**Spark & Scala Course Content (Spark-1.x & Spark-2.x)**

# 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

# Scala Basics

* Functional language
* Scala Vs Java
* Strings, Numbers
* List, Array, Map, Set
* Control Statements, collections
* Functions, methods
* Pattern matching

# 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**
* AWS with Spark
* RDS
* EC2
* IAM
* EMR
* REDSHIFT
* **Unified Stack of Spark**
* Spark Core
* Spark Sql
* Spark Streaming
* Spark MLLib
* Spark GraphX
* **Spark Architecture explanation**
* Master Slave architecture
* Spark Driver
* Workers
* Executors
* **Installation of Spark in different modes**
* Local mode
* Pseudo mode
* 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**
  + Installing sbt
  + Building a Spark Project with sbt
  + Running Spark Project with sbt
* **MAVEN**
  + Installing maven
  + Building a Spark project with maven
  + Running Spark project with maven

# Resilient Distributed Dataset (RDD)

* What is RDD
* Creating RDDs
* **RDD Operations**
* Transformations
* Actions

# Working with Key/Value Pairs

* Creating Pair RDDs
* **Transformations on Pair RDDs**
* Aggregations
* Grouping Data
* Joins
* Sorting Data
* **Data Partitioning**
* Determining an RDD’s Partitioner
* Custom Partitioners

# Loading and Saving Your Data

* **File Formats**
* Text, json, csv, tsv, Object files
* 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**
* Introduction to Accumulators
* Practical Examples on Accumulators
* Creating Custom Accumulators
* **Broadcast variables**
* Introduction to Broadcast variables
* Practical Examples on Broadcast variables
* Optimizing Broadcasts

# 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**
* Spark SQL and DataFrame Uses
* DataFrame / SQL APIs
* Catalyst Query Optimization
* Creating (CSV, JSON) DataFrames
* Querying with DataFrame API and SQL
* Caching and Re-using DataFrames
* Process Hive data in Spark
* Power of Dataset API in Spark
* Serialization concept in DataSet
* Creating DataSet API
* Process CSV, JSON, XML, Text data
* DataSet Operation
* **Integrating Spark SQL with**
* Hive
* Phoenix
* RDBMS
* 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**
* UpdateStateByKey Operation
* Transform Operation
* Window Operations
* Join Operations
* Output Operations on DStreams
* Streaming UI explanation
* **Spark Streaming Sources**
* Basic Sources
* Advanced Sources
* **Integrating Spark Streaming with**
* Kafka
* Performance Considerations
* Practical examples on Spark Streaming

# Apache Kafka

* Introduction to Apache Kafka
* Installing Apache Kafka
* Apache Kafka Architecture explanation
* Practical Examples on Apache Kafka

# Apache Phoenix

* Introduction to Phoenix
* Installing Phoenix
* Integrating with Hbase
* Practical Examples on Phoenix

# Apache Zeppelin

* Introduction to Zeppelin
* Installing Zeppelin
* Practical Examples on Zeppelin
* Data Visualization using Zeppelin

# Real Time POC-pipeline using Hadoop, Kafka, Spark, Scala, Zeppelin

# Real Time POC-pipeline using Hadoop, Sqoop, Hive, Spark, Scala, Zeppelin

**Important Points:-**

* Will give whatsapp support.
* After training will assign a task and 2 questions on CCA 175 Cloudera Spark and Hadoop Developer
* Minimum 3 months support after training through whatsapp.