1. Introduction to Scala for Apache Spark

In this module, you will understand the basics of Scala that are required for programming Spark applications. You can learn about the basic constructs of Scala such as variable types, control structures, collections, and more.

- What is Scala?
- Why Scala for Spark?
- Scala in other frameworks
- Introduction to Scala REPL, basic Scala operations, Variable Types in Scala, Control Structures in Scala, Foreach loop, Functions, Procedures, Collections in Scala- Array, ArrayBuffer, Map, Tuples, Lists.

2. OOPS and Functional Programming in Scala

In this module, you will learn about object oriented programming and functional programming techniques in Scala.

- Class in Scala, Getters and Setters, Custom Getters and Setters, Properties with only Getters, Auxiliary Constructor, Primary Constructor, Singletons.

- Companion Objects, Extending a Class, Overriding Methods, Traits as Interfaces, Layered Traits, Functional Programming, Higher Order Functions, Anonymous Functions

3. Introduction to Big Data and Apache Spark

In this module, you will understand about big data, challenges associated with it and the different frameworks available. The module also includes a first-hand introduction to Spark

- Introduction to big data
- Challenges with big data
- Batch Vs. Real Time big data analytics
- Batch Analytics - Hadoop Ecosystem Overview
- Real-time Analytics Options
- Streaming Data – Spark
- In-memory data – Spark
- What is Spark?, Spark Ecosystem, modes of Spark,
- Spark installation demo, overview of Spark on a cluster
- Spark Standalone cluster, Spark Web UI

4. Spark Common Operations

In this module, you will learn how to invoke Spark Shell and use it for various common operations.

- Invoking Spark Shell

- Creating the Spark Context, loading a file in Shell, performing basic Operations on files in Spark Shell
- Overview of SBT, building a Spark project with SBT, running Spark project with SBT local mode, Spark mode, caching overview
- Distributed Persistence

5. Playing with RDDs

In this module, you will learn one of the fundamental building blocks of Spark - RDDs and related manipulations for implementing business logics.

- RDDs, transformations in RDD, actions in RDD, loading data in RDD, saving data through RDD
- Key-Value Pair RDD
- MapReduce and Pair RDD Operations
- Spark and Hadoop Integration-HDFS
- Spark and Hadoop Integration-Yarn
- Handling Sequence Files, Partitioner.

6. Spark Streaming and MLlib

In this module, you will learn about the major APIs that Spark offers. You will get an opportunity to work on Spark streaming which makes it easy to build scalable fault-tolerant streaming applications, MLlib which is Spark's machine learning library.

- Spark Streaming Architecture
- First Spark Streaming Program
- Transformations in Spark Streaming
- Fault tolerance in Spark Streaming,
- Check pointing
- Parallelism level, machine learning with Spark, data types,
- Algorithms – statistics, classification and regression, clustering, collaborative filtering.

7. GraphX, SparkSQL and Performance Tuning in Spark

In this module, you will learn about Spark SQL that is used to process structured data with SQL queries, graph analysis with Spark, GraphX for graphs and graph-parallel computation. You will also0 get a chance to learn the various ways to optimize performance in Spark.

- Analyze Hive and Spark SQL architecture
- SQLContext in Spark SQL
- Working with DataFrames
- Implementing an example for Spark SQL
- Integrating hive and Spark SQL
- Support for JSON and Parquet File Formats
- Implement data visualization in Spark
- Loading of data
- Hive queries through Spark, testing tips in Scala, performance tuning tips in Spark,
- Shared variables: Broadcast Variables, Shared Variables: Accumulators.

8. A complete project on Apache Spark

In this module, you will get an opportunity to work on a live Spark project where you can implement the learnings from previous modules hands-on, and solve a real-time use case.

Design a system to replay the real time replay of transactions in HDFS using Spark.

Technologies Used:

1. Spark Streaming
2. Kafka (for messaging)
3. HDFS (for storage)
4. Core Spark API (for aggregation)