Developing and Deploying AI/ML Applications on Red Hat OpenShift AI

Duration: 3 Days

Course Description:

Developing and Deploying AI/ML Applications on Red Hat OpenShift AI (AI267) provides students with the fundamental knowledge to manage the complete life cycle of modern AI applications. This course helps students build core skills for using Red Hat OpenShift AI to efficiently train, test, deploy, and monitor both predictive and generative AI models at scale.

This course is based on Red Hat OpenShift ® 4.20, and Red Hat OpenShift AI 3.3.

Target Audience:

- ML Engineers responsible for handling the operational tasks of the MLOps/LLMOps lifecycle, such as deployment, automation, and monitoring.
- Data Scientists who train, deploy, and track their own models.

Prerequisites:

- A basic understanding of machine learning principles and workflows.
- A basic understanding of Generative AI and Large Language Models (LLMs).
- Basic experience with Git is required
- Experience in Python development is required, or completion of the Python Programming with Red Hat (AD141) course
- Experience in Red Hat OpenShift is required, or completion of the Red Hat OpenShift Developer II: Building and Deploying Cloud-native Applications (DO288) course

Course Outlines:

Introduction to Red Hat OpenShift AI
- Identify how Red Hat OpenShift AI provides a complete MLOps and GenAIOps platform and how to use it to configure data science projects for team collaboration.

Using Workbenches for AI/ML Development
- Use workbench environments for AI/ML development and connect them to data sources and stores.

Fundamentals of Model Serving
- Prepare, deploy, and serve models by using OpenShift AI model serving capabilities.

Serving Predictive AI Models
- Deploy and serve predictive AI models with specific runtimes, including OpenVINO.

Monitoring AI Models
- Monitor deployed models for bias, data drift, and performance by using TrustyAI and observability tools to ensure reliable and ethical AI performance in production.

Introduction to AI Pipelines
- Create and manage basic data science pipelines by using Elyra and Kubeflow SDK to automate fundamental AI/ML workflows.

Advanced AI Pipelines Development and Experiments
- Implement advanced pipeline features including container components, artifacts management, Kubernetes configuration, and systematic experimentation for production MLOps workflows.

Gen AI Model Optimization and Evaluation
- Systematically optimize and evaluate large language models by using RHOAI's compression techniques and evaluation frameworks.

Building GenAI Applications
- Build production-ready GenAI applications by using industry patterns including RAG, agentic workflows, and trustworthy AI practices, and move beyond basic model serving to ship complete intelligent solutions.