

Spark & Scala Course Content (Spark-1.x & Spark-2.x) (Development and Administration)

Introduction to Big Data and Hadoop

❖ Big Data

- What is Big Data?
- Why all industries are talking about Big Data?
- What are the issues in Big Data?
 - Storage
 - What are the challenges for storing big data?
 - Processing
 - What are the challenges for processing big data?
- What are the technologies support big data?
 - Hadoop
 - Spark
 - Data Bases
 - Traditional
 - NOSQL

❖ Hadoop

- What is Hadoop?
- Why Hadoop?
- History of Hadoop
- Hadoop Use cases
- Advantages and Disadvantages of Hadoop
- ❖ Importance of Different Ecosystems of Hadoop
- ❖ Importance of Integration with other Big Data solutions
- ❖ Big Data Real time Use Cases
- ❖ Batch vs Real Time Big Data Analytics
- ❖ Real Time Analytics
 - Streaming Data – Storm / Kafka / Flume
 - In Memory Data - Spark

Introduction to Spark

- ❖ What is Spark
- ❖ Why Spark
- ❖ Who Uses Spark
- ❖ Brief History of Spark
- ❖ Storage Layers for Spark
- ❖ Why Spark is 100 times faster than MapReduce
- ❖ **Difference between Spark-1.x and Spark-2.x**
- ❖ **Unified Stack of Spark**
 - o Spark Core

- o Spark Sql
- o Spark Streaming
- o Spark MLLib
- o Spark GraphX
- ❖ **Spark Architecture explanation**
 - o Master Slave architecture
 - o Spark Driver
 - o Workers
 - o Executors
- ❖ **Installation of Spark in different modes**
 - o Local mode
 - o Pseudo mode
 - o Cluster mode

Basics of Spark

- ❖ Creating the **Spark Context**
- ❖ Creating the **Spark Conf**
- ❖ Creating the **Spark Session**
- ❖ Configuring **Spark Context** with **Spark Conf**
- ❖ **Caching** Overview
- ❖ Distributed Persistence
- ❖ Combine scala and java seamlessly
- ❖ Deploying Applications with **spark-submit**
- ❖ Verify spark jobs in **Spark Web UI**
- ❖ **SBT**
 - o Installing sbt
 - o Building a Spark Project with sbt
 - o Running Spark Project with sbt
- ❖ **MAVEN**
 - o Installing maven
 - o Building a Spark Project with maven
 - o Running Spark Project with maven

Resilient Distributed Dataset (RDD)

- ❖ What is RDD
- ❖ Creating RDDs
- ❖ **RDD Operations**
 - o Transformations
 - o Actions
- ❖ **Passing Functions to Spark**
 - o Python, Java, Scala

Working with Key/Value Pairs

- ❖ Creating Pair RDDs
- ❖ **Transformations on Pair RDDs**
 - o Aggregations
 - o Grouping Data
 - o Joins
 - o Sorting Data
- ❖ **Data Partitioning**
 - o Determining an RDD's Partitioner
 - o Custom Partitioners

Loading and Saving Your Data

- ❖ **File Formats**
 - o Text, json, csv, tsv, Object files
 - o Hadoop Input and Output Formats
- ❖ Loading Data using RDD
- ❖ Saving Data using RDD
- ❖ MapReduce and Pair RDD Operations
- ❖ Scala and Hadoop Integrations

Broadcast and Accumulators

- ❖ **Accumulators**
 - o Introduction to Accumulators
 - o Practical Examples on Accumulators
 - o Creating Custom Accumulators
- ❖ **Broadcast variables**
 - o Introduction to Broadcast variables
 - o Practical Examples on Broadcast variables
 - o Optimizing Broadcasts

Working with Spark in different programming languages

- ❖ **Python**
 - o Installing Python
 - o How to use 'pyspark'
 - o Practical examples on spark in python
- ❖ **Scala**
 - o Installing Scala
 - o How to use 'spark-shell'
 - o Practical examples on spark in Scala
- ❖ **Java**
 - o Installing Java
 - o How to use 'Java'
 - o Practical examples on spark in Java

- ❖ **R**
 - o Installing R
 - o How to use '**SparkR**'
 - o Practical examples on spark in R

Apache Spark SQL

- ❖ Spark SQL & Hive Architecture explanation
- ❖ Working with Spark SQL **DataSets**
- ❖ Working with Spark SQL **DataFrames**
- ❖ Practice on Spark **SQL Context**
- ❖ Practical examples on **Spark SQL**
- ❖ **Integrating Spark SQL with**
 - o Hive
 - o Phoenix
 - o Cassandra
 - o RDBMS
- ❖ **Processing different files using Spark**
 - o Text
 - o Json
 - o Csv
 - o Tsv
 - o Parquet
- ❖ Spark SQL **UDFs**
- ❖ Spark SQL **Performance Tuning** Options
- ❖ JDBC/ODBC Server

Apache Spark Streaming

- ❖ Spark Streaming Architecture explanation
- ❖ Creating the **Streaming Context**
- ❖ Discretized Streams (**DStreams**)
- ❖ **Transformations on Dstreams**
 - o UpdateStateByKey Operation
 - o Transform Operation
 - o Window Operations
 - o Join Operations
- ❖ Output Operations on DStreams
- ❖ Streaming UI explanation
- ❖ **Spark Streaming Sources**
 - o Basic Sources
 - o Advanced Sources

- ❖ **Integrating Spark Streaming with**
 - o Flume
 - o Kafka
 - o Twitter
 - o HDFS
- ❖ Performance Considerations
- ❖ Practical examples on Spark Streaming

Apache Spark MLlib

- ❖ Machine Learning Basics
- ❖ **Machine Learning Algorithms**
 - o Classification
 - o Clustering
 - o Collaborative Filtering
- ❖ Performance Considerations
- ❖ Practical examples on Spark MLlib

Apache Spark Graphx

- ❖ Introduction to Spark Graphx
- ❖ Practical Examples on Spark Graphx

Apache Mesos

- ❖ Introduction to Apache Mesos
- ❖ Apache Mesos Architecture explanation
- ❖ Practical Examples on Apache Mesos

Apache Mahout

- ❖ Introduction to Apache Mahout
- ❖ Apache Mahout Architecture explanation
- ❖ Practical Examples on Apache Mahout

Apache Storm

- ❖ Introduction to Apache Storm
- ❖ Apache Storm Architecture explanation
- ❖ Practical Examples on Apache Storm

Apache Kafka

- ❖ Introduction to Apache Kafka
- ❖ Installing Apache Kafka
- ❖ Apache Kafka Architecture explanation
- ❖ Practical Examples on Apache Kafka



*Mr. Kalyan, Big Data Solution Architect,
Apache Contributor, 11+ years of IT exp, 7+ years of Big Data exp,
Cloudera CCA175 Certified Consultant, IIT Kharagpur, Gold Medalist*

Apache FLUME

- ❖ Introduction to flume
- ❖ Flume installation
- ❖ Flume Architecture
 - o Agent
 - o Sources
 - o Channels
 - o Sinks
- ❖ Practical Examples on Flume

Apache Phoenix

- ❖ Introduction to Phoenix
- ❖ Installing Phoenix
- ❖ Integrating with Hbase
- ❖ Practical Examples on Phoenix

Apache Cassandra

- ❖ Introduction to Cassandra
- ❖ Installing Cassandra
- ❖ Practical Examples on Cassandra

Apache Drill

- ❖ Introduction to Drill
- ❖ Installing Drill
- ❖ Practical Examples on Drill

Apache Zeppelin

- ❖ Introduction to Zeppelin
- ❖ Installing Zeppelin
- ❖ Practical Examples on Zeppelin
- ❖ Data Visualization using Zeppelin

Play Framework

- ❖ Introduction to Play Framework
- ❖ Installing Play Framework
- ❖ Practical Examples on Play Framework
- ❖ Spark Project using Play Framework



*Mr. Kalyan, Big Data Solution Architect,
Apache Contributor, 11+ years of IT exp, 7+ years of Big Data exp,
Cloudera CCA175 Certified Consultant, IIT Kharagpur, Gold Medalist*

Introduction of Scala

- ❖ What is Scala?
- ❖ Why Scala?
- ❖ Advantages of Scala?
- ❖ Using the Scala REPL(Read Evaluate print loop)
- ❖ What is Type Inference
- ❖ Interoperability between Scala and Java

Scala using Command Line

- ❖ Installing Java & Scala
- ❖ Interactive Scala
- ❖ Writing Scala Scripts
- ❖ Compiling Scala Programs

Basics of Scala

- ❖ Defining Variables
- ❖ Defining Functions
- ❖ String Interpolation
- ❖ IDE for Scala

Scala Type Less, Do More

- ❖ Semicolons
- ❖ Variable Declarations
- ❖ Method Declarations
- ❖ Type Inference
- ❖ Immutability
- ❖ Reserved Words
- ❖ Operators
- ❖ Precedence Rules
- ❖ Literals
- ❖ Options
- ❖ Arrays, Lists, Ranges, Tuples

Expressions and Conditionals

- ❖ If expressions
- ❖ If-Else expressions
- ❖ Match Expressions
- ❖ For Loops
- ❖ While Loops
- ❖ Do-While Loops
- ❖ Conditional Operators



*Mr. Kalyan, Big Data Solution Architect,
Apache Contributor, 11+ years of IT exp, 7+ years of Big Data exp,
Cloudera CCA175 Certified Consultant, IIT Kharapur, Gold Medalist*

- ❖ Enumerations
- ❖ Pattern Matching
- ❖ Using try, catch, and finally Clauses

Functional Programming in Scala

- ❖ What is Functional Programming?
- ❖ Functional Literals and Closures
- ❖ Recursions
- ❖ Tail Calls
- ❖ Currying
- ❖ Functional Data Structures
- ❖ Sequences, Maps, Sets
- ❖ Traversing
- ❖ Traversal, Mapping, Filtering, Folding and Reducing
- ❖ Implicit Function Parameters
- ❖ Call by Name, Call by Value

Object-Oriented Programming in Scala

- ❖ Class and Object Basics
- ❖ Value Classes
- ❖ Parent Types
- ❖ Constructors in Scala
- ❖ Fields in Classes
- ❖ Nested Types
- ❖ Traits as Mixins
- ❖ Stackable Traits
- ❖ Creating Traits
- ❖ Visibility Rules

Scala for Big Data

- ❖ Improving MapReduce with Scala
- ❖ Moving Beyond MapReduce
- ❖ Categories for Mathematics
- ❖ A List of Scala-Based Data Tools

Spark with Big Data Integrations:

- o Spark and Hive integration
- o Spark and Phoenix integration
- o Spark and Cassandra integration
- o Spark and Flume integration
- o Spark and Kafka integration
- o Spark and RDBMS integration

Real Time Big Data Projects

- ❖ We will be sharing **End-to-End Big Data Projects**
- ❖ We are providing **Big Data Project Practice on Our Lab**
- ❖ We are providing **Important Recorded Videos on Our YouTube Channel**
- ❖ Any information search in **Google / YouTube** by keyword is '**Kalyan Hadoop**'

Spark Administration topics:

- ❖ Hadoop Installation
- ❖ Hive Installation
- ❖ Hbase Installation
- ❖ Zookeeper Installation
- ❖ Phoenix Installation
- ❖ Kafka Installation
- ❖ Flume Installation
- ❖ Zeppelin Installation
- ❖ Play Framework Installation
- ❖ MySql Installation
- ❖ Java Installation
- ❖ Scala Installation
- ❖ Python Installation
- ❖ R Installation
- ❖ Eclipse Installation
- ❖ Cloudera Distribution installation

Free Big Data Workshops:

- Spark & Scala
- Cassandra
- MongoDB
- Search engine & E-commerce solutions
- Big Data Analytics (R, Mahout, Spark ML)

What we are offering to you:

- ✓ Hands on Practice on **Cloudera CCA175 Spark and Hadoop Developer Certification**
- ✓ Tips to **Crack the CCA175 Certification**
- ✓ Hands on Practice on **Spark & Scala Real-Time Examples**
- ✓ Providing 1 **Major project** on Spark.
- ✓ Providing 2 **Mini projects** on Spark.
- ✓ **Real Time Big Data projects will be shared**
- ✓ **Free Big Data Workshops on new & advanced technologies**
- ✓ **Free Weekly Online Hadoop Certification**
- ✓ Hands on installation Spark and it's relative software's in your laptop.
- ✓ Well documented Spark & Scala material with all the topics covering in the course.
- ✓ Well documented Spark blog contains frequent interview questions along with the answers and latest updates on BigData technology.
- ✓ Discussing about Spark & Scala interview questions daily base.
- ✓ **Resume preparation** with POC's or Project's based on your experience.