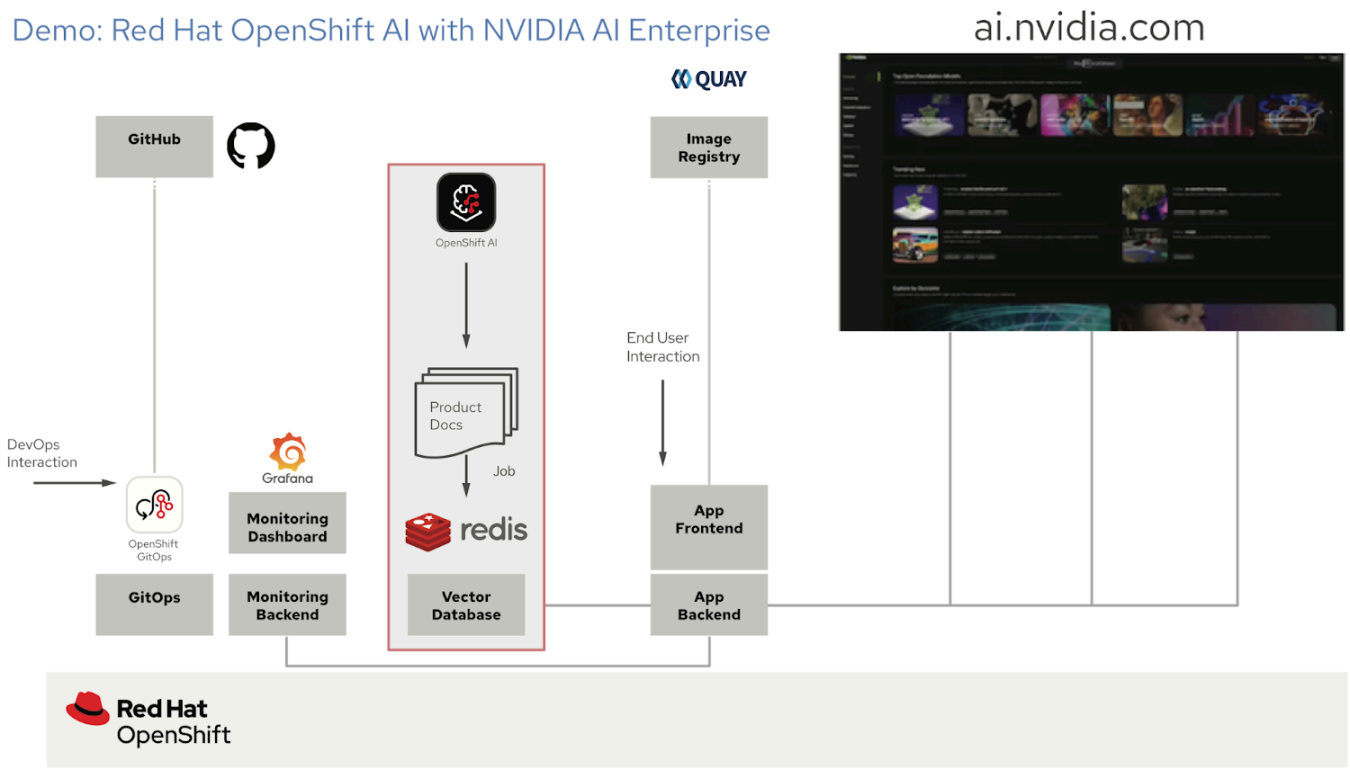

Estrutura da solução

Aplicações de IA com a Red Hat e o NVIDIA AI Enterprise

Crie uma aplicação usando o método RAG

Red Hat OpenShift AI é uma plataforma para desenvolver projetos de ciência de dados e oferecer aplicações habilitadas para IA. Você pode integrar todas as ferramentas necessárias para suporte ao método **retrieval-augmented generation (RAG)**, que permite obter respostas de IA a partir dos seus próprios documentos de referência. Ao conectar o **OpenShift AI** com o **NVIDIA AI Enterprise**, você pode experimentar modelos de linguagem de grande escala (LLMs) para encontrar o modelo ideal para sua aplicação.

Crie um pipeline de documentos

Para usar o método RAG, primeiro você precisa inserir seus documentos em um banco de dados vetorial. Na nossa app de exemplo, inserimos um conjunto de documentos da solução em um banco de dados Redis. Como esses documentos mudam com frequência, criamos um pipeline para esse processo, que executaremos periodicamente. Dessa forma, sempre teremos as versões mais recentes dos documentos.

Confira o catálogo de LLM

O NVIDIA AI Enterprise oferece um acesso a um catálogo de diferentes LLMs para que você tenha diversas opções e selecione o modelo com os melhores resultados. Os modelos são hospedados no catálogo da API da NVIDIA. Após configurar um token de API, é possível implantar um modelo usando o NVIDIA NIM que disponibiliza a plataforma diretamente a partir do OpenShift AI.

Como escolher o modelo ideal

Conforme você testa diferentes LLMs, seus usuários podem avaliar cada resposta gerada. Você pode configurar um dashboard de monitoramento do Grafana para comparar as avaliações e o tempo de resposta e latência de cada modelo. Depois, use esses dados para escolher o melhor LLM para a produção.