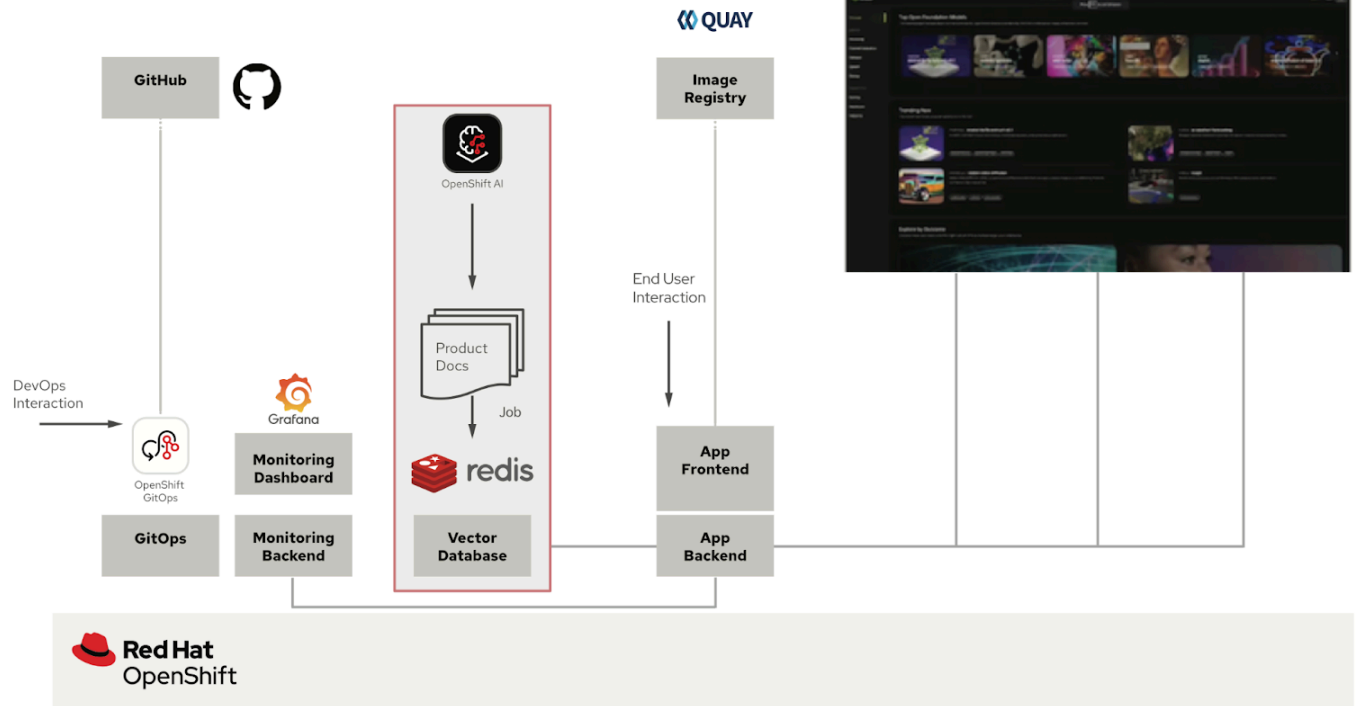


Demo: Red Hat OpenShift AI with NVIDIA AI Enterprise

ai.nvidia.com



ソリューションパターン

Red Hat と NVIDIA AI Enterprise による AI アプリケーション

RAG アプリケーションを作成する

Red Hat OpenShift AI は、データサイエンス・プロジェクトを構築し、AI 対応アプリケーションを提供するためのプラットフォームです。独自の参照ドキュメントから AI の回答を取得する手法の1つである 検索拡張生成 (RAG) をサポートするために必要なすべてのツールを統合できます。OpenShift AI を NVIDIA AI Enterprise に接続すると、大規模言語モデル (LLM) を試して、アプリケーションに最適なモデルを見つけることができます。

ドキュメントのパイプラインを構築する

RAG を利用するには、まずドキュメントをベクトルデータベースに取り込む必要があります。サンプルアプリでは、一連の製品ドキュメントを Redis データベースに埋め込んでいます。これらのドキュメントは頻繁に変更されるため、定期的に行うこのプロセス用のパイプラインを作成すると、常に最新バージョンのドキュメントを入手できます。

LLM カタログを閲覧する

NVIDIA AI Enterprise ではさまざまな LLM のカタログにアクセスできるため、さまざまな選択肢を試して、最良の結果が得られるモデルを選択できます。モデルは NVIDIA API カタログでホストされています。API トークンを設定したら、OpenShift AI から直接 NVIDIA NIM モデル提供プラットフォームを使用してモデルをデプロイできます。

適切なモデルを選択する

さまざまな LLM をテストする際、ユーザーは生成される応答をそれぞれ評価することができます。Grafana モニタリング・ダッシュボードを設定して、各モデルの評価だけでなく、レイテンシーと応答時間も比較できます。そのデータを使用して、プロダクションで使用する最適な LLM を選択できます。