# NVIDIA AI Enterprise Administration Public Training

## TRAINING OVERVIEW

The NVIDIA AI Enterprise training provides an overview of the NVIDIA AI Enterprise solution for IT Professionals. The course covers the platform and solution overview, hardware and software architecture, deployment options, licensing, temporal and spatial GPU partitioning, scaling, comprehensive validation, management, maintenance, monitoring, and troubleshooting. The related instruction and guidance are based on NVIDIA's best practices and cover the critical knowledge and skills to deploy, administer, and manage your NVIDIA AI Enterprise solution.

## TRAINING DELIVERY METHOD

Instructor-led remote sessions.

## TARGET AUDIENCE

The target audience for this course are IT professionals, which include system administrators and DevOps, who are expected to successfully deploy and administer the NVIDIA AI Enterprise solution.

## TRAINING DURATION

Remote: 3 sessions of 4 hours

## Learning Objectives:

After completing this course, the learner will be able to:

- Recall the NVIDIA AI Enterprise solution architecture
- Summarize the use cases
- Contrast the deployment options
- Demonstrate deployment on VMware vSphere
- Formulate a deployment workflow
- Interpret licensing
- Make use of solution validation techniques
- Recall vGPU partitioning options
- Utilize monitoring capabilities across the stack
- Develop management capabilities
- Translate AI stack software to use cases
- Solve common problems with troubleshooting

# NVIDIA AI Enterprise Administration Public Training

## TRAINING OUTLINE

### Introduction to NVIDIA AI Enterprise

- Software/Hardware platform overview
- NVIDIA AI Enterprise software suite
- NVIDIA Certified Systems overview
    - Overview of topologies
    - Hardware by use cases
- NVIDIA AI Enterprise features
    - Virtualization
    - Containers and Private registry
    - NGC access and exploration
- Brief Introduction to orchestration methods
    - VMWare vSphere
    - Tanzu
    - Bare Metal
    - CPU Only
    - OpenShift

### Deployment of NVIDIA AI Enterprise on VMWare vSphere

- Set the stage for deployment
    - Software stack overview
    - Identify NVIDIA Enterprise personas
    - Walk through NVIDIA licensing
    - Deployment steps
- vGPU introduction
    - Benefits of vGPU for NVIDIA AI Enterprise
    - GPU partitioning (MIG Mode)
- Downloading and Installing AI Enterprise Host software
    - Setting up and accessing NGC to download software
- vGPU profiles
- Creating a VM for NVIDIA AI Enterprise
    - VM Configuration options overview
    - Installing an operating system on a VM
    - Configuring vGPU options
    - Installing guest OS drivers
    - Install vGPU Software drivers
- Configuring NVIDIA Licensing

- Overview of NVIDIA License System
- Configuring a cloud license server (CLS)

- Installing the Docker and NVIDIA Container Toolkit
  - Docker and use of containers
  - NVIDIA Container Toolkit introduction (docker integration)
  - Verifying successful docker installation

Accessing and Installing AI Software

- Introduction to NGC and private registry
  - Overview of NGC and private registry
  - Benefits of NGC
- Generating an API key
- Navigating and reviewing entities in Enterprise Catalog
- Startup scripts and validation procedures

Management and Maintenance of NVIDIA AI Enterprise on VMware

- Management domains
  - Hardware and software component management
  - Common management tasks
  - Component monitoring
    - Virtual machine monitoring
    - AI stack monitoring using System Management Interface
  - NVIDIA vGPU Management
    - GPU partitioning overview
    - vGPU modes
      - Temporal partitioning - Time sliced scheduler
      - Spatial partitioning - vGPU profiles
  - NVIDIA AI Enterprise scaling options
    - Scaling options for deep learning training
    - Clustered multi-node options
  - Troubleshooting
    - Common problems and solutions
    - Common troubleshooting steps
    - NVIDIA Bug Report script
  - Upgrade path
- Resources & Documentation