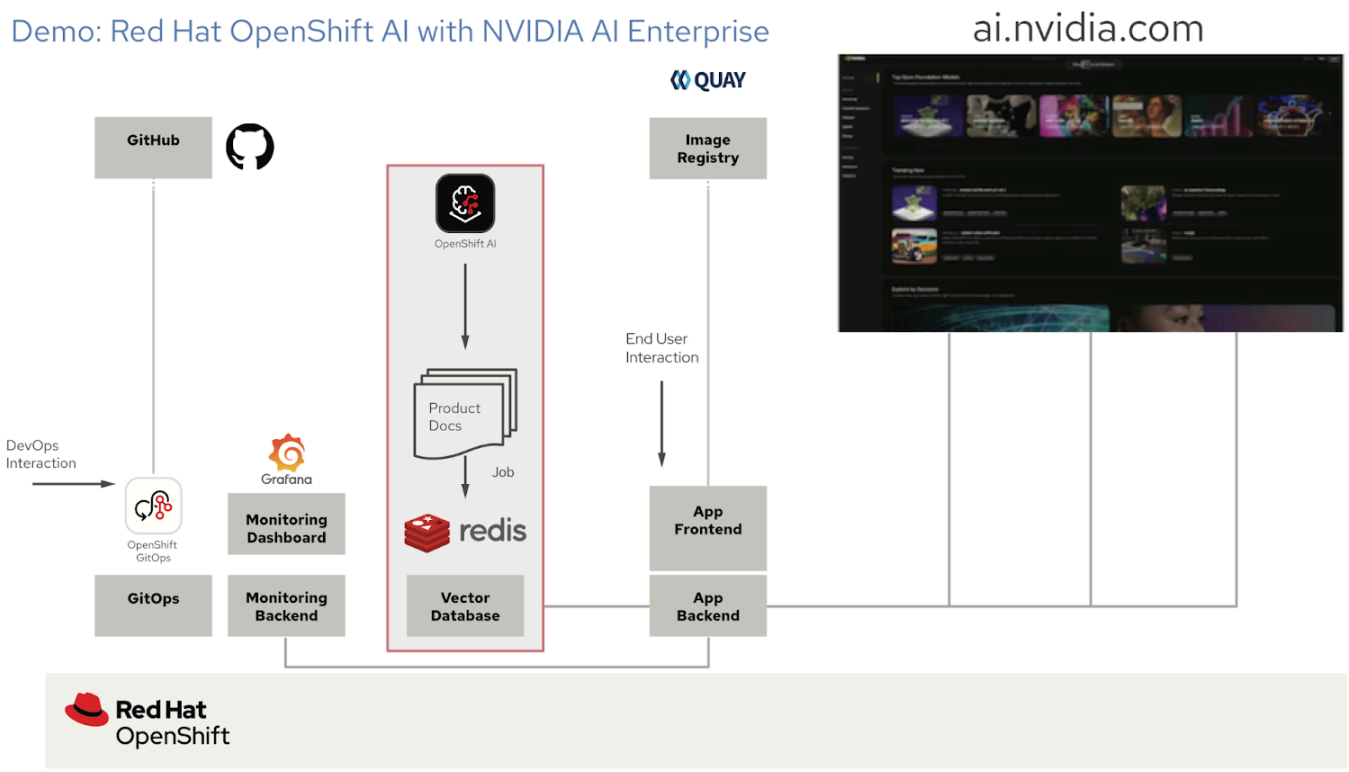

Modèle de solution

Applications d'IA avec Red Hat et NVIDIA AI Enterprise

Créer une application RAG

Red Hat OpenShift AI est une plateforme qui sert à réaliser des projets de science des données et à servir des applications basées sur l'IA. Vous pouvez intégrer tous les outils dont vous avez besoin pour prendre en charge la génération augmentée de récupération (RAG), un moyen d'obtenir des réponses d'une IA basées sur vos propres documents de référence. L'association d'OpenShift AI à NVIDIA AI Enterprise vous permet d'utiliser des grands modèles de langage (LLM) afin de trouver le modèle optimal pour votre application.

Concevoir un pipeline pour les documents

Pour tirer parti de la RAG, il est nécessaire, dans un premier temps, d'ajouter des documents dans une base de données vectorielle. Dans notre exemple d'application, nous intégrons un ensemble de documents relatifs à des produits dans une base de données Redis. Puisque ces documents changent fréquemment, nous avons créé un pipeline pour ce processus que nous exécuterons régulièrement, afin de nous assurer que l'IA dispose toujours des dernières versions des documents.

Parcourir le catalogue de LLM

NVIDIA AI Enterprise donne accès à un catalogue varié de LLM. Il est donc possible de tester plusieurs modèles et de sélectionner celui qui offre les meilleurs résultats. Les modèles sont hébergés dans le catalogue d'API de NVIDIA. Une fois le jeton textuel API configuré, un modèle peut être déployé directement à partir d'OpenShift AI, en utilisant la plateforme de service de modèles NVIDIA NIM.

Choisir le modèle le plus adapté

Lors du test de différents LLM, les utilisateurs peuvent noter chaque réponse générée. Il est possible de configurer un tableau de bord de surveillance Grafana pour comparer les notes ainsi que la latence et le temps de réponse pour chaque modèle. Ensuite, ces données peuvent être utilisées pour choisir le meilleur LLM à utiliser en production.