Faysal Aziz

Apache Spark Programming

with Databricks using PySpark

FA-0102 · 3-day

Software Development · Data & Analytics

Introduction

This is an intensive 3-day course on Apache Spark Programming with Databricks, tailored for Python developers who are new to both Databricks and Apache Spark. This course is designed to provide a comprehensive understanding of Apache Spark's capabilities in big data processing and analytics, leveraging the Databricks platform.

We'll dive into the essentials of Spark and Databricks, focusing on practical applications and hands-on lab sessions using Ubuntu remote servers. The course will enable participants to harness the power of Spark within the Databricks environment, empowering them to tackle big data challenges with advanced analytics techniques.

Learning outcomes

By the end of this course, participants will:

1. Understand the fundamentals of Apache Spark and its role in big data processing.
2. Gain proficiency in using Databricks for Spark programming.
3. Be able to build and deploy data processing and analytics tasks on Spark.
4. Learn to optimize Spark applications for performance and efficiency.
5. Develop skills to manage and analyze big data in a scalable way.

Prerequisites

- Experience in Python programming.
- Basic understanding of big data concepts.
- Familiarity with cloud computing and distributed systems is beneficial.

Training outline

1. Introduction to Apache Spark and Databricks:

1. Overview of Big Data and Distributed Computing.
2. Understanding Apache Spark and its Ecosystem.
3. Introduction to Databricks and its Integration with Spark.
4. Setting up the Environment on Ubuntu Remote Servers.

2. Spark Core Concepts and Architecture:

1. Understanding RDDs (Resilient Distributed Datasets).
2. Spark Architecture and Components (Driver, Executors).
3. Deep Dive into Spark's Execution Model.
4. Data Partitioning and Distribution.

3. Data Processing with Spark:

1. Loading and Transforming Data with Spark.
2. Performing Aggregations and Joins on Large Datasets.
3. Spark SQL for Structured Data Processing.
4. Using DataFrames and Datasets in Spark.

4. Databricks Essentials:

1. Navigating the Databricks Workspace.
2. Interactive Analysis with Databricks Notebooks.
3. Integrating Databricks with Various Data Sources.
4. Databricks Utilities and Features for Enhanced Productivity.

5. Advanced Spark Programming:

1. Advanced Data Processing Techniques in Spark.
2. Spark Streaming for Real-time Data Processing.
3. Graph Processing with GraphX.
4. Machine Learning with MLlib.

6. Optimizing and Tuning Spark Applications:

1. Understanding Spark Configuration and Tuning.
2. Debugging and Monitoring Spark Applications.
3. Best Practices for Performance Optimization.
4. Resource Management in Spark.

7. Building End-to-End Spark Applications:

1. Designing and Building a Complete Spark Application.
2. Deploying Spark Applications on the Cluster.
3. Managing and Scheduling Spark Jobs.
4. Hands-on Lab: Building a Real-world Data Processing Application.

The course concludes with a comprehensive lab session where participants will apply their learning to build and deploy a real-world data processing application using Spark and Databricks on Ubuntu remote servers.

© 2026 Faysal Aziz · Kuala Lumpur