Skillsoft
global knowledge™

DEPLOYING RAG PIPELINES FOR PRODUCTION AT SCALE

Course Code: 847005

Gain hands-on experience deploying, monitoring, and scaling RAG pipelines with the NIM Operator and learn best practices for infrastructure optimization, performance monitoring, and handling high traffic volumes.

This course begins by building a simple RAG pipeline using the NVIDIA API catalog. Participants will deploy and test individual components in a local environment using Docker Compose. Once familiar with the basics, the focus will shift to deploying NIMs, such as LLM, NeMo Retriever Text Embedding, and NeMo Retriever Text Reranking, in a Kubernetes cluster using the NIM Operator. This will include managing the deployment, monitoring, and scalability of NVIDIA NIM microservices. Building on these deployments, the workshop will cover constructing a production-grade RAG pipeline using the deployed NIMs and explore NVIDIA's blueprint for PDF ingestion, learning how to integrate it into the RAG pipeline.

To ensure operational efficiency, the workshop will introduce Prometheus and Grafana for monitoring pipeline performance, cluster health, and resource utilization. Scalability will be addressed through the use of the Kubernetes Horizontal Pod Autoscaler (HPA) for dynamically scaling NIMs based on custom metrics in conjunction with the NIM Operator. Custom dashboards will be created to visualize key metrics and interpret performance insights.

What You'll Learn

In service of teaching and demonstrating how to deploy enterprise-scale LLM-based agentic and RAG applications this course will cover the following topics and technologies:

- The current landscape of enterprise generative AI applications
- NVIDIA NIM microservices
- Components and architecture of enterprise-grade RAG applications
- At-scale inference considerations and optimations
- Kubernetes, Helm, and the NVIDIA RAG operator to deploy, manage, and scale RAG application services
- Prometheus and Grafana for cluster-wide application behavior and performance visibility
- Techniques for deploying and scaling multimodal RAG applications at scale

Date created: 4/23/2026 11:11:55 PM