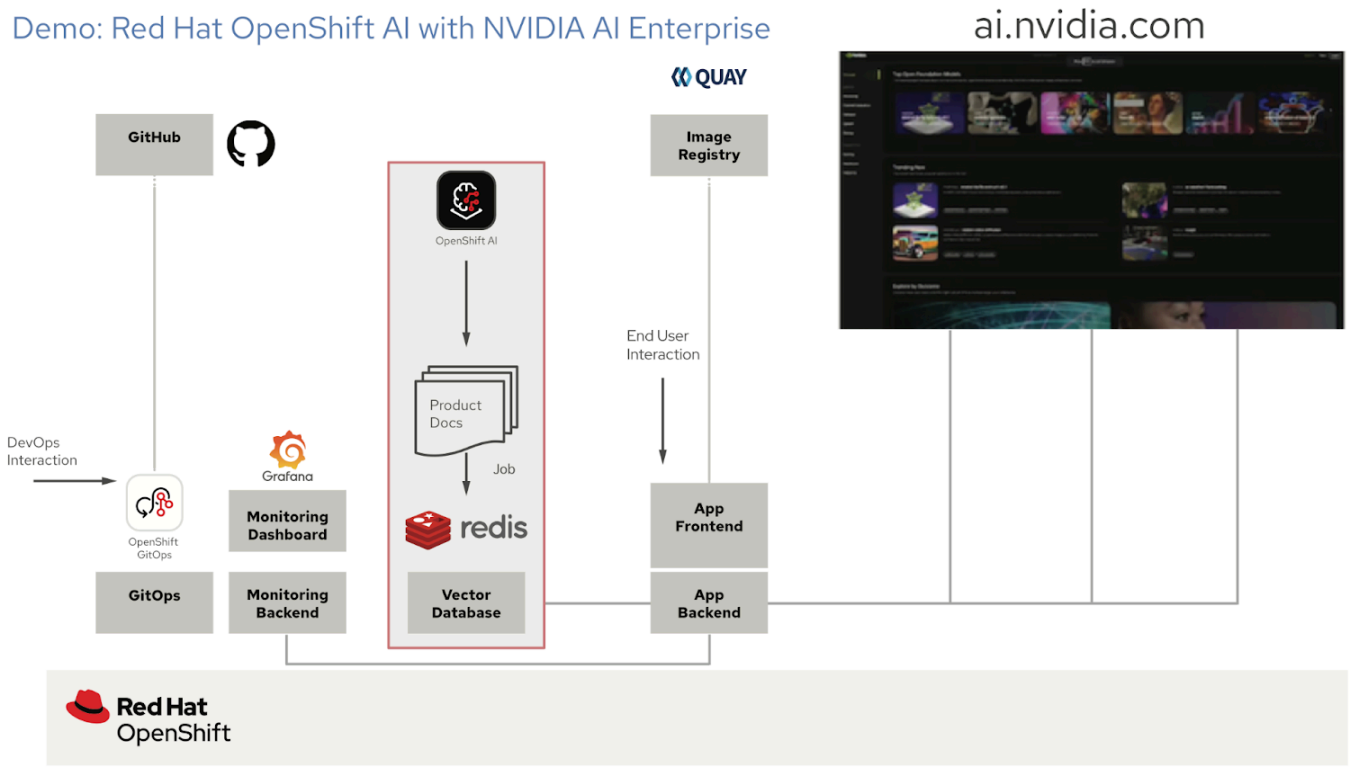

솔루션 패턴

Red Hat과 NVIDIA AI Enterprise를 활용한 AI 애플리케이션

RAG 애플리케이션 생성

Red Hat OpenShift AI는 데이터 사이언스 프로젝트를 빌드하고 AI 기능을 탑재한 애플리케이션을 제공하기 위한 플랫폼입니다. 자체 참조 문서에서 AI 답변을 얻는 방법인 검색 증강 생성(Retrieval-Augmented Generation, RAG)을 지원하는 데 필요한 모든 도구를 통합할 수 있습니다. OpenShift AI를 NVIDIA AI Enterprise와 연결하면 대규모 언어 모델(Large Language Model, LLM)을 실험하여 애플리케이션에 대한 최적의 모델을 찾을 수 있습니다.

문서 파이프라인 구축

RAG를 활용하려면 먼저 문서를 벡터 데이터베이스에 수집해야 합니다. 예시 애플리케이션에서는 제품 문서 세트를 Redis 데이터베이스에 임베드합니다. 이러한 문서는 자주 변경되기 때문에, 이 프로세스에 대한 파이프라인을 구축하여 주기적으로 실행하면 항상 최신 버전의 문서를 확보할 수 있습니다.

LLM 카탈로그 살펴보기

NVIDIA AI Enterprise에서 다양한 LLM 카탈로그에 액세스하여 여러 가지 옵션을 사용해 보고 그중에서 최상의 결과를 제공하는 모델을 선택할 수 있습니다. 모델은 NVIDIA API 카탈로그에 호스팅됩니다. API 토큰을 설정한 후에는 OpenShift AI에서 바로 NVIDIA NIM 모델 서빙 플랫폼을 사용하여 모델을 배포할 수 있습니다.

적합한 모델 선택

다양한 LLM을 테스트하면서 사용자는 생성된 각 응답을 평가할 수 있습니다. Grafana 모니터링 대시보드를 설정하면 각 모델의 평점, 대기 시간, 응답 시간을 비교할 수 있습니다. 그런 다음 해당 데이터를 기반으로 프로덕션에서 사용하기에 가장 적합한 LLM을 선택할 수 있습니다.