

INFOSOFT IT SOLUTIONS

Training | Projects | Placements

Revathi Apartments, Ameerpet, 1st Floor, Opposite Annapurna Block, Infosoft It solutions,
Software Training & Development Institute, +91-9059683947|+91-9182540872

Apache Spark

Module 1: Introduction

- 1. Overview of Hadoop
- 2. Architecture of HDFS & YARN
- 3. Overview of Spark version 2.2.0
- 4. Spark Architecture
- 5. Spark Components
- 6. Comparison of Spark & Hadoop
- 7. Installation of Spark v 2.2.0 on Linux 64 bit

Module 2: Spark Core

- 1. Exploring the Spark shell
- 2. Creating Spark Context
- 3. Operations on Resilient Distributed Dataset – RDD
- 4. Transformations & Actions
- 5. Loading Data and Saving Data

Module 3: Spark SQL & Hive SQL

- 1. Introduction to SQL Operations
- 2. SQL Context

- 3. Data Frame
- 4. Working with Hive
- 5. Loading Partitioned Tables
- 6. Processing CSV, Json ,Parquet files

Module 4: Scala Programming

- 1. Introduction to Scala
- 2. Feature of Scala
- 3. Scala vs Java Comparison
- 4. Data types
- 5. Data Structure
- 6. Arrays
- 7. Literals
- 8. Logical Operators
- 9. Mutable & Immutable variables
- 10. Type interface

Module 5: Scala Functions

- 1. OOps vs Functions
- 2. Anonymous
- 3. Recursive
- 4. Call-by-name
- 5. Currying
- 6. Conditional statement

Module 6: Scala Collections

- 1. List

- 2. Map
- 3. Sets
- 4. Options
- 5. Tuples
- 6. Mutable collection
- 7. Immutable collection
- 8. Iterating
- 9. Filtering and counting
- 10. Group By
- 11. Flat Map
- 12. Word count
- 13. File Access

Module 7: Scala Object Oriented Programming

- 1. Classes ,Objects & Properties
- 2. Inheritance

Module 8: Spark Submit

- 1. Maven build tool implementation
- 2. Build Libraries
- 3. Create Jar files
- 4. Spark-Submit

Module 9: Spark Streaming

- 1. Overview of Spark Streaming
- 2. Architecture of Spark Streaming
- 3. File streaming

- 4. Twitter Streaming

Module 10: Kafka Streaming

- 1. Overview of Kafka Streaming
- 2. Architecture of Kafka Streaming
- 3. Kafka Installation
- 4. Topic
- 5. Producer
- 6. Consumer
- 7. File streaming
- 8. Twitter Streaming

Module 11: Spark Mlib

- 1. Overview of Machine Learning Algorithm
- 2. Linear Regression
- 3. Logistic Regression

Module 12: Spark GraphX

- 1. GraphX overview
- 2. Vertices
- 3. Edges
- 4. Triplets
- 5. Page Rank
- 6. Pregel

Module 13: Performance Tuning

- 1. On-Off-heap memory tuning

- 2. Kryo Serialization
- 3. Broadcast Variable
- 4. Accumulator Variable
- 5. DAG Scheduler
- 6. Data Locality
- 7. Check Pointing
- 8. Speculative Execution
- 9. Garbage Collection

Module 14: Project Planning, Monitoring Trouble Shooting

- 1. Master – Driver Node capacity
- 2. Slave – Worker Node capacity
- 3. Executor capacity
- 4. Executor core capacity
- 5. Project scenario and execution
- 6. Out-of-memory error handling
- 7. Master logs, Worker logs, Driver logs
- 8. Monitoring Web UI
- 9. Heap memory dump