NVIDIA AI Enterprise Administration Training

NVIDIA ACADEMY

TRAINING OVERVIEW

NVIDIA AI Enterprise technical training provides an overview of the NVIDIA AI Enterprise solution for the IT professional persona. 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 instruction and guidance are based on NVIDIA's best practices and cover the critical knowledge and skills required to deploy, administer, and manage your NVIDIA AI Enterprise solution.

Please note that while this course provides a comprehensive overview of all the deployment types available for NVIDIA AI Enterprise, **it focuses specifically on the VMware on vSphere with Docker containers deployment method.** It is important to understand that this course serves as an introduction to the various other deployment types; future versions of this course will delve deeper into each of the other deployment types.

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.

COURSE PREREQUISITES

To gain the most value from this course, the target audience should have a working knowledge of the following domains:

- Data Center Infrastructure
  - Servers
  - Storage
  - Networking
  - GPUs
  - Operating systems
- Virtualization
  - VMware vSphere
- Containerization
  - Docker

TRAINING DURATION

Remote: 4 sessions of 4 hours

Learning Objectives:

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

1. **Understand AI Platform Building Options:**
   - Explain the various options available for building an AI platform.
   - Outline the key considerations for selecting an appropriate building option.
2. **Understand the NVIDIA AI Enterprise Stack:**
   - Describe the components comprising the NVIDIA AI Enterprise stack.
   - Identify the role of each component within the stack for AI deployment.
3. **Gain Proficiency in NVIDIA AI Enterprise Hardware and Software:**
   - Analyze NVIDIA AI Enterprise hardware solutions, including certified systems and GPU sizing options.
   - Explain the software suite including AI workflows, frameworks, and deployment tools provided by NVIDIA.
4. **Understand Deployment Methods and Techniques:**
   - Introduce different deployment methods, such as VMWare vSphere, Tanzu, bare metal, and public cloud.
   - Understand the deployment workflow from a high level, covering essential steps and considerations.
   - Demonstrate deployment on VMware vSphere.
5. **Learn the Management and Maintenance of NVIDIA AI Enterprise:**
   - Outline management domains, encompassing hardware and software component management and common management tasks.
   - Describe NVIDIA vGPU management, including GPU partitioning, vGPU modes, and scaling options.
   - Address troubleshooting techniques for identifying and resolving common issues in NVIDIA AI Enterprise deployments.

TRAINING OUTLINE

Session #1

Exercise 1: Connect to vCenter

- Preflight check validation of access to assigned lab system

Solution Overview

- Options for building an AI platform
  - Outline the options for building an AI platform
- NVIDIA AI Enterprise stack
  - Describe the components of the NVIDIA AI Enterprise stack
- NVIDIA AI Enterprise Support
  - Summarize NVIDIA Enterprise Support
- Getting started with NVIDIA AI Enterprise
  - Evaluate NVIDIA AI Enterprise trial options (LaunchPad+Evals)
- Resources and documentation

Hardware Solutions

- NVIDIA AI Enterprise Hardware Solutions
  - NVIDIA Certified Systems overview
    - Overview of topologies
    - Hardware by use cases
  - DGX System types
  - Cloud options
  - GPU sizing options for workloads
  - NVIDIA networking platforms overview
    - Spectrum (Ethernet)
    - DPU
    - Quantum (InfiniBand)
  - GH200
- Resources and documentation

Software Solutions

- NVIDIA AI Enterprise software suite overview
- AI workflows, frameworks, and pretrained models
- AI and data science development and deployment tools
- Infrastructure optimization layer
- Cloud management native and orchestration layer
- Resources and documentation

Session #2

Deployment Methods

- Introduction to deployment methods
    - VMWare vSphere
    - Tanzu
    - Bare metal
    - CPU only
    - Red Hat OpenShift
    - Public cloud (AWS/Azure/GCP/OCI)
    - Upstream Kubernetes
    - Managed Kubernetes
        - EKS
        - GKS
        - AKS
    - Base Command Manager Essentials
    - Hypervisors (KVM variants + Nutanix AHV)
- High-level deployment workflow
- Guest operating systems
- Resources and documentation

Deployment on VMware vSphere with Docker Containers

- Overview of NVIDIA AI Enterprise deployment on VMware vSphere
- Creating a virtual machine for NVIDIA AI Enterprise
- Deploy Ubuntu and add a vGPU device
- NGC API and NGC CLI
- NVIDIA vGPU driver installation
- NVIDIA License Server (NLS)
- Docker and NVIDIA Container Toolkit
- Resources and documentation

Exercise 2: Create your First NVIDIA AI Enterprise VM

- Create virtual machine that will host your NVIDIA AI Enterprise stack
    - Virtual machine GPU configuration
    - Guest operating system installation
    - Guest driver installation
    - Validation

Session #3

Accessing and Installing AI Software

- Introduction to the NGC & NVIDIA AI Enterprise catalogs
- Downloading software from the NGC and NVIDIA AI Enterprise catalogs
- Sharing access to NVIDIA AI Enterprise software via the NGC Enterprise catalog
- Store, share, and collaborate using the NGC private registry
- Resources and documentation

Exercise 3: Installing Docker & Docker Utility for NVIDIA GPUs

- Deploy Docker Community Edition
- Deploy NVIDIA Container Toolkit
- Validation

AI Workflows

- What is an AI workflow?
- What are the benefits to AI workflows?
- What are the AI workflows offered by NVIDIA?
- What’s included in an AI workflow?
- Where can I get more information about AI workflows?
- Resources and documentation

Session #4

AI Inference

- Overview of Triton Inference Server, Triton Management Services, and NeMo Inference Microservice:
  - What are they and how do they work?
  - Architectures
- Deployment overview
- Basic example deployment demos
- Resources and documentation

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
    - Software branches
    - Upgrade path
- Resources and documentation

Exercise 4: Installing AI & Data Science Applications & Frameworks

- Create a dataset directory
- Make use of NGC to pull container
- Build BERT container on top of TensorFlow container
- Develop startup scripts for use by AI practitioners