PROFESSIONAL COURSE OUTLINE

Deploying a Model for Inference at Production Scale

NVIDIA

Program aligned

This NVIDIA DLI course teaches teams how to deploy machine learning models on a GPU server using NVIDIA Triton Inference Server. It is especially useful for organizations that have moved beyond experimentation and need serving capability.

AI | Intermediate | NVIDIA Triton Inference Server

PROGRAM CODE	DELIVERY	DURATION	CERTIFICATION
Program aligned	Virtual, On-site, or Hybrid	4 hours	Available on request

AUDIENCE PROFILE

Who This Program Is For

Built for practitioners who already train models and now need deployment and inference capability on GPU-based serving infrastructure.

PROGRAM SUMMARY

What This Course Covers

Official NVIDIA DLI program focused on deploying machine learning models to GPU servers with NVIDIA Triton Inference Server.

TAILORED DELIVERY

Adapt the program around your team.

This outline can be adapted for virtual, on-site, or hybrid delivery, with emphasis adjusted for your team's platform priorities, role mix, and implementation goals.

Enterprise-ready delivery format

VNode ITeS can align labs, examples, delivery pace, and assessment checkpoints to the required audience profile while preserving the official program sequence where applicable.

COMPLETE MODULE SEQUENCE

Module Flow and Topic Coverage

The structure below presents the current delivery flow for this program, including the associated topics covered under each module.

TRAINING PROVIDER
VNode ITeS
MICROSOFT CERTIFIED TRAINER
vnodeites.com
linkedin.com/company/vnodeites

TALK TO US
info@vnodeites.com
+91 9419 11 4792
+91 9419 11 4792
+91 7780 81 1685

OFFICE
DLF Cyber City, Gurugram
Haryana — India
Azure · AI · Data · Power Platform

1

MODULE 1

Build the foundation for production inference

Understand the core deployment patterns and operational considerations involved in moving trained models into production inference environments.

- Inference deployment foundations
- GPU-backed deployment workflows

2

MODULE 2

Serve and manage models with Triton

Use NVIDIA Triton to expose models for inference while improving deployment readiness and scalability for AI applications.

- Serving models with Triton
- Production inference considerations

SCHEDULE A SESSION

Ready to run this program with your team?

Get in touch to discuss delivery format, timeline, and team size. This outline can be customised to your platform priorities and role mix.

Book a Discovery Call

vnodeites.com

TRAINING PROVIDER
VNode ITeS
MICROSOFT CERTIFIED TRAINER
vnodeites.com
linkedin.com/company/vnodeites

TALK TO US
info@vnodeites.com
+91 9419 11 4792
+91 9419 11 4792
+91 7780 81 1685

OFFICE
DLF Cyber City, Gurugram
Haryana — India
Azure · AI · Data · Power Platform