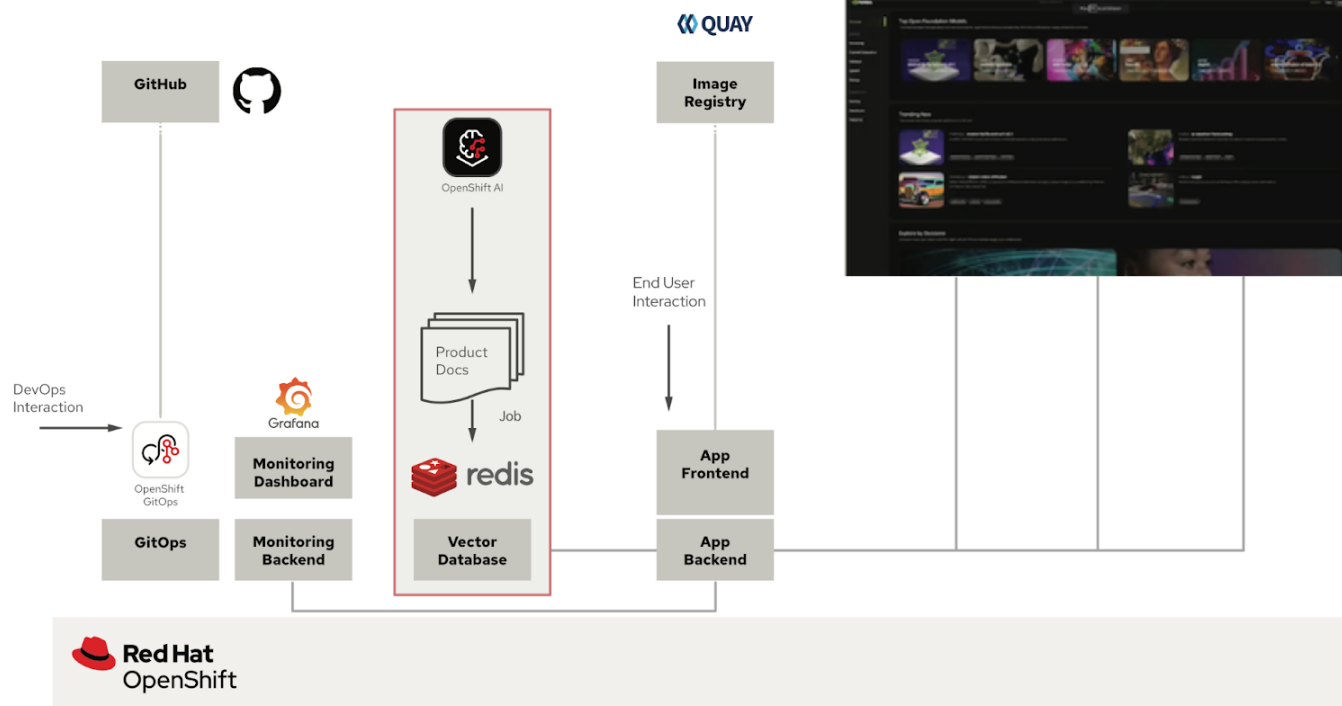


Demo: Red Hat OpenShift AI with NVIDIA AI Enterprise



解决方案模式

使用红帽技术和 NVIDIA AI Enterprise 构建的 AI 应用

创建 RAG 应用

红帽 OpenShift AI 是一个用于构建数据科学项目并为依托 AI 的应用提供服务的平台。您可以集成支持检索增强生成（RAG）所需的所有工具，这是一种从自己的参考文档中获取 AI 答案的方法。将 OpenShift AI 与 NVIDIA AI Enterprise 关联后，您可以尝试各种大语言模型（LLM），以找到适用于您应用的最佳模型。

构建文档处理管道

要使用 RAG，需要先将文档导入向量数据库。在示例应用中，我们将一组产品文档嵌入了 Redis 数据库。由于这些文档经常更改，我们可以为此过程创建一个管道并定期运行，以确保始终拥有最新版本的文档。

浏览 LLM 目录

NVIDIA AI Enterprise 支持访问不同 LLM 的目录，因此您可以尝试不同的选择，并选取能够提供最佳效果的模型。这些模型托管在 NVIDIA API 目录中。设置 API 令牌后，您就可以直接从 OpenShift AI 使用 NVIDIA NIM 模型服务平台来部署模型。

选择合适的模型

在测试不同的 LLM 时，您可以让用户对每个生成的响应进行评分。您还可以设置一个 Grafana 监控面板，用于比较各个模型的评分、延迟和响应时间。然后，您就可以根据这些数据来选择最适合在生产环境中使用的 LLM。