

---

## Spark and Spark streaming with Python

40 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, Python and R. 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.

This course will teach you to create applications in Spark with the implementation of Python programming. It provides a clear comparison between Spark and Hadoop and covers techniques to increasing your application performance and enabling high-speed processing.

The module Spark Streaming will explain how easy to build scalable fault-tolerant streaming applications. It will let you to work with large scale streaming data using familiar batch processing abstractions.

### 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 SQL is helpful
- Prior knowledge of Hadoop is not required

### Course Contents:

- Introduction to Big Data and Hadoop Ecosystem
- Introduction to Spark
- RDD
- Broadcast & Accumulators
- Spark Partitioning
- Spark SQL API
- Migration from Spark 1 to Spark 2
- Working with Data Sources
- DataFrame Operations
- Kafka
- Spark Streaming
- Structured Streaming
- Spark UI
- Performance tuning
- Log Management
- Shutdown streaming application