

AI Infrastructure Management

COURSE OVERVIEW

Scaling artificial intelligence from experimentation to production requires a robust, efficient, and scalable infrastructure. This course provides the essential knowledge and practical skills to design, implement, and manage the complete technology stack that powers enterprise AI. It covers key components such as GPU, CPU, and DPU hardware, along with networking and storage essentials for AI workloads. Participants will learn to navigate the complex landscape of hardware, cloud services, and orchestration tools to build a foundation that supports the entire machine learning lifecycle from data processing and model training to deployment and monitoring while optimizing for performance, cost, and reliability.

WHO SHOULD ATTEND?

This course is designed for technology professionals responsible for building and maintaining AI capabilities, including ML infrastructure engineers, DevOps and MLOps engineers, cloud architects and solutions engineers, IT infrastructure managers, and data platform engineers seeking practical skills to design, deploy, and manage scalable AI systems.

COURSE OUTCOMES

Delegates will gain the skills and knowledge to:

- Design and provision scalable compute infrastructure for training and inference workloads.
- Implement and manage GPU-accelerated computing environments efficiently.
- Select and configure optimal storage solutions for various data and model types.
- Orchestrate AI workloads using containerization and Kubernetes.
- Automate MLOps pipelines for continuous integration and delivery.
- Implement monitoring, logging, and cost optimization strategies.
- Ensure security, governance, and compliance across AI infrastructure.

KEY COURSE HIGHLIGHTS

At the end of the course, you will understand;

- The core components of AI infrastructure, including hardware, software, and cloud solutions for scalable deployment.
- How to design and manage architectures that support data processing, model training, and real-time AI inference.
- Hardware selection considerations, including GPUs, TPUs, and accelerators optimized for machine learning workloads.
- Methods to automate resource allocation, monitoring, and scaling to maximize performance and control costs.
- Approaches to ensure data integrity, security, and compliance, including risk management and privacy controls.
- Strategies for maintaining, troubleshooting, and upgrading AI infrastructure to meet evolving business and technology demands.

All our courses are dual-certificate courses. At the end of the training, the delegates will receive two certificates.

1. A GTC end-of-course certificate
2. Continuing Professional Development (CPD) Certificate of completion with earned credits awarded