

Apache Spark Course Content

Module 1-Why Spark? Explain Spark and Hadoop Distributed File System

- What is Spark
- Comparison with Hadoop
- Components of Spark

Module 2-Spark Components, Common Spark Algorithms-Iterative Algorithms, Graph Analysis, Machine Learning

- Apache Spark- Introduction, Consistency, Availability, Partition
- Unified Stack Spark
- Spark Components
- Comparison with Hadoop – Scalding example, mahout, storm, graph

Module 3-Running Spark on a Cluster, Writing Spark Applications using Python, Java, Scala

- Explain python example
- Show installing a spark
- Explain driver program
- Explaining spark context with example
- Define weakly typed variable
- Combine scala and java seamlessly.
- Explain concurrency and distribution.
- Explain what is trait.
- Explain higher order function with example.
- Define OFI scheduler.
- Advantages of Spark
- Example of Lamda using spark
- Explain Mapreduce with example

Module 4-RDD and its operation

- Difference between RISC and CISC
- Define Apache Mesos
- Cartesian product between two RDD

- Define count
- Define Filter
- Define Fold
- Define API Operations
- Define Factors

Module 5-Spark, Hadoop, and the Enterprise Data Centre, Common Spark Algorithms

- How hadoop cluster is different from spark
- Define writing data
- Explain sequence file and its usefulness
- Define protocol buffers
- Define text file, CSV, Object Files and File System
- Define sparse metrics
- Explain RDD and Compression
- Explain data stores and its usefulness

Module 6-Spark Streaming

- Define Elastic Search
- Explain Streaming and its usefulness
- Apache zookeeper
- Define Dstream
- Define mapreduce word count
- Explain Parquet
- Scala ORM
- Define Mlib
- Explain multi graphix and its usefulness
- Define property graph