Spark Programming with Python

Course 3577 – 24 Hours

Overview

Apache Spark™ is a fast and general engine for large-scale data processing, and arguably the first open source software that makes distributed programming truly accessible to data scientists.
Using Apache Spark™ you can write applications quickly using Java, Scala, Python, and R.
Python is a dynamic object-oriented programming language. Due to its powerful and flexible syntax, Python is an excellent platform for scientific computing. Versatility, simplicity of use, high portability and the large number of open source modules and packages make it very popular for scientific use.
This course covers all the fundamentals about Apache Spark with Python and guides you through everything you need to know about developing Spark applications with Python.
At the end of this course, you will gain deep understanding about the Spark architecture and general big data analysis and manipulations skills.

Who Should Attend

- Developers who would like to start using Spark

Prerequisites

- Previous programming experience

Course Contents

Python Crash Course

- Basic Syntax
- Tuples, Lists, and Dictionaries
- Lambda functions

Spark Core

- What is Spark?
- RDD Basics.
- Transformations and Actions.
- Chaining Transformations
- Using Anonymous Functions
- Map Reduce and Shuffles.
- Caching.
- Web Monitoring.
- Usecases.
- Serialization.
- Troubleshooting

Spark Streaming

- Introduction to Spark Streaming.
- DStream.
- Usecases.

Spark SQL

- Introduction to Spark SQL & DataFrames.
- Integration with Different Data Sources.
- Integration with Hive.

Spark MLlib

- Introduction to Machine Learning.
- The MLlib API.
- Basic Usecases.