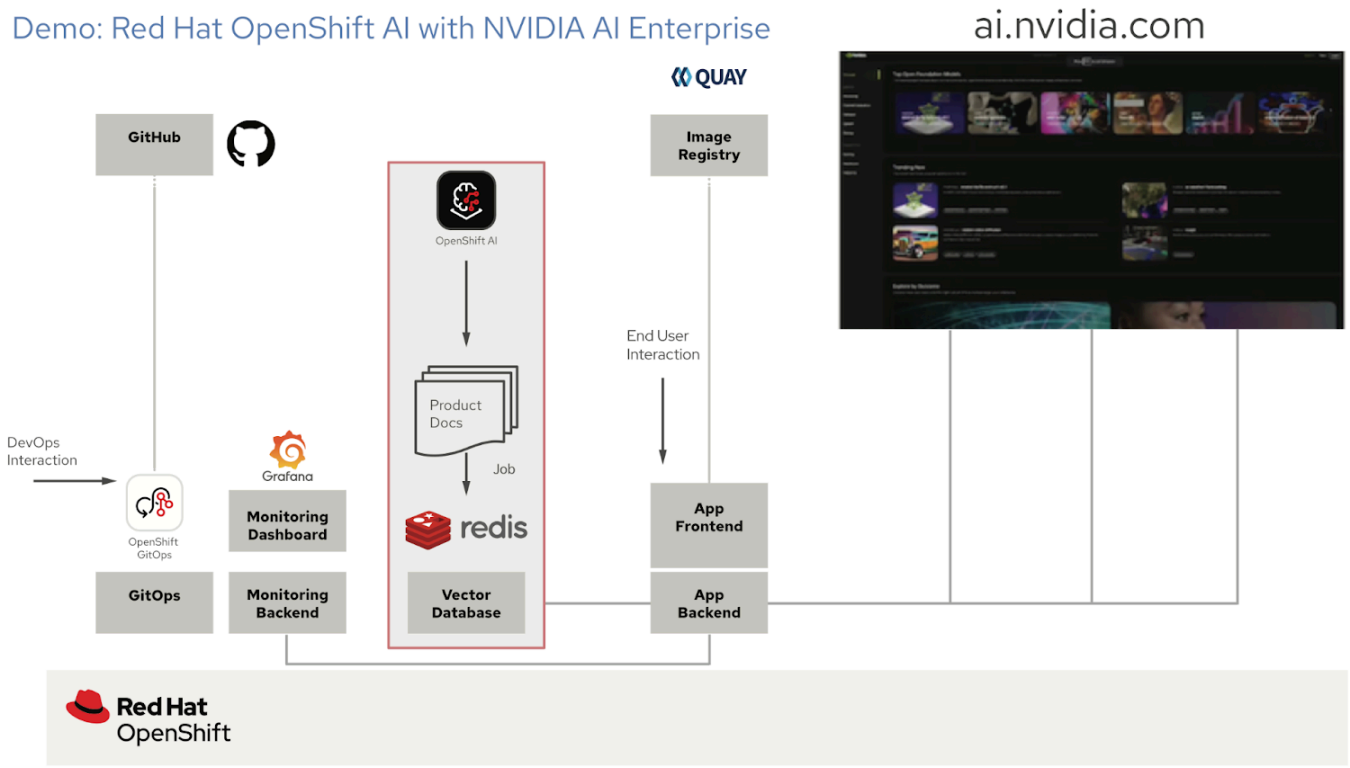

Lösungs-Pattern

KI-Anwendungen mit Red Hat und NVIDIA AI Enterprise

Entwickeln einer RAG-Anwendung

Red Hat OpenShift AI ist eine Plattform für das Entwickeln von Data Science-Projekten und Bereitstellen von KI-gestützten Anwendungen. Sie können sämtliche für die Unterstützung von Retrieval-Augmented Generation (RAG), einer Methode zum Abrufen von KI-Antworten aus Ihren eigenen Referenzdokumenten, erforderlichen Tools integrieren. Wenn Sie OpenShift AI mit NVIDIA AI Enterprise kombinieren, können Sie mit Large Language Models (LLMs) experimentieren und so das optimale Modell für Ihre Anwendung finden.

Erstellen einer Pipeline für Dokumente

Damit Sie RAG nutzen können, müssen Sie Ihre Dokumente zunächst in eine Vektordatenbank aufnehmen. In unserer Beispielanwendung integrieren wir eine Anzahl von Produktdokumenten in eine Redis-Datenbank. Da sich diese Dokumente häufig ändern, können wir für diesen Prozess eine Pipeline erstellen, die wir regelmäßig ausführen, damit wir immer die aktuellsten Versionen der Dokumente zur Verfügung haben.

Durchsuchen des LLM-Katalogs

Mit NVIDIA AI Enterprise können Sie auf einen Katalog verschiedener LLMs zugreifen. So können Sie verschiedene Möglichkeiten ausprobieren und das Modell auswählen, das die optimalen Ergebnisse erzielt. Die Modelle werden im NVIDIA API-Katalog gehostet. Sobald Sie ein API-Token eingerichtet haben, können Sie ein Modell mit der NVIDIA NIM Model Serving-Plattform direkt über OpenShift AI bereitstellen.

Auswählen des richtigen Modells

Beim Testen verschiedener LLMs können Ihre Nutzerinnen und Nutzer die einzelnen generierten Antworten bewerten. Sie können ein Grafana Monitoring Dashboard einrichten, um die Bewertungen sowie die Latenz- und Antwortzeiten der einzelnen Modelle zu vergleichen. Anhand dieser Daten können Sie dann das optimale LLM für den Produktionseinsatz auswählen.