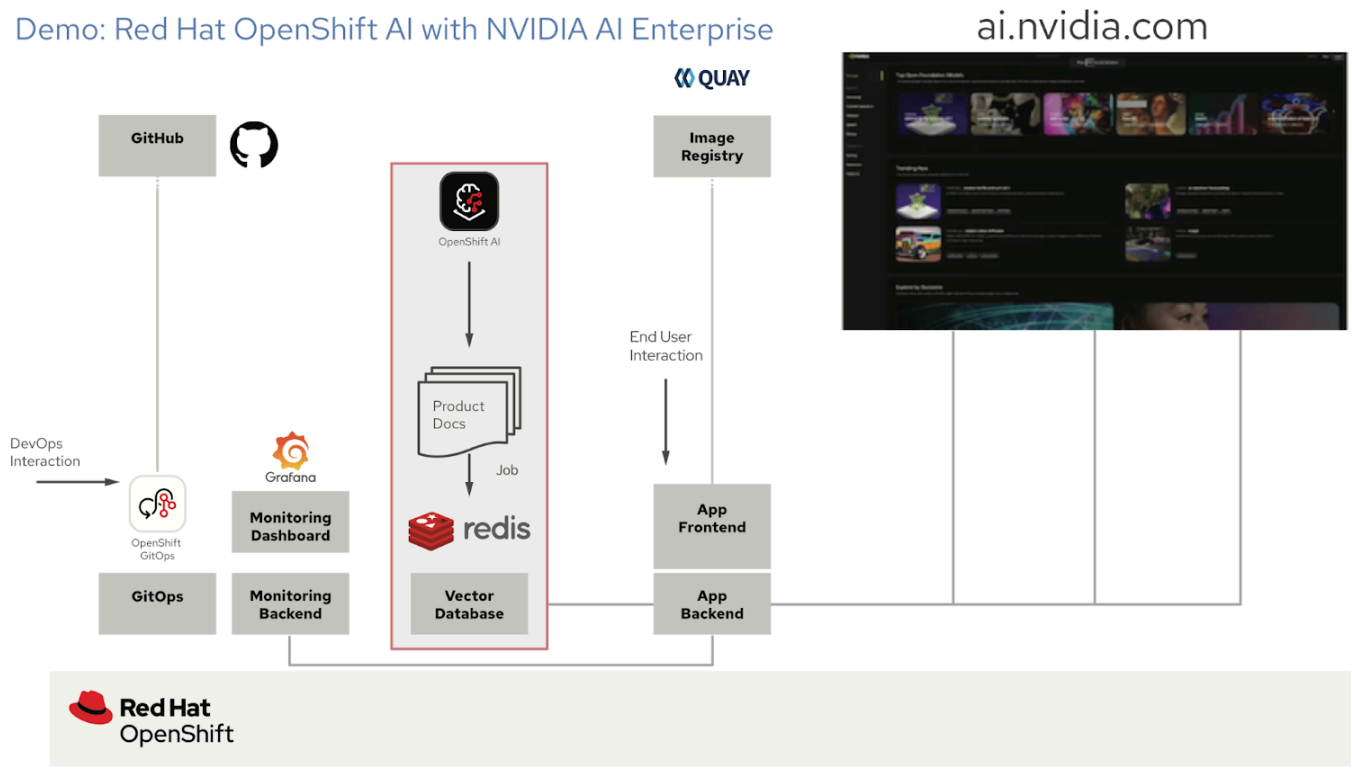

Analisi della soluzione

Applicazioni IA con Red Hat e NVIDIA AI Enterprise

Crea un'applicazione RAG

Red Hat OpenShift AI è una piattaforma per creare progetti di data science e distribuire applicazioni abilitate all'IA. Puoi integrare tutti gli strumenti che ti servono per supportare la tecnologia retrieval augmented generation (RAG), un metodo per ottenere le risposte dell'IA dai tuoi documenti di riferimento. Quando connetti OpenShift AI con NVIDIA AI Enterprise, puoi fare alcune prove con gli LLM (Large Language Model) per trovare il modello ideale per la tua applicazione.

Sviluppa una pipeline per i documenti

Se vuoi usare la tecnologia RAG, per prima cosa devi inserire i tuoi documenti in un database vettoriale. Nella nostra app di esempio, integriamo una serie di documenti di prodotto in un database Redis. Visto che questi documenti vengono modificati spesso, possiamo creare una pipeline da eseguire periodicamente per tale processo in modo da disporre sempre della versione più recente.

Sfoglia il catalogo degli LLM

NVIDIA AI Enterprise ti offre l'accesso a un catalogo di diversi modelli LLM, così potrai fare diverse prove e scegliere quello che offre risultati migliori. I modelli sono ospitati nel catalogo delle API di NVIDIA. Una volta che hai impostato un token API, puoi eseguire il deployment di un modello utilizzando la piattaforma di distribuzione dei modelli NVIDIA NIM direttamente da OpenShift AI.

Scegli il modello giusto

Mentre fai alcune prove con i diversi modelli LLM, i tuoi utenti possono valutare ogni risposta generata. Puoi configurare una dashboard di monitoraggio Grafana per confrontare le valutazioni, oltre che i tempi di risposta e di latenza di ogni modello. Potrai quindi sfruttare questi dati per scegliere il miglior modello LLM da usare nel tuo ambiente di produzione.