CLOUD x LAB

BIG DATA WITH HADOOP & 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.

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 Big Data with Hadoop and Apache Spark.

LIVE SESSIONS

60+ 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).

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

ZOOKEEPER

- Why Do we need it?
- Understanding Data Model
- Hands-On
- Quiz & Assessment
- How does election happen - Paxos Algorithm?
- Use cases
- When not to use
- Quiz & Assessment

HDFS

- 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)

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?

YARN

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

MAPREDUCE BASICS

- Why MapReduce?
- Understanding MapReduce Framework
- 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

MAPREDUCE ADVANCED

- Example 6 - Secondary Sorting (Word Recommendation)
- Example 7 - Partitioner
- 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

ANALYZING DATA WITH PIG

- Why Pig?
- Basic Structure of Pig Latin
- Getting Started
- Example - NYSE Stock Exchange
- Concept - Lazy Evaluation

PROCESSING DATA WITH HIVE

- Why Hive?
- Hive Architecture Overview
- 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 Analyses of Twitter Data
- Advanced - Partition Tables
- Understanding HCatalog & Impala
- Quiz

NOSQL AND HBASE

- Case Study: The days before NoSQL
- What is NoSQL?
- 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

IMPORTING DATA WITH SQOOP AND FLUME, OOZIE

- Sqoop Overview
- Import From MySQL to HDFS, Hive, HBase
- 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

CLOUD x LAB

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

- Hive - Sentiment Analysis
- Processing the NSE (National Stock Exchange) data with Hive for various insights
- Doing Analytics on the MovieLens data to generate the movie ratings
- 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 Hadoop & Spark course is compatible with the following certifications:

- Cloudera Certified Professional (CCP): Data Engineer
- Cloudera Certified Associate (CCA): Spark & Hadoop Developer
- Hortonworks Certified Developer (HDPCD)
- 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