SPRINTZEAL
Empowering Professionals

Big Data Hadoop and Spark Developer Certification Training | Course Agenda

Lesson 1: Introduction to Bigdata and Hadoop Ecosystem

In this lesson you will learn about traditional systems, problems associated with traditional large scale systems, what is Hadoop and it’s ecosystem. Topics covered are:

- Traditional models
- Problems with Traditional Large-scale Systems
- What is Hadoop?
- The Hadoop EcoSystem

Lesson 2: HDFS and Hadoop Architecture

In this lesson you will learn about distributed processing on cluster, HDFS architecture, how to use HDFS, YARN as a resource manager, yarn architecture and how to work with YARN. Topics covered are:

- Distributed Processing on a Cluster
- Storage: HDFS Architecture
- Storage: Using HDFS
- Resource Management: YARN

- Resource Management: YARN Architecture
- Resource Management: Working with YARN

Lesson 3: MapReduce and Sqoop

In this lesson you will learn about Mapreduce and its characteristics, advance MapReduce concepts, overview of Sqoop, basic import and exports in Sqoop, improving Sqoop's performance, limitations of Sqoop and Sqoop2. Topics covered are:

- Mapreduce
- Mapreduce characterstics
- Advance mapreduce concepts
- Sqoop Overview
- Basic Imports and Exports
- Improving Sqoop's Performance
- Limitations of Sqoop
- Sqoop 2

Lesson 4: Basics of Impala and Hive

In this lesson you will be introduced to Hive and Impala, why to use Hive and Impala, differences between Hive and Impala, how Hive and Impala works and comparison of Hive to traditional databases. Topics covered are:

- Introduction to Impala and Hive
- Why Use Impala and Hive?
- Difference between Hive and Impala
- How Hive and Impala works?

Comparing Hive to Traditional Databases

Lesson 5: Working with Impala and Hive

In this lesson you will learn about metastore, how to create databases and table in Hive and Impala, loading data into tables of Hive and Impala, HCatalog and how impala works on cluster. Topics covered are:

- Metastore
- Creating Databases and Tables
- Loading Data into Tables
- HCatalog
- Impala on cluster

Lesson 6: Type of Data Formats

In this lesson you will learn about different tyoes of file formats which are available, Hadoop tool support for file format, avro schemas, using avro with Hive and Swoop and Avro schema evolution. Topics covered are:

- File Format
- Hadoop Tool Support for File Formats

- Avro Schemas
- Using Avro with Hive and Sqoop
- Avro Schema Evolution

Lesson 7: Advance HIVE concept and Data File Partitioning

In this lesson you will learn about portioning in Hive and Impala, portioning in Impala and Hive, when to use partition, bucketing in Hive and more advance concepts in Hive. Topics covered are:

- Partitioning Overview
- When to use Partition?
- Partitioning in Impala and Hive
- Bucketing in Hive
- Advance concepts in Hive

Lesson 8: Apache Flume and HBase

In this lesson you will learn about apache flume, flume artitecture, flume sources, flume sinks, flume sinks, flume channels, flume configurations, introction to HBase, HBase artitecture, data storage in HBase, HBase vs RDBMS. Topics covered are:

- What is Apache Flume?
- Basic Flume Architecture
- Flume Sources
- Flume Sinks
- Flume Channels

- Flume Configuration
- What is HBase
- HBase Architecture
- Data storage in HBase
- HBase vs RDBMS
- Working with HBase

Lesson 9: Apache Pig

In this lesson you will learn about pig, components of Pig, Pig vs SQL and we will learn how to work with Pig. Topics covered are:

- What is Pig
- Components of Pig
- Pig vs SQL
- Working with Pig

Lesson 10: Basics of Apache Spark

In this lesson you will learn about apache spark, how to use spark shell, RDDs, functional programing in Spark. Topics covered are:

- What is Apache Spark?
- Using the Spark Shell
- RDDs (Resilient Distributed Datasets)

- Functional Programming in Spark

Lesson 11: RDDs in Spark

In this lesson you will learn RDD in detail and all operation associated with it, key value Pair RDD and few more other pair RDD operations. Topics covered are:

- A Closer Look at RDDs
- Key-Value Pair RDDs
- Other Pair RDD Operations

Lesson 12: Implementation of Spark Applications

In this lesson you will learn about spark applications vs spark shell, how to create a sparkcontext, building a spark application, how spark run on YARN in client and cluster mode, dynamic resource allocation and configuring spark properties. Topics covered are:

- Spark Applications vs. Spark Shell
- Creating the SparkContext
- Building a Spark Application (Scala and Java)
- How Spark Runs on YARN: Client Mode
- How Spark Runs on YARN: Cluster Mode
- Dynamic Resource Allocation
- Configuring Spark Properties

Lesson 13: Spark Parallel Processing

In this lesson you will learn about how spark run on cluster, RDD partitions, how to create partitioning on File based RDD, HDFS and data locality, parallel operations on spark, spark and stages and how to control the level of parallelism. Topics covered are:

- Spark on a Cluster
- RDD Partitions
- Partitioning of File-based RDDs
- HDFS and Data Locality
- Parallel Operations on Partitions
- Stages and Tasks
- Controlling the Level of Parallelism

Lesson 14: Spark RDD optimization techniques

In this lesson you will learn about RDD lineage, overview on caching, distributed persistence, storage levels of RDD persistence, how to choose the correct RDD persistence storage level and RDD fault tolerance. Topics covered are:

- RDD Lineage
- Caching Overview
- Distributed Persistence
- Storage Levels of RDD Persistence

- Choosing the Correct RDD Persistence Storage Level
- RDD Fault tolerance

Lesson 15: Spark Algorithm

In this lesson you will learn common spark use cases, interactive algorithms in spark, graph processing and analysis, machine learning and k-means algorithm. Topics covered are:

- Common Spark Use Cases
- Iterative Algorithms in Spark
- Graph Processing and Analysis
- Machine Learning
- Example: k-means

Lesson 16: Spark SQL

In this lesson you will learn about Spark SQL and SQL Context, creating dataframes, transforming and querying datframes and comraing spark SQL with Impala. Topics covered are:

- Spark SQL and the SQL Context
- Creating DataFrames
- Transforming and Querying DataFrames
- Comparing Spark SQL with Impala