

BIG DATA WITH SPARK

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.

LIVE SESSIONS RECORDINGS

45+ hours of long duration videos from the industry expert to help you with the course content.

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 a certificate in Big Data with Spark.

BIG DATA WITH SPARK - COURSE SYLLABUS

INTRODUCTION

- What is Big Data?
- Why Now?
- Big Data Use Cases
- Various Solutions
- 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)

HDFS and Yarn Introduction

- Why HDFS or Why not existing file systems?
- Understanding the architecture
- Quiz
- Advance HDFS Concepts (HA, Federation)
- Quiz
- Hands-on with HDFS (Upload, Download, SetRep)
- Quiz & Assessment
- Data Locality (Rack Awareness)

- Computing - Why not existing

tools?

- MapReduce 1.0
- Resource Management: YARN Architecture
- Advance Concepts - Speculative Execution
- Quiz

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

1. Generate movie recommendations using Spark MLlib
2. Derive the importance of various handles at Twitter using Spark GraphX
3. Churn the logs of NASA Kennedy Space Center WWW server using Spark to find out useful business and devops metrics
4. Write end-to-end Spark application starting from writing code on your local machine to deploying to the cluster
5. Build real-time analytics dashboard for an e-commerce company using Apache Spark, Kafka, Spark Streaming, Node.js, Socket.IO and Highcharts

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

- Cloudera Certified Professional (CCP): Data Engineer

- Cloudera Certified Associate (CCA): Spark Developer
- Hortonworks Certified Developer (HDPCD): Spark

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

Regards, The
CloudxLab Team