Apache Spark Syllabus

Module 1: Introduction to Spark

- Introduction to Spark
- Spark overcomes the drawbacks of working on MapReduce
- Understanding in-memory MapReduce
- Interactive operations on MapReduce
- Spark stack, fine vs. coarse-grained update, Spark Hadoop YARN, HDFS Revision, and YARN Revision
- The overview of Spark and how it is better than Hadoop
- Deploying Spark without Hadoop
- Spark history server and Cloudera distribution

Module 2: Spark Basics

- Spark installation guide
- Spark configuration
- Memory management
- Executor memory vs. driver memory
- Working with Spark Shell
- The concept of resilient distributed datasets (RDD)
- Learning to do functional programming in Spark
- The architecture of Spark

Module 3: Working with RDDs in Spark

- Spark RDD
- Creating RDDs
- RDD partitioning
- Operations and transformation in RDD
- Deep dive into Spark RDDs
- The RDD general operations
- Read-only partitioned collection of records
- Using the concept of RDD for faster and efficient data processing

- RDD action for the collect, count, collects map, save-as-text-files, and pair RDD functions

Module 4: Aggregating Data with Pair RDDs

- Understanding the concept of key-value pair in RDDs
- Learning how Spark makes MapReduce operations faster
- Various operations of RDD
- MapReduce interactive operations
- Fine and coarse-grained update
- Spark stack

Module 5: Writing and Deploying Spark Applications

- Comparing the Spark applications with Spark Shell
- Creating a Spark application using Scala or Java
- Deploying a Spark application
- Scala built application
- Creation of the mutable list, set and set operations, list, tuple, and concatenating list
- Creating an application using SBT
- Deploying an application using Maven
- The web user interface of Spark application
- A real-world example of Spark
- Configuring of Spark

Module 6: Parallel Processing

- Learning about Spark parallel processing
- Deploying on a cluster
- Introduction to Spark partitions
- File-based partitioning of RDDs
- Understanding of HDFS and data locality
- Mastering the technique of parallel operations
- Comparing repartition and coalesce
- RDD actions

Module 7: Spark RDD Persistence

- The execution flow in Spark

- Understanding the RDD persistence overview
- Spark execution flow, and Spark terminology
- Distribution shared memory vs RDD
- RDD limitations
- Spark shell arguments
- Distributed persistence
- RDD lineage
- Key-value pair for sorting implicit conversions like CountByKey, ReduceByKey, SortByKey, and AggregateByKey

Module 8: Spark MLlib

- Introduction to Machine Learning
- Types of Machine Learning
- Introduction to MLlib
- Various ML algorithms supported by MLlib
- Linear regression, logistic regression, decision tree, random forest, and K-means clustering techniques

Module 9: Integrating Apache Flume and Apache Kafka

- Why Kafka and what is Kafka?
- Kafka architecture
- Kafka workflow
- Configuring Kafka cluster
- Operations
- Kafka monitoring tools
- Integrating Apache Flume and Apache Kafka

Module 10: Spark Streaming

- Introduction to Spark Streaming
- Features of Spark Streaming
- Spark Streaming workflow
- Initializing StreamingContext, discretized Streams (DStreams), input DStreams and Receivers

- Transformations on DStreams, output operations on DStreams, windowed operators and why it is useful
- Important windowed operators and stateful operators

Module 11: Improving Spark Performance

- Introduction to various variables in Spark like shared variables and broadcast variables
- Learning about accumulators
- The common performance issues
- Troubleshooting the performance problems

Module 12: Spark SQL and Data Frames

- Learning about Spark SQL
- The context of SQL in Spark for providing structured data processing
- JSON support in Spark SQL
- Working with XML data
- Parquet files
- Creating Hive context
- Writing data frame to Hive
- Reading JDBC files
- Understanding the data frames in Spark
- Creating Data Frames
- Manual inferring of schema
- Working with CSV files
- Reading JDBC tables
- Data frame to JDBC
- User-defined functions in Spark SQL
- Shared variables and accumulators
- Learning to query and transform data in data frames
- Data frame provides the benefit of both Spark RDD and Spark SQL
- Deploying Hive on Spark as the execution engine

Module 13: Scheduling/Partitioning

- Learning about the scheduling and partitioning in Spark

- Hash partition
- Range partition
- Scheduling within and around applications
- Static partitioning, dynamic sharing, and fair scheduling
- Map partition with index, the Zip, and GroupByKey
- Spark master high availability, standby masters with ZooKeeper
- Single-node recovery with the local file system and high order functions