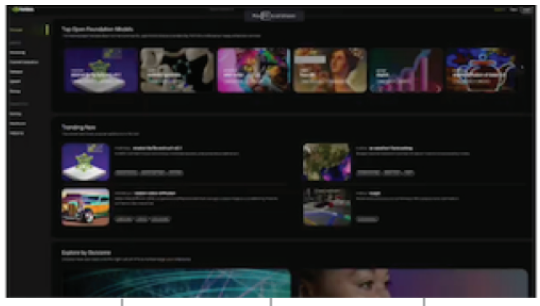
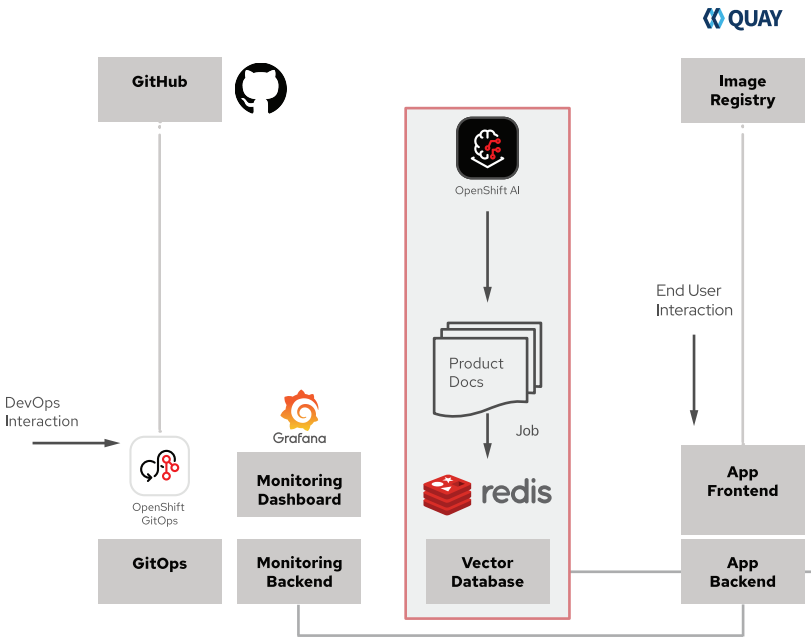
Demo: Red Hat OpenShift AI with NVIDIA AI Enterprise

ai.nvidia.com





SOLUTION PATTERN

AI applications with Red Hat and NVIDIA AI Enterprise

Create a RAG application

Red Hat OpenShift AI is a platform for building data science projects and serving AI-enabled applications. You can integrate all the tools you need to support retrieval-augmented generation (RAG), a method for getting AI answers from your own reference documents. When you connect OpenShift AI with NVIDIA AI Enterprise, you can experiment with large language models (LLMs) to find the optimal model for your application.

Build a pipeline for documents

To make use of RAG, you first need to ingest your documents into a vector database. In our example app, we embed a set of product documents in a Redis database. Since these documents change frequently, we can create a pipeline for this process that we'll run periodically, so we always have the latest versions of the documents.

Browse the LLM catalog

NVIDIA AI Enterprise gives you access to a catalog of different LLMs, so you can try different choices and select the model that delivers the best results. The models are hosted in the NVIDIA API catalog. Once you've set up an API token, you can deploy a model using the NVIDIA NIM model serving platform directly from OpenShift AI.

Choose the right model

As you test different LLMs, your users can rate each generated response. You can set up a Grafana monitoring dashboard to compare the ratings, as well as latency and response time for each model. Then you can use that data to choose the best LLM to use in production.