Red Hat OpenShift AI

Eine offene Plattform für KI-Modelle in der Hybrid Cloud

Highlights

Operationalisieren und skalieren Sie KI-Inferenz und agentische KI auf einer bewährten Basis für Anwendungsplattformen.

Die operative Effizienz von KI/ML in den Teams wird durch eine konsistente Benutzeroberfläche verbessert, die Data Scientists, Data Engineers, Anwendungsentwicklungsteams und DevOps-Teams unterstützt.

Profitieren Sie von der Flexibilität der Hybrid Cloud, indem Sie KI/ML-Workloads lokal, in einer Cloud oder am Edge erstellen, trainieren, bereitstellen und überwachen.

Intelligente Anwendungen und generative KI

Künstliche Intelligenz (KI), Machine Learning (ML) und Deep Learning (DL) beeinflussen zunehmend die Modernisierung von Anwendungen in verschiedenen Unternehmen und Branchen. Die Notwendigkeit, innovativ zu sein und strategischen Mehrwert sowie neue Insights aus Daten zu gewinnen, führt zu einer zunehmenden Nutzung KI-gestützter, cloudnativer Anwendungen sowie von MLOps und GenAIOps-Methoden. Die Herausforderungen in dieser „schönen neuen Welt" können komplex sein – sie reichen von schnell steigenden Modellkosten beim Produktivstart über komplexe Anpassungen und starre Deployment-Beschränkungen bis hin zu den erforderlichen Abläufen, um mit dem Innovationstempo Schritt zu halten. Unternehmen benötigen Lösungen, mit denen Sie Inferenzkosten senken, die Skalierung und das Monitoring vereinfachen und sich an ständige Veränderungen anpassen können.

Red Hat® AI beschleunigt die Entwicklung und Bereitstellung von Unternehmens-KI-Lösungen in Hybrid Cloud-Umgebungen. Die Lösung dient als umfassende Plattform für das Management des gesamten KI/ML-Lifecycles und bietet MLOps- sowie GenAI Ops-Funktionen. Red Hat AI konzentriert sich speziell auf 4 wichtige Aspekte:

- Steigerung der Effizienz durch schnelle, flexible und effiziente Inferenz
- Vereinfachtes Verbinden von Modellen mit Daten
- Beschleunigte Innovation agentischer KI
- Verlässliche Flexibilität und Konsistenz bei der KI-Skalierung in der Hybrid Cloud

Red Hat OpenShift® AI basiert auf Red Hat OpenShift, einer führenden Hybrid Cloud-Anwendungsplattform, und ist das Spitzenprodukt im Portfolio von Red Hat AI. Die KI-Plattform bietet KI-Engineers, Data Scientists und Entwicklungsteams eine leistungsstarke KI/ML-Basis für die Entwicklung und Bereitstellung generativer und prädiktiver Modelle sowie KI-gestützter Anwendungen in großem Umfang. Über eine einzige gemeinsame Plattform können Unternehmen mit verschiedenen Tools ihrer Wahl experimentieren, zusammenarbeiten und Markteinführungszeiten verkürzen. Red Hat OpenShift AI kombiniert die Self Service-Umgebung, die sich Data Scientists und Entwicklungsteams wünschen, mit der Zuverlässigkeit, die eine Unternehmens-IT erfordert.

Beschleunigte Entwicklungs-, Trainings-, Test- und Deployment-Prozesse

Red Hat OpenShift AI ist eine flexible, skalierbare MLOps-Plattform, die mit Open Source-Technologien entwickelt wurde und zuverlässige sowie operativ konsistente Funktionen bietet, mit denen Teams experimentieren, sowie Modelle und innovative Anwendungen bereitstellen können. OpenShift AI beschleunigt die Bereitstellung KI-gestützter Anwendungen und unterstützt Unternehmen dabei, schneller und kontrollierter von frühen Pilotprojekten zu operativ robusten Deployments überzugehen.

Die Plattform bietet eine integrierte Benutzeroberfläche (UI) mit Tools zum Entwickeln, Trainieren, Tuning, Deployment und Monitoring von prädiktiven und gen KI-Modellen. Sie können Modelle in Hybrid Cloud-Umgebungen bereitstellen, wobei der besondere Schwerpunkt auf der Bereitstellung eines kontrollierten und geschützten Footprints für souveräne und private KI liegt. Durch diesen Ansatz lässt sich sicherstellen, dass sensible Daten und KI-Modelle innerhalb festgelegter geografischer oder organisatorischer Grenzen verbleiben und dabei strenge regulatorische und Compliance-Anforderungen eingehalten werden.

Red Hat
OpenShift AI

Gen KI Analystenprognose

„KI gilt als wichtiger Faktor für die Budgets für digitale Infrastruktur im Jahr 2026, da Unternehmen ihre Workloads und Datenanforderungen an hybride Infrastrukturlösungen anpassen wollen. 90 % der Entscheidungstragenden glauben, dass KI ein wichtiger Faktor für ihre Budgets für digitale Infrastruktur und ihre Technologieentscheidungen für 2026 sein wird."[1]

Vereinfachte KI-Einführung

OpenShift AI ist als Add-on zu Red Hat OpenShift verfügbar und bietet eine Plattform, die für eine beschleunigte KI-Einführung und ein gestärktes Vertrauen in KI-Initiativen entwickelt wurde, indem Open Source Communities mit einem robusten KI-Ökosystem kombiniert werden. Dies bietet mehr Flexibilität und Freiheit bei der Auswahl der richtigen KI/ML-Technologie für Ihr Unternehmen. Nutzende können ihre prädiktiven Modelle erstellen oder mit einem externen gen KI-Modell beginnen und es dann mit Retrieval-Augmented Generation (RAG) erweitern, indem sie einen der verschiedenen in der Plattform bereitgestellten Modellserver verwenden. Die Plattform bietet schnellen Zugriff auf optimierte und validierte Drittanbieter-Modelle wie Llama, Mistral, DeepSeek und Granite, die effizient auf vLLM ausgeführt werden, welches im Red Hat AI Repository auf Hugging Face verfügbar ist. Im Katalog können Nutzende diese Modelle untersuchen und eigene hinzufügen.

Verbesserte operative Konsistenz in Teams

Red Hat OpenShift AI sorgt für ein konsistentes Benutzererlebnis, das Data Scientists, KI-Engineers, Entwicklungs- und DevOps-Teams die effektive Zusammenarbeit und so die zügige Bereitstellung von KI-Lösungen ermöglicht. Die Plattform bietet Self Service-Zugriff auf kollaborative Workflows, GPU-Beschleunigung (Graphic Processing Unit) und optimierte Abläufe. Sie ermöglicht eine konsistente Bereitstellung von KI-Lösungen in großem Umfang in Hybrid Cloud-Umgebungen und am Netzwerkrand.

IT-Abläufe profitieren von vereinfachten Konfigurationen und stärker automatisierten Workflows auf einer bewährten Plattform, die sich mit geringem Aufwand vertikal skalieren lässt und gleichzeitig bessere Governance und Sicherheit bietet.

Flexibilität der Hybrid Cloud

Red Hat OpenShift AI ermöglicht das Trainieren, Deployment und Monitoring von KI/ML-Workloads in verschiedenen Umgebungen – in Clouds, On-Premise-Rechenzentren oder Air Gap-Umgebungen – zur Einhaltung von gesetzlichen, Sicherheits- und Datenanforderungen. Die Plattform ist mit mehreren KI-Beschleunigern von Anbietern wie NVIDIA, AMD und Intel kompatibel. Diese Funktion kann erweitert werden, um eine GPU as a Service-Umgebung zu erstellen, mit der Unternehmen GPU-Ressourcen zentral verwalten, partitionieren und planen sowie gleichzeitig deren Verwendung detailliert beobachten können.

Gen KI und agentische KI

Für gen KI-Projekte werden dedizierte Benutzererlebnisse durch Komponenten wie AI Hub (Vorschau für Entwicklungsteams) angeboten. Dabei handelt es sich um ein Dashboard für Platform Engineers, das Katalog, Registry und Modellbereitstellungen zum Einrichten und Deployment von Modellen und MCP-Servern konsolidiert. Gen AI Studio (Vorschau für Entwicklungsteams) bietet KI-Asset-Endpunkte und eine KI-Plattform, in der KI-Engineers und Anwendungsentwicklungsteams auf bereitgestellte Modelle und MCP-Server zugreifen und experimentieren, diese vergleichen, bewerten und testen können.

OpenShift AI beschleunigt die agentische KI durch die Bereitstellung einer einheitlichen API-Schicht und einer flexiblen, skalierbaren Basis. Der Support für Llama Stack API und MCP (Technologievorschau) umfasst eine unternehmensgerechte Implementierung der Llama Stack API, die einen einzigen, standardisierten Einstiegspunkt für verschiedene KI-Funktionen bietet.

Mit zusätzlichen Tools wie LLM-Evaluierung (LM Eval) und LLM-Benchmarking können Sie reale Inferenz-Deployments unterstützen. LLM Compressor bietet Algorithmen, mit denen die benutzerdefinierten Modelle eines Unternehmens verkleinert werden können, und zwar mit ähnlichen Methoden, mit denen Red Hat validierte und optimierte Modelle im Red Hat AI Repository auf Hugging Face erstellt.

---

1 *IDC Tech Supplier. „AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Dok.-Nr.: US53418426, Okt. 2025. (erfordert Client-Anmeldung)*

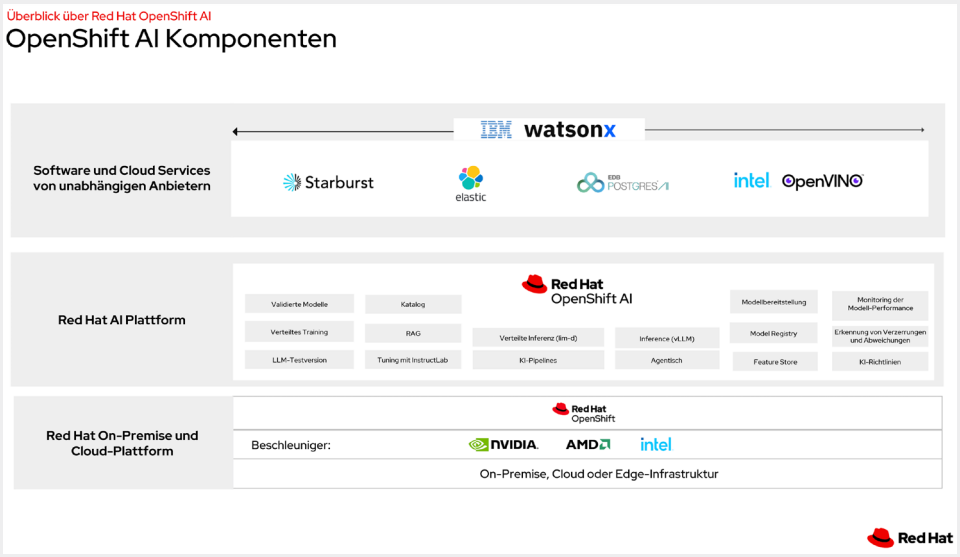

Abbildung 1. Red Hat OpenShift AI und optionale Partnerkomponenten

Mehrere andere zentrale Tools und Funktionen von Red Hat OpenShift AI bieten eine solide Basis:

- **Modellerstellung und -anpassung:** Data Scientists können explorative Data Science in einer JupyterLab-Benutzeroberfläche durchführen, die sofort einsatzbereite, sicher erstellte Notebook Images mit gängigen Python-Libraries bietet. Für gen KI-Projekte ermöglicht OpenShift AI Retrieval Augmented Generation (RAG) und ein verteiltes InstructLab-Training – mit Tools für die Modellanpassung, um Kompetenzen und Wissen effizienter in gen KI-Modelle einzubringen.
- **Modellbereitstellung:** Red Hat OpenShift AI bietet eine Vielzahl von Frameworks, die KServe als Kern-Engine für die Modellbereitstellung nutzen, um das Deployment prädiktiver Machine Learning- oder Basismodelle in Produktivumgebungen zu vereinfachen. Für LLMs, die maximale Skalierbarkeit erfordern, bietet OpenShift AI parallelisierte Bereitstellung mit vLLM-Runtimes. LIm-d bietet ein Framework zum Optimieren der LLM-Inferenz durch Disaggregieren der Pipeline in modulare Dienste, die intelligente automatische Skalierung und effizientes Anfrage-Routing unterstützen.
- **KI-Pipelines:** Red Hat OpenShift AI enthält auch eine Komponente für Pipelines, mit der Sie KI-Aufgaben in Pipelines orchestrieren und Pipelines über ein grafisches Frontend erstellen können. Unternehmen können Prozesse wie Datenaufbereitung, Modellentwicklung und Modellbereitstellung miteinander verknüpfen.
- **Modell-Monitoring:** Red Hat OpenShift AI unterstützt Ops-orientierte Nutzende beim Monitoring von Betriebs- und Performance-Metriken für Modellserver und bereitgestellte Modelle. Nutzende können sofort auf einsatzbereite Visualisierungen für Performance- und Ablaufsmetriken zugreifen oder Daten mit anderen Beobachtbarkeitsservices integrieren.
- **Verteilte Workloads:** Mit verteilten Workloads können Teams die Datenverarbeitung sowie das Training, Tuning und die Bereitstellung von Modellen beschleunigen. Diese Fähigkeit unterstützt die Priorisierung und Verteilung der Aufgabenausführung zusammen mit einer optimalen Knotenauslastung. Der fortschrittliche GPU-Support hilft bei der Bewältigung der Workload-Anforderungen von Basismodellen.

- **KI-Richtlinien, Erkennung von Verzerrungen und Abweichungen:** Red Hat OpenShift AI bietet Tools, mit denen Data Scientists und KI-Engineers überwachen können, ob die Modelle auf der Basis der Trainingsdaten korrekt und unvoreingenommen arbeiten, aber auch, ob sie im realen Deployment korrekt funktionieren. KI-Richtlinien bieten ein anpassbares Framework zum Implementieren wichtiger Sicherheitskontrollen und tragen dazu bei, dass die Modelle in der Produktion transparent, fair und zuverlässig sind. Zu den Tools zur Abweichungserkennung gehören Eingabedatenverteilungen für bereitgestellte ML-Modelle, um zu erkennen, wenn die für die Modellinferenz verwendeten Live-Daten erheblich von den Daten abweichen, mit denen das Modell trainiert wurde.
- **Katalog und Registry:** Red Hat OpenShift AI bietet einen internen Modellkatalog und einen kuratierten Katalog, in dem Platform Engineers optimierte genKI-Modelle finden, vergleichen und bewerten können. Die Plattform bietet auch eine zentrale Registry, die Data Scientists und KI-Engineers beim Teilen, Modifizieren, Bereitstellen und Nachverfolgen von prädiktiven und gen KI-Modellen, Metadaten und Modellartefakten unterstützt.
- **Feature Store:** Verwalten Sie bereinigte, klar definierte Daten-Features für ML-Modelle, um die Performance zu verbessern und Workflows zu beschleunigen.

Tools für den kompletten KI-Lifecycle

Red Hat OpenShift bietet die Services und Software, mit denen Unternehmen ihre Modelle erfolgreich trainieren, bereitstellen und in die Produktivumgebung verschieben können (siehe Abbildung 2).

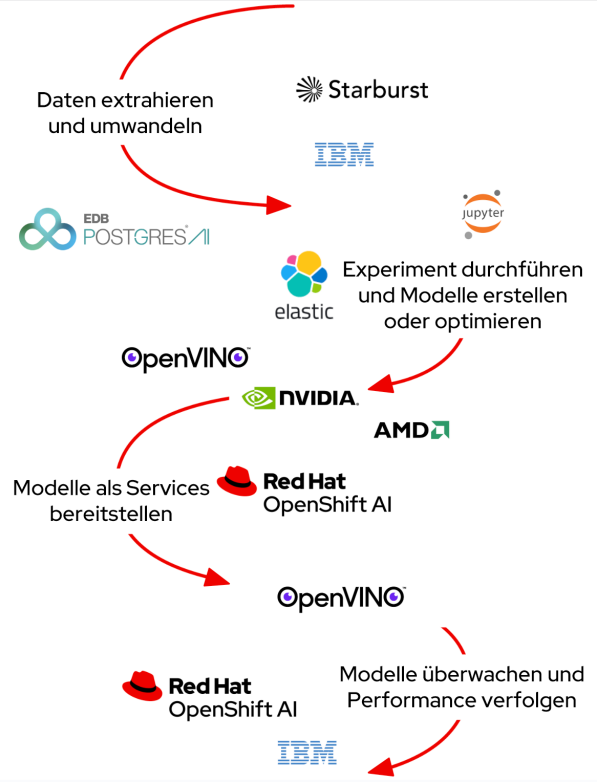

Abbildung 2. Typischer Lifecycle für die Modelloperationalisierung

Red Hat OpenShift AI

Das Dashboard von Red Hat OpenShift AI ist ein zentraler Ort zum Auffinden von und Zugreifen auf Anwendungen und Dokumentationen und erleichtert so die Einführung. Die Smart Start-Tutorials bieten Best Practice-Anleitungen für gängige Komponenten und integrierte Partnersoftware und sind direkt über das Dashboard verfügbar. So können Data Scientists in kürzester Zeit lernen und starten. In den folgenden Abschnitten werden die in Red Hat OpenShift AI integrierten Tools der Technologiepartner beschrieben. Für einige Tools ist eine zusätzliche Lizenz des jeweiligen Technologiepartners erforderlich.

Starburst

Starburst

Starburst beschleunigt Analysen, indem es Ihren Teams die schnelle und einfache Nutzung Ihrer Daten ermöglicht, um Prozessabläufe im Unternehmen zu verbessern. Starburst wird als selbst gemanagtes Produkt oder als vollständig gemanagter Service bereitgestellt. Es demokratisiert den Datenzugriff und bietet Datenkonsumenten umfassende Insights. Das Tool basiert auf Open Source Trino (früher bekannt als PrestoSQL), der ersten MPP-SQL-Engine (Massively Parallel Processing, Structured Query Language). Starburst wurde von Expertinnen und -Experten von Trino entwickelt und wird von ihnen betrieben. Mit diesem Tool können Sie flexibel verschiedene Datensätze unabhängig vom jeweiligen Speicherort abfragen, ohne Ihre Daten verschieben zu müssen.

Starburst lässt sich in die skalierbaren Cloud Storage- und Computing-Services von Red Hat OpenShift integrieren und bietet damit eine stabile, sicherheitsorientierte, effiziente und kostengünstige Möglichkeit zur Abfrage Ihrer Unternehmensdaten. Dies bietet unter anderem folgende Vorteile:

- **Automatisierung:** Operatoren von Starburst und Red Hat OpenShift bieten automatische Konfiguration, automatische Optimierung und automatisches Management von Clustern.
- **Hohe Verfügbarkeit und schrittweises Herunterskalieren:** Red Hat OpenShift Load Balancer kann Services wie den Trino-Koordinator in einem Always On-Zustand halten.
- **Elastische Skalierbarkeit:** Red Hat OpenShift kann den Trino-Worker-Cluster basierend auf der Abfragelast automatisch skalieren.

Hewlett Packard Enterprise

HPE Machine Learning Data Management Software

Unternehmen benötigen Datenverwaltungslösungen, die unterschiedliche Vorgänge, von Laptop-Experimenten bis hin zu kritischen Unternehmensbereitstellungen, erleichtern. HPE Machine Learning Data Management Software (vormals Pachyderm) ermöglicht Data Scientists containerisierte, datengesteuerte ML-Pipelines mit einer garantierten Datenherkunft zu erstellen und zu skalieren, die durch automatische Datenversionierung bereitgestellt wird. HPE Machine Learning Data Management Software wurde zur Lösung realer Data Science-Probleme entwickelt und bietet die Datenbasis, mit der Teams ihren ML-Lifecycle mit Reproduzierbarkeit automatisieren und skalieren können. Mit Use Cases, die von unstrukturierten Daten bis hin zu Data Warehouses,Natural Language Processing (NLP), Video- und Bild-ETL (Extrahieren, Transformieren und Laden), Finanzdienstleistungen und Life Sciences reichen, bietet die HPE Machine Learning Data Management Software:

- Automatisierte Datenversionierung, die Teams eine leistungsstarke Möglichkeit bietet, den Überblick über Datenänderungen zu behalten
- Datengesteuerte, containerisierte Pipelines zur Beschleunigung der Datenverarbeitung und Verringerung der Computing-Kosten
- Eine unveränderliche Datenherkunft, die einen festen Datensatz für Aktivitäten und Assets im ML-Lifecycle bereitstellt
- Eine Konsole zur intuitiven Visualisierung Ihres Directed Acyclical Graphs (DAG) und Hilfestellung beim Debuggen sowie bei der Reproduzierbarkeit
- Support für Jupyter Notebooks mit der JupyterLab Mount Extension für eine Point and Click-Schnittstelle zu versionierten Daten

- Enterprise Administration mit robusten Tools für die Bereitstellung und Verwaltung von HPE Machine Learning Data Management Software in großem Umfang und in verschiedenen Teams innerhalb des Unternehmens

NVIDIA beschleunigt das Deployment von KI-Lösungen

Da KI/ML-Anwendungen zunehmend wichtiger für geschäftlichen Erfolg werden, benötigen Unternehmen Plattformen, die komplexe Workloads bewältigen, die Hardwarenutzung optimieren und Skalierbarkeit bieten. Skalierbare Datenverarbeitung, Datenanalyse, ML-Training und Inferenz stellen äußerst ressourcenintensive Rechenaufgaben dar. Mit Software von NVIDIA können Sie sämtliche Aspekte der End-to-End-Data Science beschleunigen, indem die Parallelverarbeitungsfähigkeiten von GPUs genutzt werden.

NVIDIA NIM verbessert die Verwaltung und Performance von NVIDIA-GPUs in der Umgebung von Red Hat OpenShift, sodass KI-Anwendungen das volle Potenzial der KI-Software und -Hardware von NVIDIA nutzen können. Durch die Integration von NVIDIA NIM und Red Hat OpenShift AI lassen sich Ressourcen besser zuweisen, die Effizienz steigern und KI-Workloads produktiver ausführen.

Intel OpenVINO Toolkit

Das Intel OpenVINO Toolkit beschleunigt die Entwicklung und Bereitstellung leistungsstarker DL-Inferenzanwendungen auf Intel-Plattformen. Mit dem Toolkit können Sie neuronale Netzmodelle virtuell übernehmen, optimieren und anpassen sowie umfassende KI-Inferenzprozesse mit dem OpenVINO-Ökosystem von Entwicklungstools ausführen.

- **Modellierung:** Softwareentwicklungsteams haben die Flexibilität, ihre eigenen DL-Modelle zu verwenden. Für eine schnellere Markteinführung können die Teams auch vortrainierte und optimierte Modelle verwenden, die durch die Zusammenarbeit von Intel mit Hugging Face für das OpenVINO-Toolkit verfügbar sind. OpenVINO unterstützt Pytorch, ONNX, TensorFlow und andere gängige Modellformate.
- **Optimierung:** Das OpenVINO-Toolkit bietet mehrere Möglichkeiten zur Konvertierung von Modellen für mehr Komfort und Performance und unterstützt Softwareentwicklungsteams dabei, KI-Modelle schneller und effizienter auszuführen. Entwicklungsteams können die Modellkonvertierung überspringen und die Inferenz direkt aus den Formaten PyTorch, ONNX, TensorFlow, TensorFlow Lite, JAX, oder PaddlePaddle ausführen. Die Konvertierung in OpenVINO IR bietet optimale Performance, die durch Verwenden von Funktionen zur Gewichtskomprimierung und Quantisierung, die im Neural Network Compression Framework von OpenVINO verfügbar sind, weiter optimiert werden kann. Dieselben Features reduzieren auch den Storage und Runtime Footprint.
- **Deployment:** Die OpenVINO Runtime Inference Engine ist eine API (Application Programming Interface), die zur Beschleunigung des Inferenzprozesses in Ihre Anwendungen integriert werden kann. Der Ansatz „einmal erstellen, umgebungsunabhängig bereitstellen" ermöglicht die effiziente Ausführung von Inferenzaufgaben auf verschiedener Hardware von Intel, einschließlich CPUs (Central Processing Units), GPUs, NPUs und FPGAs. Die OpenVINO GenAI-Erweiterungs-Library vereinfacht das Deployment von gen KI-Workloads und reduziert in vielen Fällen den benötigten Code auf nur 3 bis 5 Zeilen. OpenVINO Model Server bietet mehrere Features für agentische und Modellbereitstellungsszenarien, mit denen der Entwicklungsaufwand noch weiter reduziert wird.

EDB

EDB Postgres AI ist eine leistungsstarke und intelligente Plattform, die für transaktionale, analytische sowie KI-Workloads entwickelt wurde und beispiellose Flexibilität bietet, unabhängig davon, ob die Daten On-Premise oder in einer beliebigen Cloud-Umgebung gespeichert werden. Als weltweit führender Anbieter von Postgres-Datenbanklösungen für Unternehmen bietet EDB eine offene, unternehmensgerechte souveräne Daten- und KI-Plattform, mit der KI-Projekte bis zu 3-mal schneller produktiv werden können. Durch die Integration mit Red Hat OpenShift AI ermöglicht EDB Postgres AI Nutzenden den Aufbau robuster KI-Wissensdatenbanken für Retrieval-Augmented Generation (RAG) und vereint KI-Daten, Modelle und Anwendungen in einer souveränen Full-Stack-KI-Plattform, die

umgebungsunabhängig bereitgestellt werden kann. Diese Umwandlung wichtiger Betriebsdaten in ein KI-fähiges Asset kann die Effizienz um bis zu 30 % steigern und die Verwendung vertraulicher Daten, einschließlich unstrukturierter Daten, vereinfachen, um Modellausgaben in der Wissensdatenbank eines Unternehmens zu verankern.

Elastic

Elastic Search AI Platform (auf dem ELK Stack[2] basierend) kombiniert die Präzision einer Suche mit der Intelligenz von KI, sodass Nutzende Prototypen schneller erstellen und mit LLMs integrieren und gen KI zur Erstellung skalierbarer, kosteneffizienter Anwendungen einsetzen können. Mit dieser Plattform können Nutzende transformative RAG-Anwendungen (Retrieval Augmented Generation) erstellen, proaktiv Beobachtbarkeitsprobleme lösen und komplexe Sicherheitsbedrohungen beseitigen. Elasticsearch kann dort eingesetzt werden, wo sich Ihre Anwendungen befinden: lokal, bei dem von Ihnen gewählten Cloud-Anbieter oder in Air Gapped-Umgebungen.

Elastic lässt sich mit Einbettungsmodellen aus dem IT-Ökosystem wie Red Hat OpenShift AI, Hugging Face, Cohere, OpenAI und anderen über einen einzigen, unkomplizierten API-Aufruf integrieren. Dieser Ansatz gewährleistet einen sauberen Code zur Verwaltung hybrider Inferenzen für RAG-Workloads und bietet unter anderem folgende Funktionen:

- Chunking, Connectors und Webcrawler für die Aufnahme verschiedener Datensätze in Ihre Suchschicht
- Semantische Suche mit ELSER (Elastic Learned Sparse EncodeR), dem integrierten ML-Modell und dem E5-Einbettungsmodell, das eine mehrsprachige Vektorsuche ermöglicht
- Sicherheit auf Dokumenten- und Feldebene, Implementierung von Zugriffsrechten, die der Role-based Access Control (RBAC) Ihres Unternehmens entsprechen

Mit Elastic Search AI Platform sind Sie Teil einer weltweiten Community von Entwicklerinnen und Entwicklern, in der Inspiration und Support stets in Reichweite sind. Besuchen Sie die Elastic Community auf Slack, in unseren Diskussionsforen oder in den sozialen Medien.

Fazit

Mit Red Hat OpenShift AI können Unternehmen experimentieren, zusammenarbeiten und letztendlich die Entwicklung KI-gestützter Anwendungen beschleunigen. Data Scientists und KI-Engineers erhalten die Flexibilität, mit Red Hat OpenShift AI Modelle in der Hybrid Cloud zu entwickeln und bereitzustellen. IT-Operations- und Platform Engineers profitieren von MLOps- und GenAIOps-Funktionen, wodurch Modelle schneller in der Produktion eingesetzt werden können. Self Service für Entwicklungsteams, KI-Engineers und Data Scientists, einschließlich Zugriff auf GPUs, fördert Innovationen auf einer Anwendungsplattform, die bereits von Unternehmens-IT genutzt wird und deren Zuverlässigkeit uneingeschränkt anerkannt ist. Red Hat OpenShift AI stellt weiterhin eine umfassende, vertrauenswürdige und konsistente Plattform bereit, die einzigartige Unterscheidungsmerkmale für effiziente Inferenz, agentische KI und skalierbare Hybrid Cloud-Abläufe bietet – unterstützt von einem robusten Partnernetzwerk.

Weitere Informationen

Machen Sie noch heute den ersten Schritt und besuchen Sie uns unter Red Hat OpenShift AI.

Über Red Hat

Red Hat unterstützt Kunden dabei, ihre Umgebungen zu standardisieren, cloudnative Anwendungen zu entwickeln und komplexe Umgebungen mit vielfach ausgezeichnetem Support, Training und Consulting Services zu integrieren, zu automatisieren, zu sichern und zu verwalten.

facebook.com/redhatinc
@RedHatDACH
linkedin.com/company/red-hat

de.redhat.com
Nr. 2876428_1125

2 *DerELK-Stack besteht aus Elasticsearch, Kibana, Beats und Logstash.*

EUROPA, NAHOST, UND AFRIKA (EMEA)
00800 7334 2835
de.redhat.com
europe@redhat.com

TÜRKEI
00800 448820640

ISRAEL
1 809 449548

VAE
8000-4449549