Certification Course in

Big Data Engineering with Hadoop & Spark

Online Self-paced Course | 60+ Hours of Training

About CloudxLab

At Cloudxlab, we are building one of the best gamified learning environments to make technology learning fun and for life. More than 50,000 users across the world have been benefited by our signature courses on Machine Learning and Big Data. Our vision is to upskill people on high-end technologies like Deep Learning, Machine Learning, Big Data and make them employable.

As humans, we are immersed in data in our everyday lives. As per IBM, the data doubles every two years on this planet. The value that data holds can only be understood when we can start to identify patterns and trends in the data. Normal computing principles do not work when data becomes huge.

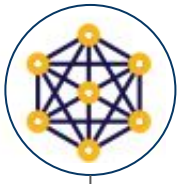
There is massive growth in the big data space, and job opportunities are skyrocketing, making this the perfect time to launch your career in this space. In this course, you will learn Hadoop and Spark to drive better business decisions and solve real-world problems.

Sandeep Giri

Founder at CloudxLab

Why CloudxLab

Earn a Verified Certificate from CloudxLab



Learn Big Data with Hadoop and PySpark from industry experts and become expert in Big Data domain

Online cloud lab for hands-on for real-world experience

Best-in-class support Throughout your learning journey

Lifetime course access

Work on real-world projects.

Interact with the international community of peers via the discussion forum.

Course Creators

Sandeep Giri

Founder at CloudxLab

Past: Amazon, InMobi, D.E.Shaw

Course Developer

Know More

Abhinav Singh

Co-Founder at CloudxLab

Past: Byjus

Course Developer

Know More

Jatin Shah

Ex-LinkedIn, Yahoo,

Yale CS Ph.D. IIT-B

Course Advisor

Know More

Praveen Pavithran

Co-Founder at Yatis

Past: YourCabs, Cypress Semiconductor

Course Advisor

Know More

Course Curriculum

Course 1: Big Data with Hadoop

1. Introduction

- Big Data Introduction
- Distributed systems
- Big Data Use Cases
- Various Solutions
- Overview of Hadoop Ecosystem
- Spark Ecosystem Walkthrough

2. Foundation & Environment

- Understanding the Cloudxlab
- Cloudxlab Hands-on
- Hadoop & Spark Hands-on
- Basics of Linux - Quick Hands-on
- Understanding Regular Expressions

3. Zookeeper

- ZooKeeper - Race Condition
- ZooKeeper - Deadlock
- Hands-On
- How does election happen - Paxos Algorithm?
- Use cases
- When not to use

Course Curriculum

Course 1: Big Data with Hadoop

4. HDFS

- Why HDFS or Why not existing file systems?
- HDFS - NameNode & DataNodes
- Advance HDFS Concepts (HA, Federation)
- Hands-on with HDFS (Upload, Download, SetRep)
- Data Locality (Rack Awareness)

5. YARN

- YARN - Why not existing tools?
- YARN - Evolution from MapReduce 1.0
- Resource Management: YARN Architecture
- Advance Concepts - Speculative Execution

6. MapReduce Basics

- MapReduce - Understanding Sorting
- MapReduce - Overview & Quiz
- Example 0 - Word Frequency Problem - Without MR
- Example 1 - Only Mapper - Image Resizing
- Example 2 - Word Frequency Problem
- Example 3 - Temperature Problem
- Example 4 - Multiple Reducer
- Example 5 - Java MapReduce Walkthrough & Quiz

Course Curriculum

Course 1: Big Data with Hadoop

7. MapReduce Advanced

- Writing MapReduce Code Using Java
- Building MapReduce project using Apache Ant
- Concept - Associative & Commutative
- Example 8 - Combiner
- Example 9 - Hadoop Streaming
- Example 10 - Adv. Problem Solving - Anagrams
- Example 11 - Adv. Problem Solving - Same DNA
- Example 12 - Adv. Problem Solving - Similar DNA
- Example 12 - Joins - Voting
- Limitations of MapReduce

8. Analyzing Data with Pig

- Pig - Introduction
- Pig - Modes
- Getting Started
- Example - NYSE Stock Exchange
- Concept - Lazy Evaluation

Course Curriculum

Course 1: Big Data with Hadoop

9. Processing Data with Hive

- Hive - Introduction
- Hive - Data Types
- Getting Started
- Loading Data in Hive (Tables)
- Example: Movielens Data Processing
- Advance Concepts: Views
- Connecting Tableau and HiveServer 2
- Connecting Microsoft Excel and HiveServer 2
- Project: Sentiment Analysis of Twitter Data
- Advanced - Partition Tables

10. NoSQL and HBase

- NoSQL - Scaling Out / Up
- NoSQL - ACID Properties and RDBMS Story
- CAP Theorem
- HBase Architecture - Region Servers etc
- Hbase Data Model - Column Family Orientedness
- Getting Started - Create table, Adding Data
- Adv Example - Google Links Storage
- Concept - Bloom Filter
- Comparison of NOSQL Databases

Course Curriculum

Course 1: Big Data with Hadoop

11. Importing Data with Sqoop and Flume, Oozie

- Sqoop - Introduction
- Sqoop Import - MySQL to HDFS
- Exporting to MySQL from HDFS
- Concept - Unbounding Dataset Processing or Stream Processing
- Flume Overview: Agents - Source, Sink, Channel
- Example 1 - Data from Local network service into HDFS
- Introduction to Oozie

Course Curriculum

Course 2: Big Data with Spark

1. Introduction

- Apache Spark ecosystem walkthrough
- Spark Introduction - Why Spark?

2. Scala Basics

- Scala - Quick Introduction - Access Scala on CloudxLab
- Scala - Quick Introduction - Variables and Methods
- Getting Started: Interactive, Compilation, SBT
- Types, Variables & Values
- Functions
- Collections
- Classes
- Parameters
- More Features

3. Spark Basics

- Apache Spark ecosystem walkthrough
- Spark Introduction - 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

Course Curriculum

Course 2: Big Data with Spark

- RDD lineage
- RDD Persistence Overview
- Distributed Persistence.

4. 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

5. 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

Course Curriculum

Course 2: Big Data with Spark

6. Data Formats and 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?

7. DataFrames and Spark SQL

- Spark SQL - Introduction
- Spark SQL - Dataframe Introduction
- Transforming and Querying DataFrames
- Saving DataFrames
- DataFrames and RDDs
- Comparing Spark SQL, Impala, and Hive-on-Spark

8. Machine Learning with Spark

- Machine Learning Introduction
- Applications Of Machine Learning
- MILib Example: k-means
- SparkR Example

Projects

1. Sentiment analysis

Sentiment analysis of "Iron Man 3" movie using Hive and visualizing the sentiment data using BI tools such as Tableau

2. Process the NYSE

Process the NYSE (New York Stock Exchange) data using Hive for various insights

3. MovieLens Project

Analyze MovieLens data using Hive

4. Spark MLlib

Generate movie recommendations using Spark MLlib

5. Spark GraphX

Derive the importance of various handles at Twitter using Spark GraphX

6. Churn the logs

Churn the logs of NASA Kennedy Space Center WWW server using Spark to find out useful business and devops metrics

7. Spark application

Write end-to-end Spark application starting from writing code on your local machine to deploying to the cluster

8. Analytics Dashboard

Real-time analytics dashboard for an e-commerce company using Apache Spark, Kafka, Spark Streaming, Node.js, Socket.IO and Highcharts

Course Details and Fees

Please find more information about the course and fees here:
https://cloudxlab.com/course/1/big-data-with-hadoop-and-spark-scala

Mode of Learning

Online Self-Paced Learning

Our Esteemed Customers

For Further Details

Contact us at +080-4920-2224 or +1 412-568-3901 or contact:

For Business

For corporate training and bulk enrollments, write to us at reachus@cloudxlab.com

Headquarters - United States

2035, Sunset Lake Road Suite B-2, 19702
Newark, New Castle
Delaware, United States

R&D Center - India

Issimo Technology Private Limited
#215, Arcade, Brigade Metropolis,
Mahadevpura, Bangalore, India - 560 048