PySpark for Data Testing Automation

Course Description :

This PySpark training course is designed to equip you with the essential skills and knowledge needed to master PySpark for data testing automation. The course covers fundamental concepts, advanced techniques, and practical hands-on labs to ensure you can effectively test and automate data processes using PySpark. You'll learn how to write robust tests, set up CI pipelines, and handle real-world data testing scenarios.

Duration : 24 hours

Prerequisites :

- **Basic knowledge of Python programming**: Understanding of Python syntax and basic programming concepts.
- **Understanding of data processing concepts**: Familiarity with data processing, SQL, and database systems.
- **Familiarity with SQL**: Basic knowledge of SQL queries and database operations.
- **Experience with Big Data tools**: Optional but beneficial to have prior experience with tools like Hadoop or Spark.

Table of Contents :

Module 01 : Introduction to PySpark

- Overview of PySpark
- Setting up PySpark
- Basic concepts: RDDs, DataFrames, and Datasets

Module 02 : PySpark Basics

- Creating and manipulating DataFrames
- Working with RDDs
- PySpark SQL and DataFrames

Module 03 : Data Testing in PySpark

- Importance of data testing
- Writing unit tests for PySpark code
- Using PySpark testing utilities

Module 04 : Testing PySpark Applications

- Setting up a testing environment
- Writing test cases for PySpark transformations
- Using built-in PySpark test utility functions

Module 05 : Advanced Testing Techniques

- Mocking data for testing
- Using pytest with PySpark
- Integration testing with PySpark

Module 06 : Continuous Integration (CI) for PySpark

- Setting up CI pipelines
- Automated testing in CI
- Best practices for PySpark CI