GTC Transforming Emerging Economies

GTC Training Consulting Group Ltd,
22 Kumasi Crescent,
Off Aminu Kano Crescent,
Wuse 2, Abuja.
Tel: +234(0) 9056761232
Email: enquiries@thegtcgroup.com
Web: www.thegtcgroup.com

Scalable AI with Apache Spark and MLlib

COURSE OVERVIEW

This course equips participants with the knowledge and relevant skills required to build, train, and deploy scalable machine learning models using distributed computing. It introduces the core concepts of Spark's architecture, explores data preprocessing and feature engineering at scale, and covers supervised and unsupervised learning techniques implemented with MLlib. The course emphasizes practical applications, performance optimization, and integration of Spark with real world AI pipelines, equipping participants to handle large scale datasets and deliver efficient, production ready machine learning solutions.

WHO SHOULD ATTEND?

This course is designed for data scientists, machine learning engineers, big data practitioners, and software developers who want to leverage Spark and MLlib to scale AI workloads. It is also valuable for technical managers and decision-makers seeking to understand the capabilities and limitations of scalable machine learning systems for business or research applications. Prior knowledge of Python, basic machine learning concepts, and familiarity with distributed systems will be beneficial.

COURSE OUTCOMES

Delegates will gain the skills and knowledge to:

- Know the fundamentals of distributed computing with Apache Spark.
- Build and train scalable machine learning models using Mllib.
- Apply classification, regression, clustering, and recommendation techniques.
- Optimize ML workflows and manage model persistence.
- Use Spark ML pipelines for automation and reproducibility.
- Integrate Spark with cloud platforms and data lakes.
- Troubleshoot performance issues and tune distributed systems.

KEY COURSE HIGHLIGHTS

At the end of the course, you will understand;

- Spark architecture and distributed data processing.
- Data wrangling and feature engineering at scale.
- Building ML pipelines for automation and deployment.
- Distributed model evaluation and parameter tuning.
- Integration with Hadoop, HDFS, and cloud services (e.g., AWS, Azure).
- Case studies in fraud detection, predictive analytics, and recommendations.
- Labs using Databricks or standalone Spark environments.

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