte wie Vektorspeicher oder Vektordatenbanken aneinander. Die Verwendung eines Einbettungsendpunkts und eines Vektorspeichers für nachgelagerte Konnektoren (z. B. zu einer Vektordatenbank) stellt eine bedeutende Weiterentwicklung in der Art und Weise dar, wie Daten gespeichert und abgerufen werden. Im Gegensatz zu herkömmlichen ML-Modellen, die von Grund auf neu entwickelt werden, basieren LLMs häufig auf Transferlernen, da diese Modelle mit FMs beginnen, die mit neuen Daten feinabgestimmt werden, um die Leistung in einem spezifischeren Bereich zu verbessern. Daher ist es von entscheidender Bedeutung, dass LLMOps die Funktionen des Risikomanagements und der Überwachung des Modellverfalls bereitstellt.

Risiken und Ethik im Zeitalter der Generativen KI

„ChatGPT – Es ist schick, aber es verbreitet trotzdem Unsinn.“ – MIT Tech Review. „Garbage in – Garbage out“ war schon immer eine Herausforderung bei der Datenverarbeitung. Der einzige Unterschied bei generativer KI besteht darin, dass sie den Müll hervorragend glaubwürdig erscheinen lässt, was zu ungenauen Ergebnissen führt. LLMs neigen dazu, Fakten zu erfinden, die zu der Erzählung passen, die sie aufbauen. Daher müssen Unternehmen, die in generativer KI eine großartige Möglichkeit sehen, ihre Kosten mit KI-Äquivalenten zu senken, Deep Fakes effizient erkennen, Voreingenommenheit reduzieren und Risiken senken, um die Ehrlichkeit und Ethik der Systeme zu gewährleisten. Eine frei fließende Datenpipeline mit einer robusten KI-Infrastruktur, die Datenmobilität, Datenqualität, Datenverwaltung und Datenschutz durch End-to-End-Verschlüsselung und KI-Leitplanken unterstützt, ist für die Entwicklung verantwortungsvoller und erklärbarer generativer KI-Modelle von entscheidender Bedeutung.

Kundenszenario und NetApp

Abbildung 3: Workflow für maschinelles Lernen/große Sprachmodelle

[Abbildung 3: Workflow für maschinelles Lernen/große Sprachmodelle]

Trainieren wir oder optimieren wir? Die Frage, ob (a) ein LLM-Modell von Grund auf trainiert, ein vortrainiertes FM optimiert oder RAG verwendet werden soll, um Daten aus Dokumentrepositorys außerhalb eines Basismodells abzurufen und Eingabeaufforderungen zu erweitern, und (b) ob Open-Source-LLMs (z. B. Llama 2) oder proprietäre FMs (z. B. ChatGPT, Bard, AWS Bedrock) genutzt werden sollen, ist eine strategische Entscheidung für Unternehmen. Jeder Ansatz erfordert einen Kompromiss zwischen Kosteneffizienz, Datengravitation, Betrieb, Modellgenauigkeit und Verwaltung von LLMs.

NetApp als Unternehmen integriert KI intern in seine Arbeitskultur und in seinen Ansatz für Produktdesign und Entwicklungsbemühungen. Beispielsweise basiert der autonome Ransomware-Schutz von NetApp auf KI und maschinellem Lernen. Es ermöglicht die frühzeitige Erkennung von Dateisystemanomalien und hilft so, Bedrohungen zu identifizieren, bevor sie den Betrieb beeinträchtigen. Zweitens nutzt NetApp prädiktive KI für seine Geschäftsabläufe, beispielsweise für Verkaufs- und Bestandsprognosen, sowie Chatbots, um Kunden bei Produktsupportdiensten im Callcenter, technischen Daten, Garantien, Servicehandbüchern und vielem

mehr zu unterstützen. Drittens schafft NetApp Kundennutzen für die KI-Datenpipeline und den ML/LLM-Workflow durch Produkte und Lösungen, die Kunden beim Aufbau prädiktiver KI-Lösungen unterstützen, beispielsweise für Bedarfsprognosen, medizinische Bildgebung, Stimmungsanalysen und generative KI-Lösungen wie GANs zur Erkennung von Anomalien bei industriellen Bildern im Fertigungssektor sowie zur Geldwäschebekämpfung und Betrugserkennung im Bank- und Finanzdienstleistungssektor mit NetApp -Produkten und -Funktionen wie NetApp ONTAP AI, NetApp SnapMirror und NetApp FlexCache.

NetApp -Funktionen

Die Bewegung und Verwaltung von Daten in generativen KI-Anwendungen wie Chatbot, Codegenerierung, Bildgenerierung oder Genommodellausdruck kann sich über das Edge-, private Rechenzentrums- und hybride Multicloud-Ökosystem erstrecken. Beispielsweise kann ein Echtzeit-KI-Bot, der einem Passagier dabei hilft, sein Flugticket über eine Endbenutzer-App, die über APIs vortrainierter Modelle wie ChatGPT verfügbar ist, auf die Business Class upzugraden, diese Aufgabe nicht alleine bewältigen, da die Passagierinformationen nicht öffentlich im Internet verfügbar sind. Die API erfordert Zugriff auf die persönlichen Daten des Passagiers und die Ticketinformationen der Fluggesellschaft, die in einem Hybrid- oder Multicloud-Ökosystem vorhanden sein können. Ein ähnliches Szenario könnte für Wissenschaftler gelten, die ein Arzneimittelmolekül und Patientendaten über eine Endbenutzeranwendung austauschen, die LLMs verwendet, um klinische Studien im Rahmen der Arzneimittelforschung durchzuführen, an denen eine oder mehrere biomedizinische Forschungseinrichtungen beteiligt sind. Zu den sensiblen Daten, die an FMs oder LLMs weitergegeben werden, können personenbezogene Daten, Finanzinformationen, Gesundheitsinformationen, biometrische Daten, Standortdaten, Kommunikationsdaten, Daten zum Online-Verhalten und rechtliche Informationen gehören. Bei einem solchen Fall von Echtzeit-Rendering, sofortiger Ausführung und Edge-Inferenz werden Daten von der Endbenutzer-App über Open Source- oder proprietäre LLM-Modelle zu Speicherendpunkten in ein Rechenzentrum vor Ort oder auf öffentliche Cloud-Plattformen verschoben. In all diesen Szenarien sind Datenmobilität und Datenschutz für die KI-Operationen mit LLMs von entscheidender Bedeutung, die auf großen Trainingsdatensätzen und der Bewegung solcher Daten beruhen.

Abbildung 4: Generative KI – LLM-Datenpipeline

[Abbildung 4: Generative AI-LLM-Datenpipeline]

Das Portfolio von NetApp an Speicherinfrastruktur, Daten und Cloud-Diensten basiert auf intelligenter Datenverwaltungssoftware.

Datenvorbereitung: Die erste Säule des LLM-Tech-Stacks ist vom älteren traditionellen ML-Stack weitgehend unberührt. Die Datenvorverarbeitung in der KI-Pipeline ist notwendig, um die Daten vor dem Training oder der Feinabstimmung zu normalisieren und zu bereinigen. Dieser Schritt umfasst Konnektoren zum Aufnehmen von Daten, wo immer diese sich in Form einer Amazon S3-Ebene oder in lokalen Speichersystemen wie einem Dateispeicher oder einem Objektspeicher wie NetApp StorageGRID befinden.

- NetApp ONTAP* ist die grundlegende Technologie, die den kritischen Speicherlösungen von NetApp im Rechenzentrum und in der Cloud zugrunde liegt. ONTAP umfasst verschiedene Funktionen und Fähigkeiten zur Datenverwaltung und zum Schutz, darunter automatischen Ransomware-Schutz vor Cyberangriffen, integrierte Datentransportfunktionen und Speichereffizienzfunktionen für eine Reihe von Architekturen von lokalen, hybriden und Multicloud-Umgebungen in NAS-, SAN-, Objekt- und Software Defined Storage (SDS)-Situationen von LLM-Bereitstellungen.
- NetApp ONTAP AI* für das Training von Deep-Learning-Modellen. NetApp ONTAP unterstützt NVIDIA GPU Direct Storage durch die Verwendung von NFS über RDMA für NetApp -Kunden mit ONTAP -Speichercluster und NVIDIA DGX-Rechenknoten. Es bietet eine kosteneffiziente Leistung, um Quelldatensätze mehrfach aus dem Speicher zu lesen und in den Arbeitsspeicher zu übertragen, um die Intelligenz zu fördern und Organisationen Schulungen, Feinabstimmungen und Skalierungszugriff auf LLMs zu ermöglichen.

- NetApp FlexCache* ist eine Remote-Caching-Funktion, die die Dateiverteilung vereinfacht und nur die aktiv gelesenen Daten zwischenspeichert. Dies kann für LLM-Schulungen, Umschulungen und Feinabstimmungen nützlich sein und Kunden mit Geschäftsanforderungen wie Echtzeit-Rendering und LLM-Inferenz einen Mehrwert bieten.
- NetApp SnapMirror* ist eine ONTAP -Funktion, die Volume-Snapshots zwischen zwei beliebigen ONTAP Systemen repliziert. Diese Funktion überträgt Daten am Rand optimal in Ihr lokales Rechenzentrum oder in die Cloud. SnapMirror kann zum sicheren und effizienten Verschieben von Daten zwischen lokalen und Hyperscaler-Clouds verwendet werden, wenn Kunden generative KI in Clouds mit RAG entwickeln möchten, die Unternehmensdaten enthalten. Es überträgt effizient nur Änderungen, spart Bandbreite und beschleunigt die Replikation und bietet so wichtige Datenmobilitätsfunktionen während des Trainings, des erneuten Trainings und der Feinabstimmung von FMs oder LLMs.
- NetApp SnapLock* ermöglicht die unveränderliche Festplattenfunktion auf ONTAP-basierten Speichersystemen zur Datensatzversionierung. Die Microcore-Architektur ist darauf ausgelegt, Kundendaten mit der FPolicy Zero Trust-Engine zu schützen. NetApp stellt die Verfügbarkeit von Kundendaten sicher, indem es Denial-of-Service-Angriffen (DoS) widersteht, wenn ein Angreifer auf besonders ressourcenintensive Weise mit einem LLM interagiert.
- NetApp Cloud Data Sense* hilft dabei, in Unternehmensdatensätzen vorhandene persönliche Informationen zu identifizieren, abzubilden und zu klassifizieren, Richtlinien umzusetzen, Datenschutzanforderungen vor Ort oder in der Cloud zu erfüllen, die Sicherheitslage zu verbessern und Vorschriften einzuhalten.
- NetApp BlueXP*-Klassifizierung, unterstützt durch Cloud Data Sense. Kunden können Daten im gesamten Datenbestand automatisch scannen, analysieren, kategorisieren und darauf reagieren, Sicherheitsrisiken erkennen, die Speicherung optimieren und Cloud-Bereitstellungen beschleunigen. Es kombiniert Speicher- und Datendienste über seine einheitliche Steuerebene. Kunden können GPU-Instanzen für Berechnungen und hybride Multicloud-Umgebungen für Cold-Storage-Tiering sowie für Archive und Backups verwenden.
- NetApp -Datei-Objekt-Dualität*. NetApp ONTAP ermöglicht Dual-Protokoll-Zugriff für NFS und S3. Mit dieser Lösung können Kunden über S3-Buckets von NetApp Cloud Volumes ONTAP auf NFS-Daten von Amazon AWS SageMaker-Notebooks zugreifen. Dies bietet Flexibilität für Kunden, die einfachen Zugriff auf heterogene Datenquellen mit der Möglichkeit benötigen, Daten sowohl von NFS als auch von S3 gemeinsam zu nutzen. Beispielsweise zum Feinabstimmen von FMs wie Metas Llama 2-Textgenerierungsmodellen auf SageMaker mit Zugriff auf Dateiobjekt-Buckets.

Der * NetApp Cloud Sync*-Dienst bietet eine einfache und sichere Möglichkeit, Daten zu jedem Ziel in der Cloud oder vor Ort zu migrieren. Cloud Sync überträgt und synchronisiert Daten nahtlos zwischen lokalen oder Cloud-Speichern, NAS und Objektspeichern.

- NetApp XCP* ist eine Client-Software, die schnelle und zuverlässige Datenmigrationen von beliebigen zu NetApp und NetApp-zu- NetApp ermöglicht. XCP bietet außerdem die Möglichkeit, Massendaten effizient von Hadoop HDFS-Dateisystemen in ONTAP NFS, S3 oder StorageGRID zu verschieben, und die XCP-Dateianalyse bietet Einblick in das Dateisystem.
- NetApp DataOps Toolkit* ist eine Python-Bibliothek, die es Datenwissenschaftlern, DevOps und Dateningenieuren erleichtert, verschiedene Datenverwaltungsaufgaben auszuführen, wie etwa die nahezu sofortige Bereitstellung, das Klonen oder das Erstellen von Snapshots eines Datenvolumens oder JupyterLab-Arbeitsbereichs, die durch hochleistungsfähigen Scale-Out- NetApp -Speicher unterstützt werden.

Produktsicherheit von NetApp. LLMs können in ihren Antworten unbeabsichtigt vertrauliche Daten preisgeben, was für CISOs, die die Schwachstellen im Zusammenhang mit KI-Anwendungen untersuchen, die LLMs nutzen, Anlass zur Sorge gibt. Wie von OWASP (Open Worldwide Application Security Project) dargelegt, können Sicherheitsprobleme wie Datenvergiftung, Datenlecks, Denial-of-Service und Prompt-Injections innerhalb von LLMs Unternehmen beeinträchtigen, da Daten für Angreifer unberechtigten Zugriff zugänglich gemacht werden. Zu den Anforderungen an die Datenspeicherung sollten Integritätsprüfungen und

unveränderliche Snapshots für strukturierte, halbstrukturierte und unstrukturierte Daten gehören. Zur Datensatzversionierung werden NetApp Snapshots und SnapLock verwendet. Es bietet eine strenge rollenbasierte Zugriffskontrolle (RBAC) sowie sichere Protokolle und eine Verschlüsselung nach Industriestandard zum Schutz sowohl der Daten im Ruhezustand als auch während der Übertragung. Cloud Insights und Cloud Data Sense bieten gemeinsam Funktionen, die Ihnen dabei helfen, die Quelle der Bedrohung forensisch zu identifizieren und die wiederherzustellenden Daten zu priorisieren.

* ONTAP AI mit DGX BasePOD*

Die NetApp ONTAP AI-Referenzarchitektur mit NVIDIA DGX BasePOD ist eine skalierbare Architektur für Workloads im Bereich maschinelles Lernen (ML) und künstliche Intelligenz (KI). Für die kritische Trainingsphase von LLMs werden Daten typischerweise in regelmäßigen Abständen aus dem Datenspeicher in den Trainingscluster kopiert. Die in dieser Phase eingesetzten Server nutzen GPUs zur Parallelisierung von Berechnungen, wodurch ein enormer Datenhunger entsteht. Um eine hohe GPU-Auslastung aufrechtzuerhalten, ist es entscheidend, den Bedarf an Roh-E/A-Bandbreite zu decken.

* ONTAP AI mit NVIDIA AI Enterprise*

NVIDIA AI Enterprise ist eine durchgängige, Cloud-native Suite von KI- und Datenanalysesoftware, die von NVIDIA für die Ausführung auf VMware vSphere mit NVIDIA-zertifizierten Systemen optimiert, zertifiziert und unterstützt wird. Diese Software ermöglicht die einfache und schnelle Bereitstellung, Verwaltung und Skalierung von KI-Workloads in der modernen Hybrid-Cloud-Umgebung. NVIDIA AI Enterprise, unterstützt von NetApp und VMware, bietet KI-Workload- und Datenmanagement der Enterprise-Klasse in einem vereinfachten, vertrauten Paket.

1P Cloud-Plattformen

Vollständig verwaltete Cloud-Speicherangebote sind nativ auf Microsoft Azure als Azure NetApp Files (ANF), auf AWS als Amazon FSx for NetApp ONTAP (FSx ONTAP) und auf Google als Google Cloud NetApp Volumes (GNCV) verfügbar. 1P ist ein verwaltetes, leistungsstarkes Dateisystem, das es Kunden ermöglicht, hochverfügbare KI-Workloads mit verbesserter Datensicherheit in öffentlichen Clouds auszuführen, um LLMs/FMs mit Cloud-nativen ML-Plattformen wie AWS SageMaker, Azure-OpenAI Services und Googles Vertex AI zu optimieren.

NetApp Partner Solution Suite

Zusätzlich zu seinen Kerndatenprodukten, -technologien und -funktionen arbeitet NetApp auch eng mit einem robusten Netzwerk von KI-Partnern zusammen, um den Kunden einen Mehrwert zu bieten.

- NVIDIA Leitplanken* in KI-Systemen dienen als Sicherheitsvorkehrungen, um den ethischen und verantwortungsvollen Einsatz von KI-Technologien sicherzustellen. KI-Entwickler können das Verhalten von LLM-basierten Anwendungen zu bestimmten Themen definieren und sie daran hindern, sich an Diskussionen zu unerwünschten Themen zu beteiligen. Guardrails, ein Open-Source-Toolkit, bietet die Möglichkeit, ein LLM nahtlos und sicher mit anderen Diensten zu verbinden, um vertrauenswürdige und sichere LLM-Konversationssysteme aufzubauen.

Domino Data Lab bietet vielseitige Tools auf Unternehmensniveau zum Erstellen und zur Produktisierung generativer KI – schnell, sicher und wirtschaftlich, egal, wo Sie sich auf Ihrer KI-Reise befinden. Mit der Enterprise MLOps-Plattform von Domino können Datenwissenschaftler bevorzugte Tools und alle ihre Daten nutzen, Modelle überall problemlos trainieren und bereitstellen sowie Risiken und Kosten effizient verwalten – alles von einem Kontrollzentrum aus.

Modzy für Edge AI. NetApp und Modzy haben sich zusammengeschlossen, um KI im großen Maßstab für alle

Arten von Daten bereitzustellen, darunter Bilder, Audio, Text und Tabellen. Modzy ist eine MLOps-Plattform zum Bereitstellen, Integrieren und Ausführen von KI-Modellen. Sie bietet Datenwissenschaftlern die Möglichkeit zur Modellüberwachung, Drifterkennung und Erklärbarkeit mit einer integrierten Lösung für nahtlose LLM-Inferenz.

Run:AI und NetApp haben sich zusammengetan, um die einzigartigen Fähigkeiten der NetApp ONTAP AI-Lösung mit der Run:AI-Cluster-Management-Plattform zur Vereinfachung der Orchestrierung von KI-Workloads zu demonstrieren. Es teilt und verbindet GPU-Ressourcen automatisch und ist darauf ausgelegt, Ihre Datenverarbeitungs-Pipelines mit integrierten Integrationsframeworks für Spark, Ray, Dask und Rapids auf Hunderte von Maschinen zu skalieren.

Abschluss

Generative KI kann nur dann effektive Ergebnisse erzielen, wenn das Modell anhand großer Mengen hochwertiger Daten trainiert wird. Obwohl LLMs bemerkenswerte Meilensteine erreicht haben, ist es wichtig, ihre Grenzen, Designherausforderungen und Risiken im Zusammenhang mit Datenmobilität und Datenqualität zu erkennen. LLMs basieren auf großen und unterschiedlichen Trainingsdatensätzen aus heterogenen Datenquellen. Ungenaue oder verzerrte Ergebnisse der Modelle können sowohl Unternehmen als auch Verbraucher gefährden. Diese Risiken können Einschränkungen für LLMs entsprechen, die möglicherweise aus Herausforderungen im Datenmanagement im Zusammenhang mit Datenqualität, Datensicherheit und Datenmobilität entstehen. NetApp unterstützt Unternehmen dabei, die Komplexität zu bewältigen, die durch schnelles Datenwachstum, Datenmobilität, Multi-Cloud-Management und die Einführung von KI entsteht. Eine maßstabsgetreue KI-Infrastruktur und ein effizientes Datenmanagement sind entscheidend für den Erfolg von KI-Anwendungen wie generativer KI. Für die Kunden ist es entscheidend, alle Bereitstellungsszenarien abzudecken, ohne Kompromisse bei der Expansionsfähigkeit einzugehen, die die Unternehmen benötigen, und gleichzeitig Kosteneffizienz, Datenverwaltung und ethische KI-Praktiken unter Kontrolle zu halten. NetApp arbeitet ständig daran, Kunden dabei zu unterstützen, ihre KI-Bereitstellungen zu vereinfachen und zu beschleunigen.