Apache Spark and Scala Advance Course

EduNextgen

Education for Next Generation Leaders

Course Durations: 30 Hours **Course Mode: Online/Offline**

About Company:

EduNextgen extended arm of Product Innovation Academy is a growing entity in education and career transformation, specializing in today's most in-demand skills. A platform with blended learning programs supported by in-trend technology platforms for learning. Engaging organizations for learning development objectives.

Training courses are designed and updated by renowned industry experts. Our blended learning approach combines online classes, instructor-led live virtual classrooms and virtual teaching assistance.

About The Course:

With businesses generating Big Data at a rapid pace, analyzing the data to leverage meaningful business insights is the need of the hour. The demand for Analytics Skill is going up steadily but there is a huge deficit on the supply side. In spite of Big Data Analytics being a 'Hot' job, there is still a large number of unfilled jobs across the globe due to shortage of required skill. A McKinsey Global Institute study states that the US will face a shortage of about 190,000 Data Scientists and 1.5 million Managers and Analysts who can understand and make decisions using Big Data by 2018.

This is an extensive program designed to cover most important modules of Big Data required by today's Industry and as well help you achieve Certifications from Hortonworks and Cloudera. The program is bundled with modules like Spark Components i.e. RDD, SparkSQL, MLlib, Spark Streaming GraphX. And to support your learning path there are resources like our Cloud Labs, Industry Grade Projects, Assignments, Use Cases and more with our world class Technical Support post learning.

The course curriculum and contents are made by industries expert and curriculum covers based of Cloudera and Hotanwork Certification

Why This Course:

- Cover Apache Spark Components i.e. Scala, RDD, SparkSQL, Spark Streaming
- Hands-on Experience and Use Case
- Project Execution in Different Domain Data Sets and Components i.e. RDD, SBT and SparkSQL
- Pre-Installed Spark Environment (Plug and Play)
- Cloud Lab
- Live Support (24×7)

Participants will get Access to:

- LMS Access
- Cloud Lab
- 50+ Assignments
- 100+ Quizzes
- Pre-Installed Spark Environment (Plug and Play)
- 5+ Industry Grade Projects in Different Domain
- Live Support via one to one Screen Sharing, Mail and Call
- Course Completion Certificate

Batch Schedule:

Weekend: 3 Hours per day (Online), 4 Hours per day (Offline)

Weekday: 2 Hours per day (Online), 2 Hours per day (Offline)

Course Curriculum

Module 1: Introduction to Big Data & Apache Spark (4 Hours)

This module will help you to understand about the Big Data and Apache Spark. This module will cover the Big Data Characteristics, Types of Data, Big Data Application, Apache Spark and Its Components. Also this module will cover Spark Advantages and Impotence. Below topics are covered in this module:

- What is Big Data?
- Big Data Growth
- Why Need of More Data?
- Big Data Impacts
- Big Data Characteristics (5V's)
- Types of Data
- Industries who are making the Most of Big Data
- Big Data Applications
- Traditional Model and its Problems
- Introduction of Apache Spark
- Why Spark?
- Spark: Characteristics, Ecosystem/Components and Version
- Spark Advantages
- Spark vs Hadoop
- How Spark work with HDFS and Hive
- Spark Shell and REPL
- Scala IDE

Hands on/Programs/Practical:

- Spark Installation
- How to Start Spark Shell
- How to work with Cloud Lab
- Execute Word Count Program using Spark and Hadoop

Module 2: YARN and HDFS Architecture (3 Hours)

This module will help you to understand YARN and HDFS Architecture. Also, this module will cover Cluster, Nodes, Racks and HDFS Hands on. Below topics are covered in this module:

- What is Cluster?
- Details about: Blocks, Namenodes, Datanodes, Secondary NameNode

- Introduction to HDFS
- HDFS Federation
- HDFS High Availability
- Racks and Replication Factor
- Rack Awareness
- HDFS File Write Anatomy
- HDFS File Read Anatomy
- HDFS Architecture
- Hadoop 1 vs Hadoop 2
- YARN
- How YARN Runs an Application

Hands on/Programs/Practical:

- HDFC Commands: ls, mkdir, cat, put etc.
- Blocks
- Replication

Module 3: Deep Dive on Scala (6 Hours)

Apache Spark develop on Scala Language, so learning Scala is good compare with other language, Spark supports Java, Python and R apart from Scala. This module is divide into lessons and each lesions will help you deep understanding about Scala.

Hands on/Programs/Practical:

***All Lesson based on Practical. Code Execution will be for each topics.**

Lesson 1: Deep Dive on Scala – I

- Introduction to Scala
- Why Scala?
- Scala Frameworks
- Use of Scala
- Scala in the Enterprise
- Scala REPL - Read Evaluate Print Loop
- Data Types
- Scala Variables and Operators
- Functions and Lambdas
- Scala Statements and Loops
- val, var and def

Lesson 2: Deep Dive on Scala – II

- Class and Object
- Case Class
- Access Modifier
- Singletons Object

- Companion Objects
- Inheritance
- Packages and Package Objects
- Traits
- Exception Handling

Lesson 3: Deep Dive on Scala – III

- What is Functional Programming?
- Higher Order Functions
- Anonymous Functions
- Currying
- Closures

Module 4: Introduction to RDDs (4 Hours)

This module will help you to understand RDD and its Operation. Also this module cover RDD, its Feature, SparkContext, RDD Function, Transformations, Actions and Join Operations. Below topics are covered in this module:

- Introduction to RDDs
- Why RDD?
- RDDs Features
- SparkContext(sc)
- Data Loading into RDDs
- Key value pair and MapReduce
- RDDs Functions
- Transformation and Actions
- Join Operations

Hands on/Programs/Practical:

- How to Start Spark Shell
- Work Count Program using RDD
- How to Load Data from Hadoop and Local
- Transformations
- Actions
- Join Operations
- Project Execution in Entertainments Domain via RDD

Module 5: Introduction to SparkSQL (4 Hours)

This module will help you to understand working with SparkSQL. Also this module cover SparkSQL, Dataframe, Data Loading and Join Operations. Below topics are covered in this module:

- Introduction to SparkSQL
- Spark SQL Context

- Dataframe in SparkSQL
- Data Loading Technique in SparkSQL
- Working with Hive Context and JDBC
- Working with Parquet, JSON and CSV
- RDD vs SparkSQL

Hands on/Programs/Practical:

- How to Create SQLContext(sc)
- Data Loading into SparkSQL from Local and Hadoop
- Run the Queries
- Specifying the Schema
- How to Create Dataframe
- Reading JSON file and Creating Data Frames
- Reading CSV files with header while starting the Spark Shell
- Join Operations
- Project Execution in Retails Domain via SparkSQL

Module 6: Scala Build Tool (SBT) (1 hour)

This module will help you to understand working with Scala Build Tool (SBT). Also this module cover SBT, Installation, Maven and SBT Command. Below topics are covered in this module:

- Introduction to SBT
- SBT Installation
- SBT vs Maven
- SBT Commands
- Build SBT Project

Hands on/Programs/Practical:

- SBT Installation
- Create SBT Projects

Module 7: Introduction to Kafka and Spark Streaming (3 Hours)

This module will help you to understand working with Kafka and Spark Streaming. Also this module cover Kafka Architecture, Producer, Consumer, Topics, Broker etc. Below topics are covered in this module:

- Introduction to Spark Kafka
- Use of Kafka
- Kafka Architecture
- Understanding Producer, Consumer, Topics and Broker
- Replication and Partitions of Topics
- Installation and Configuration
- Spark Streaming Architecture and Abstraction

- DStreams
- Transformations
- Streaming UI
- Connecting with Kafka
- Streaming Exercise

Hands on/Programs/Practical:

- Kafka Installation
- How to work with Producer, Consumer, Topics and Broker
- Word Count Program with Spark Streaming

Module 8: Introduction to MLlib and GraphX (3 Hours)

This module will help you to understand working with MLlib and GraphX. Also this module cover what is MLlib, Graph, Graph Operator, etc. Below topics are covered in this module:

- Introduction to MLlib
- Different Algorithms and k-Means
- Running a Spark MLlib Example
- Enabling Native Acceleration for MLlib
- Introduction to Graph-Parallel
- GraphX Components, Construction and Background
- Data Parallel, Graph-Parallel, and RDDs tie in with GraphX
- Visualizing Spark GraphX
- Exploring Graph Operators
- GraphX handles Visualizations
- Create views and look alternative options

Hands on/Programs/Practical:

- How to execute MLlib program
- How to work with GraphX
- GraphX presentation

Module 9: Advance Project Execution using Spark (2 Hours)

Understand Project: What is the Project, Dataset, Use Case Execution, Output

Hands on/Programs/Practical:

- Step by step Project Execution