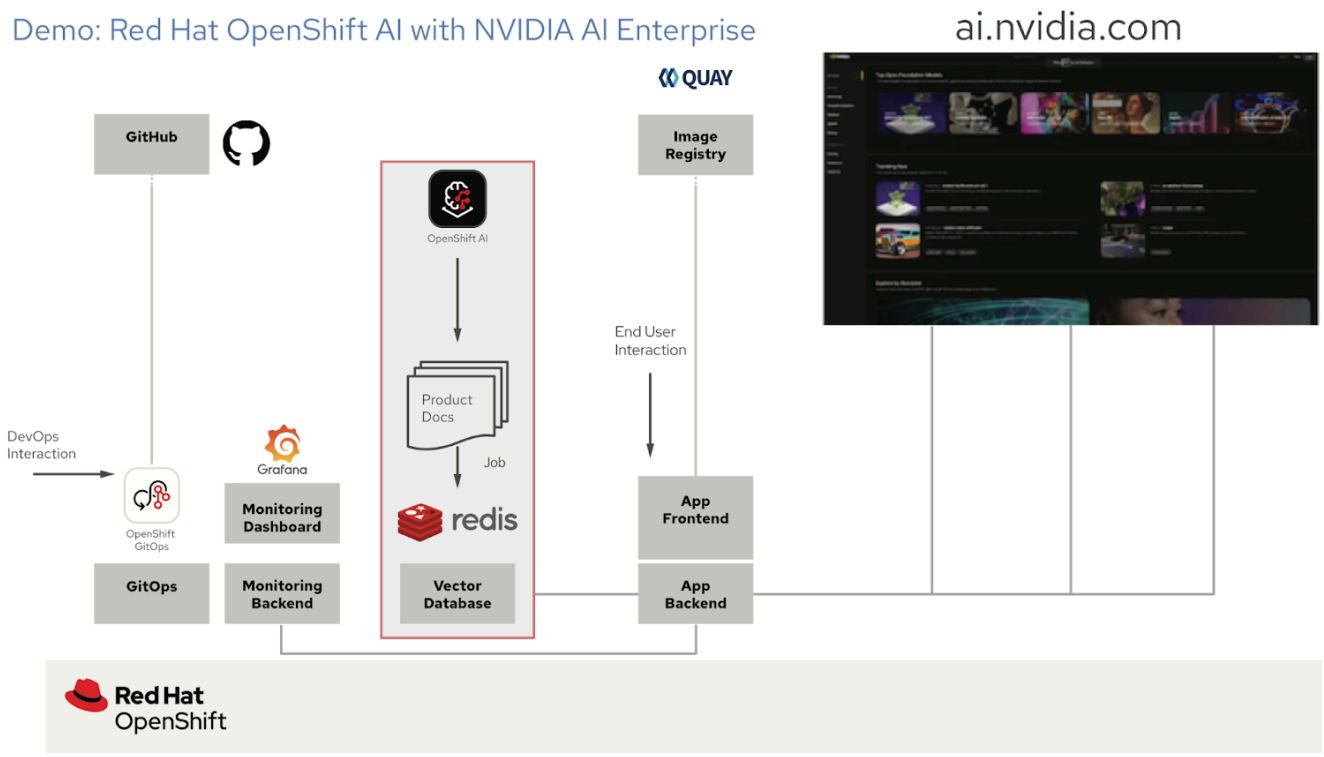

PATRÓN DE LA SOLUCIÓN

Aplicaciones de inteligencia artificial con NVIDIA AI Enterprise y Red Hat

Cree una aplicación RAG

Red Hat OpenShift AI es una plataforma para diseñar proyectos de análisis de datos y distribuir aplicaciones que utilizan la inteligencia artificial. Con ella puede integrar todas las herramientas que necesita para respaldar la generación aumentada por recuperación (RAG), un método para obtener respuestas de la inteligencia artificial desde sus documentos de referencia. Cuando conecta OpenShift AI con NVIDIA AI Enterprise, puede experimentar con los modelos de lenguaje de gran tamaño (LLM) para encontrar el más adecuado para su aplicación.

Diseñe un canal para los documentos

Para usar la RAG, primero debe incorporar los documentos en una base de datos vectorial. En nuestra aplicación de ejemplo, integramos un conjunto de documentos de productos en una base de datos de Redis. Ya que estos cambian con frecuencia, podemos crear un canal para este proceso que ejecutaremos de forma regular y, de esta forma, siempre tendremos las versiones más recientes de ellos.

Consulte el catálogo de LLM

NVIDIA AI Enterprise brinda acceso a un catálogo de distintos LLM, por lo que puede probar diferentes opciones y elegir el modelo que ofrezca los mejores resultados. Estos modelos se alojan en el catálogo de NVIDIA API. Una vez que haya configurado un token de interfaz de programación de aplicaciones (API), puede implementar un modelo utilizando la plataforma de distribución de NVIDIA NIM directamente desde OpenShift AI.

Seleccione el modelo adecuado

A medida que prueba distintos LLM, los usuarios tienen la posibilidad de calificar cada respuesta que se genera. Puede configurar un panel de control de Grafana para comparar las calificaciones, además de la latencia y los tiempos de respuesta de cada modelo. Luego, puede usar esos datos para elegir el mejor LLM para implementar en la producción.