CLOUD x LAB

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 (7 AM - 10 AM Indian Standard Time) or (6:30 PM - 09:30 PM Pacific Time).

CLOUD x LAB

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)

RECAP OF HDFS AND YARN

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
- Learn operations on Key-Value Based RDD
- Solving various problems using RDD

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

SPARK ADVANCED OPERATIONS

- Using Accumulators & Creating Custom Accumulators
- Using Broadcast variables

We will learn key performance considerations:

1.Level of Parallelism

2. Serialization Format

3.Memory Management

4.Hardware Provisioning

- Understanding Caching & Persistence
- We will Data Partitioning/Re-partitioning techniques.
- A project to consider the above optimization techniques.
- We will how to create custom partitioner.

RUNNING SPARK ON A CLUSTER

- Understand the Spark Runtime Architecture and various components such as Driver, Executor, Cluster Manager etc.
- Learn what goes inside when we launch an spark application.
- We will learn the two modes of Spark: Local and Cluster.
- How to launch a program on YARN, AWS Cluster etc.

- How to setup spark in standalone mode.
- Understand and demonstrate on how to run drive in various modes.
- Learn how to package the dependencies of your code.
- Understand how to use the Spark-Submit and various command line options.

STREAM PROCESSING WITH SPARK AND KAFKA

- 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
- Understanding and loading various Input formats: JSON, XML, AVRO, SequenceFile?, Parquet, Protocol Buffers.
- Comparing Compressions
- Understanding Row Oriented and Column Oriented Formats - RCFile?

MACHINE LEARNING WITH SPARK

- GraphX: Graph Processing and Analysis
- Understanding Machine Learning
- MlLib Example: k-means
- SparkR Example

GRAPH PROCESSING WITH GRAPHX

Basics of Graph Processing: Covers the understanding of what does it mean by graph processing in real life with examples. What are other frameworks providing graph computing?

GraphX Overview: What is GraphX? Understanding the functionalities and algorithms provided by GraphX. And how does GraphX work. Along with comparison with other similar products.

Implementing Page rank using GraphX: We will learn the basics of PageRank - the algorithm that made Google. The we learn how to implement using GraphX.

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