Run:ai for AI Practitioners

Course Outline

Course Overview

NVIDIA Run:ai is a comprehensive platform that brings advanced, AI-native scheduling and resource management to GPU infrastructure, enabling enterprises to accelerate and scale AI operations efficiently, reduce costs, and accelerate innovation

In this course, you will learn the basics of NVIDIA Run:ai and how to create AI workloads. Designed for AI practitioners, this course guides you through the practical steps of creating, deploying, and monitoring interactive, inference, and training workloads using NVIDIA Run:ai.

Delivery Method

Instructor-led remote training sessions via NVIDIA Teams platform.
Hands-on lab exercises.

Target Audience

The course is designed for AI practitioners who want to gain the knowledge and skills necessary to run AI workloads on the NVIDIA Run:ai platform.

Course Duration

Remote | 2 sessions of 4 hours.

Course Prerequisites

- Basic understanding of Linux system administration and command-line operations
- Familiarity with containerized environments (e.g., Docker)
- General knowledge of AI/ML workflows or GPU-accelerated workloads is beneficial
- Experience with cluster or platform management is recommended
- Experience programming in Python.
- Experience working with machine learning libraries and tools, such as PyTorch and Jupyter Lab.

Course Objectives

By the end of this course, participants will be able to:

- Understand core Run:ai concepts, architecture, and resource management.
- Identify Run:ai workload types (Workspace, Training, Inference) and their appropriate use cases.
- Submit and manage workloads through the Run:ai UI.
- Set up and customize working environments for development, training, and inference.
- Configure and use platform assets such as departments, projects, credentials, data sources, environments, and compute resources.
- Monitor workloads using dashboards, metrics, logs, and event history.
- Troubleshoot common workload issues and apply platform best practices to improve performance and utilization.

Topics Covered

- Overview of NVIDIA Run:ai, dashboards, and core platform building blocks
- Understanding workload types: Workspace, Training, and Inference
- Managing platform assets: projects, credentials, data sources, environments, compute resources
- Running example workloads: interactive, inference, multi-GPU, and multi-node training

- Monitoring workloads using metrics, logs, and event history
- Troubleshooting common issues (pending, suspended, invalid workloads)
- Best practices for performance, stability, and efficient resource use
- Hands-on exercises and Q&A