Faysal Aziz

Hadoop HDFS and Spark with PySpark

FA-0237 · three-day

Data & Analytics

Introduction

This three-day intensive course is designed to introduce learners to the fundamentals and advanced features of Hadoop HDFS and Spark using PySpark. The course will cover a wide range of topics, from the basic structure of Hadoop and Spark to their execution models, dataframes, transformations, jobs, stages, and more. The course will also delve into the application of these tools for machine learning.

Learning outcomes

By the end of this course, learners will be able to:

- Understand the basic structure and functionality of Hadoop HDFS and Spark.
- Use PySpark to interact with Hadoop HDFS and Spark.
- Perform a variety of data operations using Spark, such as reading, writing, transforming, and aggregating data.
- Understand the execution model of Spark including jobs, stages, and tasks.
- Implement advanced transformations and user-defined functions.
- Perform SQL queries in Spark.
- Handle unstructured data in Spark.
- Understand and apply optimization techniques in Spark.
- Use Spark for machine learning applications.

Prerequisites

- Basic knowledge of Python programming.
- Basic understanding of distributed systems.

- Familiarity with SQL and relational databases.
- Basic knowledge of machine learning principles is helpful, but not required.

Training outline

1. Day 1: Introduction to Hadoop HDFS, Spark and PySpark

- Introduction to Distributed Systems and Big Data
- Overview of Hadoop and HDFS
- Introduction to Spark and its Execution Model
- Introduction to PySpark
- Introduction to Dataframes in Spark
- Data Transformations in Spark
- Spark Jobs and Stages
- Lab: Setting up PySpark and performing basic operations

2. Day 2: Deep Dive into Spark

- Introduction to Spark's Structured API
- Resilient Distributed Dataset (RDD) in Spark
- Data Sinks in Spark
- Reading and Writing Data in Spark
- Understanding Spark Schemas
- SQL in Spark
- Handling Unstructured Data in Spark
- Lab: Data operations and SQL queries in PySpark

3. Day 3: Advanced Spark and Machine Learning with Spark

- Advanced Transformations in Spark
- User Defined Functions in Spark
- Data Aggregations, Grouping, and Joins in Spark
- Shuffles in Spark
- Optimization Techniques in Spark
- Introduction to Clustering in Spark
- Introduction to Machine Learning with Spark

- Lab: Implementing a machine learning model with Spark

By the end of this course, learners will be equipped with the knowledge and skills to use Hadoop HDFS and Spark with PySpark, enabling them to handle complex data operations, apply optimization techniques, and leverage machine learning capabilities in their big data projects.

© 2026 Faysal Aziz · Kuala Lumpur