Beschleunigte Einführung von KI für Finanzdienstleistungen mit Red Hat

Verkürzung der Markteinführungszeit für KI/ML-Lösungen mit einer End-to-End-Plattform

Komplexere KI-Modelle führen zu größeren Herausforderungen bei der Einführung

Finanzinstitute möchten die Chancen nutzen, die sich durch die Einführung von künstlicher Intelligenz (KI) ergeben. Das rasante Entwicklungstempo in Bereichen wie Deep Learning, dialogorientierter und generativer KI hat den Umfang und die Anwendbarkeit von KI-Lösungen enorm gesteigert. Gleichzeitig ergeben sich durch die zunehmende Komplexität der Modelle neue Herausforderungen bei der Ausführung, und bestehende Herausforderungen werden verstärkt. Zu diesen Herausforderungen zählen:

- **Standalone-Entwicklungsprozess:** Der Großteil der KI- und ML-Entwicklungs- und -Trainingsprozesse findet derzeit in dedizierten Umgebungen statt und erfordert spezielle Ressourcen, beispielsweise schnellere Hardware wie GPUs (Graphic Processing Units). Die Provisionierung von KI/ML-Umgebungen nimmt viel Zeit in Anspruch und behindert die Einführung neuer KI-basierter Services erheblich.
- **Skalierung, Flexibilität und Ressourcenoptimierung:** KI/ML-Lösungen erfordern Komponenten mit unterschiedlichem Ressourcenbedarf, beispielsweise die CPU (Central Processing Unit), Speicher, Disk und spezielle Hardware (GPU), TPU (Tensor Processing Unit) und FPGA (Field-Programmable Gate Array). Die Skalierung solcher Lösungen erfordert häufig einen Hybrid Cloud-Ansatz.
- **Monitoring und Drift:** KI/ML-Modelle müssen kontinuierlich überwacht und regelmäßig aktualisiert werden, um Drift zu erkennen und zu korrigieren. Red Hat® OpenShift® erleichtert die kontinuierliche Integration von Modell-Updates durch das Bereitstellen einer standardbasierten Überwachungsinfrastruktur, die anwendungsbasiertes Drift Monitoring mit der KI/ML-Entwicklungspipeline verbinden kann.
- **Sicherheit der Modelllieferkette:** Das Ökosystem der KI/ML-Entwicklungstools basiert weitgehend auf quelloffenen, von der Community betriebenen Frameworks. Die Qualitätssicherung der Softwarelieferkette stellt in dieser Umgebung eine wachsende Herausforderung dar. Entwicklungsteams wünschen sich die neuesten Tools, aber Unternehmen müssen sicherstellen, dass diese Tools sicher und sicherheitsoptimiert sind sowie keinerlei anfälligen oder bösartigen Artefakte enthalten.

Vorteile, die die Komplexität deutlich reduzieren

Die vorgeschlagene KI/ML-Lösung bietet Finanzinstituten unter anderem folgende Vorteile:

- Eine End-to-End-Plattform für Modellentwicklung, Training und Inferenz. Dies sorgt für konsistente Operationen in Public und Private Clouds und reduziert die Reibungspunkte zwischen den verschiedenen Phasen des Prozesses.
- Self Service-Funktionen beschleunigen die Wertschöpfung in ML-Umgebungen.
- Konsistente quelloffene ML-Tools und -Libraries auf dem neuesten Stand der Technik, zusammen mit einem umfangreichen IT-Ökosystem von Open Source- und von Partnern unterstützten Technologien.
- Schnelle Entwicklung und Bereitstellung von ML-Modellen, zusammen mit Monitoring- und schnellen Iterationsfunktionen, die für die Aktualität der bereitgestellten Modelle sorgen.

Case Study: Large Language Models (LLM)

Ein Beispiel, das die Herausforderungen und Vorteile im Bereich der Finanzdienstleistungen veranschaulicht, ist das Implementieren einer LLM-basierten Lösung, wie GPT-4, BLOOM, BART, DOLLY und andere. Diese Lösungen werden für die Digitalisierung von Dokumenten im Rahmen von Onboarding- oder KYC-Prozessen (Know Your Customer), für die Analyse von ESG-Berichten (Environment, Social, Governance) oder für das Implementieren von dialogorientierten Lösungen (z. B. Chatbots) verwendet.

Bei diesen Lösungen werden in der Regel große ML-Modelle mit mehreren Millionen oder Milliarden von Parametern verwendet. Aufgrund des Aufwands, der Komplexität und der erforderlichen Rechenleistung für das Erstellen dieser Modelle ist es üblich, auf vortrainierten oder Basismodellen aufzubauen. Da diese Modelle in der Regel auf allgemeinen Datensätzen trainiert werden, erfordert ihre Anwendung auf den spezifischen Kontext eines Use Case im Finanzdienstleistungsbereich ein zusätzliches domain- oder firmenspezifisches Training auf einem kleineren Satz lokaler Daten durch Feinabstimmung oder Transfer Learning. Ein Beispiel für die Architektur einer derartigen Lösung ist in Abbildung 1 dargestellt.

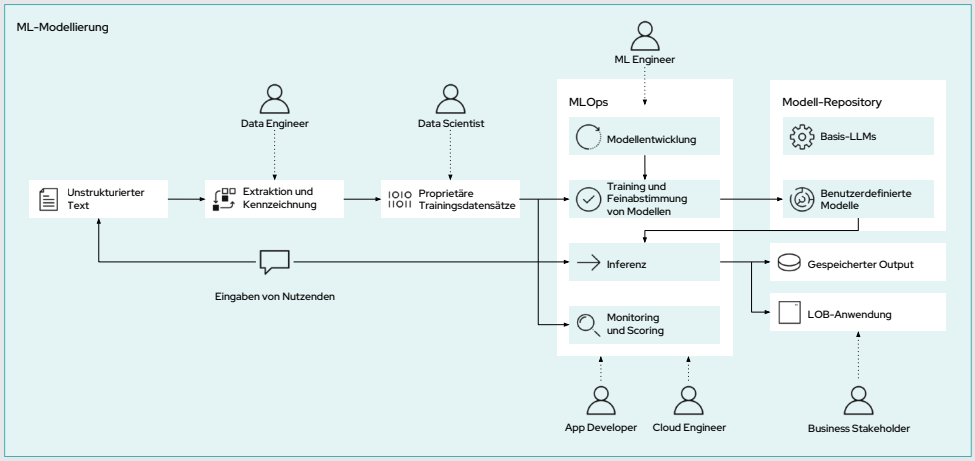

Abbildung 1. Überblick über die ML-Modellierung

Überblick über Funktionen

Lösungsarchitektur

Red Hat bietet eine Plattform, die den gesamten KI/ML-Lifecycle von der Entwicklung über das Training bis hin zur Inferenz effizient und produktiv hostet. Die Plattformtechnologie von Red Hat lässt sich auf den gängigen Infrastrukturen ausführen, von Bare Metal- und On-Premise-Virtualisierung bis hin zu großen Public Clouds. Das bedeutet, dass dieselbe Plattform mit denselben Tools und denselben MLOps-Prozessen verwendet wird.

Red Hat ist sich der Bedeutung von Open Source und des Schutzes der Softwarelieferkette bewusst und engagiert sich in Upstream Communities, um erstklassige neue Software und zuverlässige Partnerschaften zu entwickeln. In diesem Zusammenhang kuratiert, unterstützt und zertifiziert Red Hat eine große Anzahl von Upstream-Tools, die KI/ML-Entwicklungsteams benötigen. Red Hat kümmert sich um die Upstream-Lieferkette und stellt Ihrem Unternehmen ein Produkt zur Verfügung, auf das Sie sich verlassen können – mit Support rund um die Uhr.

Plattformkomponenten

Betriebssystem

Die KI/ML-Architektur von Red Hat basiert auf Red Hat Enterprise Linux®, einem Betriebssystem, das auf einer modernen Deployment-Infrastruktur On-Premise oder in einer Cloud-Umgebung, auf Bare Metal oder auf virtuellen Maschinen ausgeführt werden kann. Red Hat Enterprise Linux ist für die Ausführung auf einem umfangreichen Partnernetzwerk aus Hardware- und großen Cloud-Anbietern zertifiziert, darunter Amazon Web Service (AWS), Google Cloud, IBM Cloud for Financial Services, Oracle Cloud und Microsoft Azure. Die Linux-Plattform bietet Sicherheit, Performance, Support und erstklassige Automatisierung mit Red Hat Ansible® Automation Platform. Schließlich bietet Red Hat Enterprise Linux Unterstützung für spezielle Hardware für die KI/ML-Modellentwicklung, einschließlich GPUs und FPGAs.

Container-Orchestrierung

Neben benutzerdefinierten und kommerziell erhältlichen Anwendungen ist die überwiegende Mehrheit der in KI/ML-Prozessen verwendeten Open Source-Tools und -Libraries containerisiert. Vortrainierte oder produktive ML-Modelle werden auch als Container Images paketiert. Darüber hinaus umfassen KI/ML-Prozesse mehrere Komponenten, die miteinander interagieren und dynamisch skaliert werden müssen. Zu diesen Komponenten gehören das rechenintensive Training neuer Modelle, Inferenz-Engines mit hohem Durchsatz und Modellentwicklungsumgebungen, die von Data Scientists genutzt werden. Dafür ist eine flexible und anpassungsfähige Plattform erforderlich. Red Hat OpenShift, eine Kubernetes-Distribution, ist branchenführend beim Deployment und der Orchestrierung von containerisierten Workloads. Sie ist die gängigste Plattform für Drittanbieter- und Open Source-KI-Entwicklungstools. Somit können Ihre Entwicklungsteams auf die gewünschten KI/ML-Frameworks zugreifen, um die Wertschöpfung zu beschleunigen. Red Hat OpenShift bietet auch Operator-Technologien zur Automatisierung des Deployments von Komponenten für Self Service und reduzierte Betriebskosten.

Skalierbarer, sicherheitsgehärteter Storage

KI/ML-Projekte erfordern große Mengen an Trainingsdaten, um präzise Modelle zu erstellen. Dabei kann es sich um historische Daten oder um Live-Daten aus Quellen wie Marktdaten-Feeds, Internet of Things (IoT) und Beobachtbarkeit handeln. Die Daten müssen stets so gespeichert werden, dass sie benutzerfreundlich und für Entwicklerinnen und Entwickler immer wieder zugänglich sind. Red Hat unterstützt und integriert quelloffenen softwaredefinierten Storage in Form von Red Hat OpenShift Data Foundation, basierend auf Red Hat Ceph® Storage. OpenShift Data Foundation ist eine softwaredefinierte Storage-Lösung, die sich in Red Hat OpenShift integrieren lässt und wirtschaftlich auf Petabyte und mehr skaliert werden kann. Streaming-Daten können mit AMQ Streams, die auf Apache Kafka basieren, konsumiert werden, um Entwicklungsteams einen wiederholbaren Zugriff auf Streaming-Daten zu ermöglichen. Sowohl OpenShift Data Foundation als auch AMQ Streams, die in Containern paketiert sind, können mit Red Hat OpenShift verwaltet werden, sodass mehrere Entwicklungsteams auf Self Service-Basis arbeiten können.

Plattformfunktionen

Self Service

Mit Red Hat OpenShift können Entwicklungsteams und Projekte bei Bedarf eingebunden und die Ressourcen je nach Bedarf vertikal oder horizontal skaliert werden. Darüber hinaus kann kostspielige spezielle Hardware, wie beispielsweise GPUs, gepoolt und gemeinsam genutzt werden. Die Einhaltung von Sicherheitsvorschriften und die Sicherheit der Softwarelieferkette sind auf den verschiedenen Ebenen integriert.

Erweiterte Überwachung und Beobachtbarkeit

Red Hat OpenShift beinhaltet Überwachung nach Open Source-Industriestandard durch Prometheus und bietet Kompatibilität mit Monitoring-Tools von Drittanbietern, wie etwa Splunk. Dies ermöglicht die Integration von MLOps-Pipelines mit einer flexiblen, zentralisierten Infrastruktur für die Überwachung und zur Alarmierung in der gesamten Pipeline. Die Verfolgung der Modell-Performance kann die Skalierung automatisieren und Warnungen bei schwachen Präzisionswerten ausgeben.

Agilität

Die KI/ML-Modellierung ist ein iterativer Prozess. Data Engineers und Data Scientists erkunden die Pfade, die von den Daten hinterlassen werden, und der Modellentwicklungsprozess beinhaltet zahlreiche Starts und Stopps, unvorhergesehene Pfade, Hindernisse und Sackgassen. Zu den typischen Herausforderungen gehören der Zugang zu Qualitätsdaten aus verschiedenen Quellen wie Datenbanken, Dateisystemen, Streams, APIs (Application Programming Interfaces) sowie die Einhaltung von gesetzlichen Vorschriften und Sicherheitsstandards. Im Hinblick auf die Tools gehören das Versionieren umfangreicher Libraries sowie das Aktualisieren bestehender und das Einführen neuer Tools zu den Herausforderungen. Red Hat vereinfacht die KI/ML-Pipeline für professionelle Nutzende, indem es ihnen ein konsistentes Erlebnis in einer Hybrid Cloud-Umgebung bietet, um KI/ML-Projekte zu beschleunigen.

Ein Unterschied zwischen der herkömmlichen Anwendungsentwicklung und der Entwicklung von KI/ML-Anwendungen besteht darin, dass die Anwendungen selbst oder die den Anwendungen zugrunde liegenden KI-Modelle aktualisiert werden müssen. KI/ML-Techniken ermöglichen nicht nur das anfängliche Training eines Modells durch ML, sondern auch das kontinuierliche Aktualisieren des Modells. So bieten die Modelle Vorteile, die herkömmliche Anwendungen nicht bieten können. Allerdings müssen die Modelle in regelmäßigen Abständen aktualisiert werden, um die Performance zu verbessern. Red Hat OpenShift bietet Anwendungsteams die Möglichkeit, Komponenten der MLOps-Toolchain transparent vertikal und horizontal zu skalieren. Wenn für Ihre Anwendung eine Aktualisierung des Modells erforderlich ist, können die (teureren) Trainingsressourcen mit GPUs oder anderen speziellen Baugruppen manuell zugewiesen und erweitert werden. Wenn die Aktualisierung abgeschlossen ist, kann Red Hat OpenShift diese Ressourcen wieder den Bereichen zuweisen, in denen sie am dringendsten benötigt werden.

Skalierbarkeit und Flexibilität für Training und Inferenz

Die Trainingsphase der KI/ML-Modellierung ist eine der ressourcenintensivsten Operationen in der MLOps-Pipeline. Hier befinden sich die größten horizontal skalierbaren Instanzen von KI/ML-Tools, und hier ist die Nachfrage nach spezieller Hardware – wie GPUs, TPUs und FPGAs – von Unternehmen wie Nvidia am größten. Die einzelnen Projekte und Teams möchten Zugang zu ihren eigenen Umgebungen haben, um ihr Training durchzuführen. Die Fähigkeit der KI/ML-Architektur von Red Hat, eine gemeinsam genutzte Infrastruktur bereitzustellen, bietet erhebliche Vorteile hinsichtlich Effizienz und Wirtschaftlichkeit. Anstatt dedizierte Ressourcen zu hohen Kosten anzusammeln, bietet Red Hat OpenShift Entwicklungsteams einen virtuellen On-Demand-Zugang zum gesamten Cluster. Kubernetes orchestriert und vermittelt diesen Zugriff, um sicherzustellen, dass diese Ressourcen an den Orten und zu dem Zeitpunkt bereitgestellt werden, an dem das Unternehmen sie benötigt.

Offenes IT-Ökosystem

Die KI/ML-Plattform von Red Hat ist, wie die anderen Red Hat Produkte, vollständig quelloffen. Das Open Source-Ökosystem von Tools und Technologien, die KI/ML-Profis zur Verfügung stehen, umfasst Folgendes:

- ML-Libraries
- KI/ML-Lifecycle-Management
- Datenzugang, Datenqualität und Metadatenmanagement
- Erkennung von Bias und Erklärbarkeit
- Vortrainierte Modelle

Aufgrund des offenen Charakters des IT-Ökosystems und der Flexibilität der Plattform können diese Tools in verschiedenen Kombinationen verwendet werden, je nach den Anforderungen der jeweiligen Lösung. Zudem unterstützt eine offene Plattform die kontinuierliche Innovation, da neue Technologien, Tools und Modelle kontinuierlich in die Lösung integriert werden können.

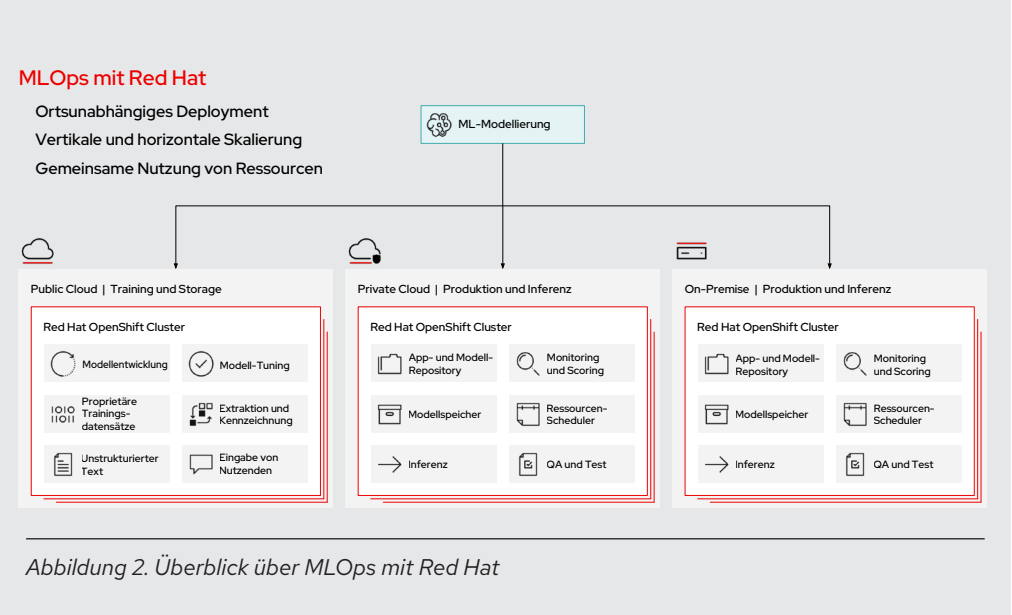

Abbildung 2. Überblick über MLOps mit Red Hat

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

TÜRKEI
00800 448820640

ISRAEL
1 809 449548

VAE
8000-4449549

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

de.redhat.com
#420450_0723