GTC Transforming Emerging Economies

GTC International Consulting Limited
Riverbank House
1 Putney Bridge Approach
Fulham, London, SW6 3BQ
T: +44(0)2037055710
E:enquiries@thegtcgroup.com
W: www.thegtcgroup.com

Scalable AI Deployments

COURSE OVERVIEW

Bridging the gap between a working model and a robust, production-grade AI service is the most significant challenge in modern AI. This course has been designed to provide the essential engineering principles and hands-on techniques to deploy, manage, and scale AI systems effectively. Participants will move beyond notebooks and local scripts to master the tools and architectures like containerization, orchestration, and MLOps required to serve models reliably to millions of users, ensuring performance, monitoring for drift, and automating the entire lifecycle.

WHO SHOULD ATTEND?

This course is ideal for data scientists, machine learning engineers, and DevOps professionals looking to operationalize AI in real-world environments. It is especially relevant for ML engineers and data scientists building production models, DevOps and MLOps engineers responsible for deployment and automation, AI/ML platform teams managing infrastructure and pipelines, and software developers creating AI-powered applications who need practical skills in scalable, reliable AI operations.

COURSE OUTCOMES

Delegates will gain the skills and knowledge to:

- Containerize and package ML models for consistent, portable deployment across any environment.
- Design and implement scalable serving architectures using REST APIs and gRPC for high-throughput inference.
- Leverage orchestration tools like Kubernetes to automate deployment, scaling, and management of AI services.
- Build continuous integration and delivery (CI/CD) pipelines specifically tailored for machine learning (MLOps).
- Implement comprehensive monitoring for model performance, data drift, and system health.
- Optimize inference latency and cost for real-time and batch processing scenarios.

KEY COURSE HIGHLIGHTS

At the end of the course, you will understand;

- How to design AI deployments that scale efficiently across varied workloads.
- Managing AI infrastructure for elasticity using auto-scaling and resource optimization.
- Ensuring security, compliance, and governance in large-scale AI deployments.
- Techniques for building modular, microservices-based AI architectures.
- Monitoring, logging, and troubleshooting AI systems in production at scale.
- Strategies for cost-effective deployment and maintenance of enterprise AI solutions.
- Best practices for containerization, versioning, and continuous integration/deployment (CI/CD) of AI models.

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