CLOUD x LAB

CERTIFICATE PROGRAM

Big Data Engineering with PySpark

Online Self Paced Course | 60+ Hours of Training

CloudxLab & Course

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.

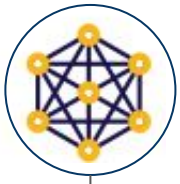
There is massive growth in the big data space, and job opportunities are skyrocketing, making this the perfect time to launch your career in this space. In this course, you will learn Hadoop and Spark to drive better business decisions and solve real-world problems.

Sandeep Giri

Founder at CloudxLab

Why CloudxLab

Earn a Verified Certificate from CloudxLab



Learn Big Data with Hadoop and Spark from industry experts and become expert in Data Science 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
- Quiz

2. Foundation & Environment

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

Course Curriculum

Course 1: Big Data with Hadoop

3. Zookeeper

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

4. HDFS

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

Course Curriculum

Course 1: Big Data with Hadoop

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

7. Map Reduce Advanced

- Writing MapReduce Code Using Java
- Building MapReduce project using Apache Ant
- Concept - Associative & Commutative
- Quiz
- 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
- Quiz

Course Curriculum

Course 1: Big Data with Hadoop

8. Analyzing Data with Pig

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

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
- Understanding HCatalog & Impala
- Quiz

Course Curriculum

Course 1: Big Data with Hadoop

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
- Quiz

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
- Example 2 - Extracting Twitter Data
- Quiz
- Example 3 - Creating workflow with Oozie

Course Curriculum

Course 2: Big Data with PySpark

1. Introduction

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

2. Apache Spark Basics with Python

- Cluster installation
- RDDs (Resilient Distributed Datasets)
- Quiz and Assessment

3. Apache Spark - Key Value RDD

- Using the Spark Shell on CloudxLab
- Quiz
- Key-value RDDs with python - reducebyKey
- Key-value RDDs with python - Computing Maximum Tempreture
- Key-value RDDs with python - Advanced Topics
- Key-value RDDs with python - project - Handling binary files

4. Spark streaming with Python

- Spark streaming with Python - Introduction
- Spark streaming with Python - Dstream
- Spark streaming with Python - Use Cases
- Quiz

Course Curriculum

Course 2: Big Data with PySpark

- Spark streaming with Python - Understand master URL
- Apache Kafka Introduction
- Integrating Apache Spark Streaming & Apache Kafka
- Apache Spark Streaming - updateStateByKey Operation
- Apache Spark Streaming - Transform and Window Operations
- Apache Spark Streaming - Join and Output Operations

5. Spark on cluster with Python

- Apache Spark - Running On Cluster - Architecture
- Apache Spark - Running On Cluster - Launching
- Apache Spark - Running On Cluster - Local Mode (Python)
- Apache Spark - Running On Cluster - Cluster Mode - Standalone
- Apache Spark - Running On Cluster - Cluster Mode - YARN
- Apache Spark - Running On Cluster - Cluster Mode - Mesos+AWS
- Apache Spark - Running On Cluster - Deployment Modes
- Apache Spark - Running On Cluster - Slides

6. Advanced Spark Programming with Python

- Adv Spark Programming - Understanding Persistence (Python)
- Adv Spark Programming - Persistence StorageLevel (Python)
- Adv Spark Programming - Data Partitioning
- Adv Spark Programming - Partitioning HandsOn (Python)
- Adv Spark Programming - Data Partitioning Example
- Adv Spark Programming - Custom Partitioner (Python)
- Adv Spark Programming - Hardware Provisioning

Course Curriculum

Course 2: Big Data with PySpark

- Adv Spark Programming - Memory Management
- Adv Spark Programming - Serialization Format
- Writing Spark Application with Python Project

7. DataFrames and Spark SQL

- Spark SQL - Introduction
- Spark SQL - Dataframe Introduction
- Transforming and Querying DataFrames
- Saving DataFrames
- DataFrames and RDDs

8. Machine Learning with Spark

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

9. Graph Processing in Spark

- GraphX Quick Walkthrough
- Graphx Example
- Resources

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 NSE

Process the NSE (National 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. Churn the logs

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

6. Spark application

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

7. 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/152/big-data-with-pyspark-incl-hadoop-spark-and-python

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:

Aswath Madhu
Program Director

programs@cloudxlab.com

Prakhar Katiyar
Chief Admissions Counsellor

admissions@cloudxlab.com

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