



APACHE SPARK AND SCALA TRAINING

COURSE DESIGN

High-quality videos, slides, hands-on examples, quizzes, automated assessments, case studies, and real-world projects.

COURSE MATERIAL

Lifetime access to cutting-edge self-paced learning content.

LAB

90 Days of [CloudxLab](#) access for hands-on practice.

SUPPORT

Email support to answer your queries and we've also launched [Discussions](#) - a Q&A site for Artificial Intelligence, Machine Learning, Deep Learning, Big Data & Data Science professionals.

CERTIFICATE

Earn certificate in Apache Spark and Scala Training.

LIVE SESSIONS

30+ hours of live online instructor-led training. Classes will be conducted every Saturday & Sunday between (8 PM - 11 PM Indian Standard Time) or (7:30 AM - 10:30 AM Pacific Time).

BIG DATA WITH HADOOP & SPARK - COURSE SYLLABUS

INTRODUCTION

- What is Big Data?
- Why Now?
- Big Data Use Cases
- Various Solutions
- Overview of Hadoop Ecosystem
- Spark Ecosystem Walkthrough
- Quiz

FOUNDATION & ENVIRONMENT

- Understanding the CloudxLab
- CloudxLab Hands-On
- Spark Hands-on
- Quiz and Assessment
- Basics of Linux - Quick Hands-On
- Understanding Regular Expressions
- Quiz and Assessment
- Setting up VM (optional)

DATA FORMATS & MANAGEMENT

- InputFormat and InputSplit
- JSON
- XML
- AVRO
- How to store many small files - SequenceFile?
- Parquet
- Protocol Buffers
- Comparing Compressions
- Understanding Row Oriented and Column Oriented Formats - RCFile?

SCALA BASICS

- Introduction to Scala?
- Accessing Scala using CloudxLab
- Getting Started: Interactive, Compilation, SBT
- Types, Variables & Values
- Functions
- Collections
- Classes
- Parameters
- More Features
- Quiz and Assessment

SPARK BASICS

- What is Apache Spark?
- Why Spark?
- Using the Spark Shell on CloudxLab
- Example 1 - Performing Word Count
- Understanding Spark Cluster Modes on YARN
- RDDs (Resilient Distributed Datasets)
- General RDD Operations: Transformations & Actions
- RDD Lineage
- RDD Persistence Overview
- Distributed Persistence

WRITING AND DEPLOYING SPARK APPLICATIONS

- Creating the SparkContext
- Building a Spark Application (Scala, Java, Python)
- The Spark Application Web UI
- Configuring Spark Properties
- Running Spark on Cluster
- RDD Partitions
- Executing Parallel Operations
- Stages and Tasks
- Project: Churning the logs of NASA Kennedy Space Center WWW server

COMMON PATTERNS IN SPARK DATA PROCESSING

- Common Spark Use Cases
- Example 1 - Data Cleaning (Movielens)
- Example 2 - Understanding Spark Streaming
- Understanding Kafka
- Example 3 - Spark Streaming from Kafka
- Iterative Algorithms in Spark
- Project: Real-time analytics of orders in an e-commerce company

DATAFRAMES AND SPARK SQL

- Spark SQL and the SQL Context
- Creating DataFrames
- Transforming and Querying DataFrames
- Saving DataFrames
- DataFrames and RDDs
- Comparing Spark SQL, Impala, and Hive-on-Spark

MACHINE LEARNING WITH SPARK

- GraphX: Graph Processing and Analysis
- Understanding Machine Learning



- MLlib Example: k-means
- SparkR Example

OTHER Topics/Content

- Java Essential
- Linux Basics
- Spark On Cluster
- Adv Spark Programming
- Hands-on videos

PROJECTS INCLUDED

- Building Real-Time Analytics Dashboard
- Churning the logs of NASA Kennedy Space Center WWW server
- Generating movie recommendations using Spark MLlib
- Deriving the importance of various Handles at Twitter using Spark GraphX
- Write end-to-end Spark application starting from writing code on your local machine to deploying to the cluster

The Big Data with Spark course is compatible with the following certifications:

- Hortonworks Certified Developer (HDPCD): Spark

[Click Here To Enroll Now!!](#)

Please feel free to email your queries to reachus@cloudxlab.com

Regards,
The CloudxLab Team