

Spark for ML and PySpark

24 hours

Course Overview:

Apache Spark is a fast and general engine for large-scale data processing. It is 100x faster than Hadoop MapReduce in memory and 10x faster on disk. Apache Spark is designed to write applications quickly in Java, Scala or Python. You can use it interactively from the Scala and Python shells. You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN. Access data in HDFS, Cassandra, HBase, Hive, and any Hadoop data source.

In this PySpark training course, we will teach you everything you need to know about the Spark Python API. This course is designed for users that already have a basic working knowledge of Python.

Who Should Attend:

This course is designed for developers, BI experts, analysts with python programming experience, working experience with datasets, including data analytics.

Required skills:

- Working experience in python programming
- Basic knowledge of machine learning
- Basic knowledge of SQL is helpful
- Prior knowledge of Hadoop is not required

Course Contents:

- Big Data, Spark & PySpark
 - Introduction to Big Data & Spark (in DataBricks)
 - PySpark (RDDs)
 - PySpark SQL (Data Frames)
- Machine Learning with Spark - PySpark ML
 - Regression (Linear regression)
 - Classification (Decision trees)
 - Clustering (k-means)