LÖSUNGSÜBERSICHT

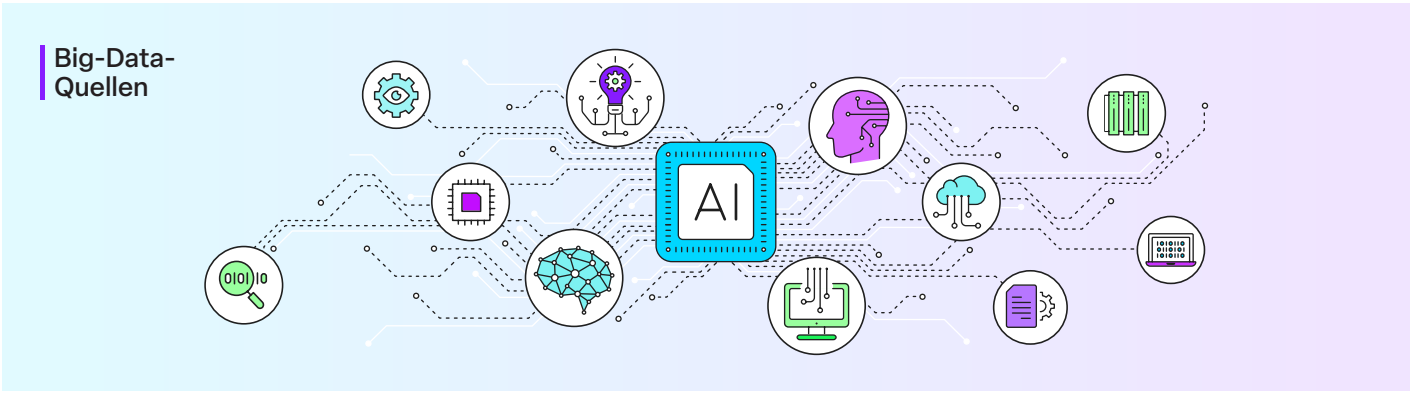
AI-Workloads nutzen die InfiniBox®

Retrieval-Augmented Generation (RAG) ist die neue "Killer-App", die maschinelles Lernen (ML) und künstliche Intelligenz (KI) vorantreibt

Unternehmen aller Größenordnungen sammeln große Mengen unterschiedlicher Datensätze in riesigen Volumina mit der Absicht, diese Daten für maschinelles Lernen, prädiktive Modellierung und andere fortschrittliche Analysen zu nutzen, um Geschäftsprobleme zu lösen und umsetzbare Erkenntnisse zu gewinnen.

Der Betrieb von KI-Sprachmodellen (Large Language Models, LLM) stellt jedoch eine Herausforderung dar, da sie eine erhebliche Menge an Rechenressourcen und Energie (Strom und Kühlung) benötigen. Aus diesem Grund investieren nicht viele Unternehmen in diese Technologie, sondern nutzen stattdessen die KI-Dienste von Cloud-Anbietern. Ein kleineres, kompaktes Modell, das sogenannte Small Language Model (SLM), ist ebenfalls aufgetaucht und populär geworden. Da weniger Parameter und Rechenressourcen erforderlich sind, sind SLMs praktischer und leichter anpassbar für Unternehmen, die die hohen Kosten für den Betrieb und das Training von LLMs nicht rechtfertigen können, insbesondere für ihr erstes KI-Projekt.

LLMs und SLMs sind in der Regel auf öffentlich zugängliche Daten geschult und haben keinen Zugang zu spezifischen Informationen und privaten Daten eines Unternehmens. Diese Einschränkungen führen oftmals zu ungenauen und unzuverlässigen Antworten, die als „Halluzinationen“ bezeichnet werden. Diese Tatsache wird von vielen öffentlich verfügbaren generativen KI-Chatbots, z. B. ChatGPT, anerkannt, die Warnungen über mögliche Fehlinformationen anzeigen. Diese Halluzinationen erscheinen oft als plausible Antworten, sind aber falsch und können die Kundenerfahrung eines Unternehmens in vielerlei Hinsicht beeinträchtigen.



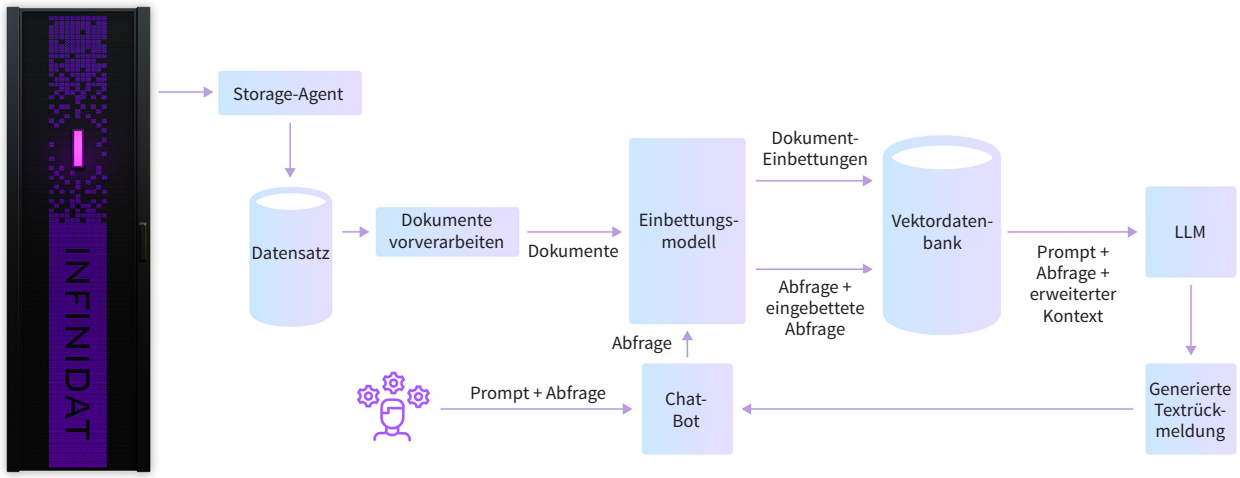
Retrieval-Augmented Generation (RAG) mit InfiniBox zur Behebung von Halluzinationen
Die Verwendung ungenauer Daten kann zu falschen Geschäftsentscheidungen führen und Kunden mit falschen Informationen versorgen, was dem Ruf des Unternehmens schadet und zu Ineffizienzen im Betrieb führt.

Generative KI (GenAI) ist zwar leistungsfähig, muss aber ständig mit neuen, zuverlässigen und genauen Daten ergänzt werden, um ihre Zuverlässigkeit zu verbessern und aufrechtzuerhalten. Retrieval-Augmented Generation (RAG) hat sich als wertvolles Tool erwiesen, um diese Lücke zu schließen und die Genauigkeit zu verbessern. RAG kombiniert die Leistungsfähigkeit von GenAI mit den privaten Daten eines Unternehmens, um neue Erkenntnisse zu gewinnen und Arbeitsabläufe zu optimieren. Durch die Integration von Informationen aus externen Quellen in LLMs und SLMs können Unternehmen genauere und spezifischere Ergebnisse erzielen.

Aber wo soll man anfangen? Das muss nicht zwangsläufig eine entmutigende Aufgabe sein. Die Wahrscheinlichkeit ist groß, dass die Benutzer Daten auf bestehenden Systemen und Speicherplätzen in ihrer eigenen Infrastruktur haben, die genutzt werden können. Für spezielle KI-Workloads ist es wichtig zu wissen, dass KI nur ein weiterer Workload für die InfiniBox ist, eine Speicherlösung für Unternehmen, die dafür bekannt ist, dass sie aufgrund ihres innovativen Designs skalierbare, konsistente E/A für verschiedene, moderne Workloads liefert.

Ein RAG-Workflow kann leicht aus vorhandenen Open-Source-Produkten und Daten erstellt werden, die sich bereits im Rechenzentrum eines Unternehmens befinden. Die Infinidat-Entwickler erstellten eine RAG-Workflow-Architektur, die den Prozess skizziert.

Infinidat RAG Workflow Architektur[1]



Die RAG-Workflow-Architektur von Infinidat läuft auf einem Kubernetes-Cluster. Anwender, die RAG mit Daten vor Ort betreiben wollen, aber keine GPU-Ressourcen zur Verfügung haben, können eine schnelle und bequeme Lösung über die Cloud nutzen. Unser Ansatz verwendet einen Kubernetes-Cluster als Grundlage für die Ausführung der RAG-Pipeline, was hohe Verfügbarkeit, Skalierbarkeit und Ressourceneffizienz ermöglicht. Mit AWS Terraform vereinfachen wir die Einrichtung eines RAG-Systems erheblich, indem wir die gesamte Automatisierung mit nur einem Befehl ausführen. Der gleiche Kerncode, der zwischen der InfiniBox vor Ort und der InfuzeOS™ Cloud Edition läuft, macht die Replikation zu einem Kinderspiel. Innerhalb von zehn Minuten ist ein voll funktionsfähiges RAG-System auf der InfuzeOS Cloud Edition für die Arbeit mit Ihren Daten bereit.

Die Entwicklung einer RAG-Pipeline ist von Natur aus ein iterativer Prozess, der aktualisiert und gepflegt werden muss, um die Genauigkeit aufrechtzuerhalten. Indem sie mit den neuesten Fortschritten Schritt halten und ihre RAG-Pipeline kontinuierlich verfeinern, können Unternehmen die Genauigkeit und Praxistauglichkeit ihrer KI-gestützten Erkenntnisse erheblich verbessern und die Wettbewerbsvorteile dieser neuen Technologie maximieren.

Der InfiniBox-Vorteil für RAG

Bestehende Infinidat-Kunden können davon ausgehen, dass die InfiniBox in ihrem Rechenzentrum bereits über das „Big Data"-Repository an Datensätzen verfügt, um die notwendigen Dokumente für die Einbettung in ein Sprachmodell zu erstellen. Die Auswahl einer Datenbank zum Speichern von Einbettungen (Vektoren) ist für Ihr RAG-System von grundlegender Bedeutung. Glücklicherweise unterstützen die aktuellen Versionen gängiger Datenbank-Engines wie Oracle, Postgres, MongoDB und DataStax Enterprise das Speichern und Abrufen von Vektordaten und sind damit RAG-fähig. Viele unserer aktuellen Kunden setzen ihre InfiniBox in geschäftskritischen Anwendungen ein, die Oracle verwenden.

Unsere solide Datenserviceschicht, die von InfuzeOS bereitgestellt wird, ermöglicht die Steuerung aller Infrastrukturkomponenten durch Software - eine sehr leistungsfähige Fähigkeit. InfuzeOS nutzt unseren innovativen Neural Cache zum dynamischen, vorausschauenden Optimieren der Datenanordnung. Die aktivsten Daten – Hot Data – werden im DRAM-Cache gespeichert, damit der Datenabruf von SSD- oder HDD-Laufwerken auf ein Minimum reduziert wird. InfuzeOS sorgt für eine kontinuierlich optimierte Leistung bei der Skalierung, eine erstklassige niedrige Latenz und eine kompetente Betriebsunterstützung für unterschiedliche moderne Workloads in On-Premises- und Cloud-Umgebungen.

Fortune Business Insights geht davon aus, dass der Markt für Big-Data-Technologien von 349,40 Milliarden Dollar im Jahr 2023 auf 1.194,35 Milliarden Dollar im Jahr 2032 wachsen wird.

Für einen RAG-Workload bedeutet dies, dass InfiniBox während der Einbettungsphase, in der hochdimensionale Vektoren in einer Datenbank gespeichert werden, eine blitzschnelle Leistung mit sehr geringer Latenz erreicht, die den Prozess nicht behindert.

Wenn ein Nutzer eine Frage stellt (z. B. ChatGPT), wird ihre Anfrage in eine Einbettung umgewandelt, die sich im gleichen Raum wie die bereits vorhandenen Einbettungen in der Vektordatenbank befindet. Bei der Ähnlichkeitssuche ermitteln Vektordatenbanken schnell die Vektoren, die der Anfrage am nächsten liegen und diese beantworten. Auch hier ermöglicht die extrem niedrige Latenz der InfiniBox schnelle Reaktionen für GenAI-Workloads.

Darüber hinaus bietet die Speicherung der Vektordatenbank auf der InfiniBox einige Vorteile. Das Einlesen der Daten kann ressourcen- und zeitintensiv sein, wobei die Ergebnisse in der Vektordatenbank gespeichert werden. Diese Datenbank kann als Snapshot gespeichert und/oder repliziert werden, sodass eine Vektordatenbank zur Verfügung steht, die in andere RAG-Workflow-Implementierungen integriert werden kann und den Ressourcenbedarf in anderen Implementierungen reduziert.

Sobald die Daten vorverarbeitet und in einem effizienten, schnell abrufbaren Datenspeichersystem (InfiniBox) gespeichert sind, kommt das Large Language Model (LLM) ins Spiel. LLMs geben kohärente und kontextbezogene Antworten. Nach dem Abruf der relevanten Daten aus der Datenbank kombiniert das LLM das abgerufene Wissen mit seinem Verständnis, um eine umfassende Antwort zu geben.

Schließlich bietet die InfuzeOS Cloud Edition für AWS und Azure Cloud-basierte Speichervorgänge, die denen vor Ort ähneln, sodass Speicheradministratoren die gleichen leistungsstarken InfuzeOS-Funktionen des Rechenzentrums in der Cloud nutzen können, um Datensätze schnell auf KI-Ressourcen zu replizieren. Diese Speicherstandardisierung reduziert die Kosten und die Komplexität der Speicherverwaltung in allen Umgebungen.

Der Vorteil von InfiniBox gegenüber RAG liegt in erster Linie darin, dass wir mit Innovationen wie Neural Cache die leistungsstärkste Speicherplattform für diesen Workload anbieten können. Die Lösung von Infinidat kann eine beliebige Anzahl von InfiniBox-Plattformen umfassen und ermöglicht die Erweiterbarkeit auf Speicherlösungen von Drittanbietern über dateibasierte Protokolle wie NFS. Aber vergessen wir nicht, dass wir auch 100-prozentige Verfügbarkeit und Cyber-Recovery garantieren, alles wesentliche Bestandteile einer effizienten KI-Infrastruktur.

Der Einsatz von GenAI und Retrieval-Augmented Generation (RAG) wird wesentlich dazu beitragen, dass KI-Modelle genauer und relevanter werden. Unternehmen können ihre bereits getätigten Investitionen in die InfiniBox-Architektur zusammen mit dem oben beschriebenen RAG-Workflow nutzen, um schnelle und reaktionsfähige KI-Modelle für ihre bestehenden privaten Datensätze zu erstellen, neue Erkenntnisse zu gewinnen und Arbeitsabläufe zu optimieren.

[1] Entwurf und Entwicklung des RAG-Arbeitsablaufs, der Andrew Wang zuzuschreiben ist

dach@infinidat.com

INFINIDAT