HADOOP & SPARK

You would learn below technologies after finishing this course:

JAVA, SCALA, SQL, LINUX, AWS - EMR, EC2, S3, Hadoop, Hive, Pig, Sqoop, Oozie, Tez, Spark, KAFKA, Cassandra, Machine Learning modules, Maven, Eclipse, IntelliJ IDEA.

Course Outline: ----------

Day1

- - INTRODUCTION Legacy of Big Data Analytics - Google way of doing things - Birth of Hadoo - Hadoop EcoSystem - Spark - Road Ahead

Day2

- Cluster Setup Cluster setup using Amazon EMR

Day3

- - HDFS HDFS architecture - Write semantics - Read semantics - HDFS CLI - WEB HDFS

Day4 & Day5

- - JAVA Java basics

Day6

- - YARN Sample YARN MR job - Understanding execution - Understanding code - Different phases of job - Hadoop distcp

Day7

- - YARN Hadoop Configuration - Different MR input formats - MR output formats - Anatomy of MR1 job run - Anatomy of MR2 job run - Handling Failures - Examples

Day8

- - HIVE Hive Cli - Hive DDL - Inserts - Deletes - ARCHITECTURE

Day9

- - HIVE File formats - Complex Data types - Views & Indexes - Grouping Functions - UDF

Day10

- - HIVE Subqueries - Joins - Windowing & Analytical functions - Best practices - HIVE Configuration - Metastore - TEZ

Day11

- - SQOOP Architecture - Basics - Import - export - sqoop metastore

Day12

- - PIG PIG basics - Data Types - Operators - Expressions - Functions - Macros & UDFs - PIG using Hcatalog

Day13

- - Oozie - Oozie Architecture - workflows - Coordinator - Examples

Day15

- - SPARK Introduction Spark History - Spark Stack - Why Spark - Architecture -Simple Spark Job

Day16, 17

- - SCALA Data Types - Operators - Collections - Functions - Classes - File I/O - Exception Handling

Day18

- - SPARK RDD Cluster Modes - Self contained application - spark-submit - Resource Allocation - RDD - Using Intellij IDEA

Day19

- - SPARK Programming Creating RDDs - Transformations - Actions - Shuffle Operations - Shared Variables

Day20

- - SPARK SQL Why Spark SQL? - DataFrames - Loading DF from DataSources - Interacting with Hive

Day21

- - SPARK DataFrames Programming with Dataframes

Day22

- - SPARK With Zeppelin Zeppelin Introduction - Interpreters - Visualization - Notebooks

Day23

- - SPARK Streaming Spark Streaming basics - Dstreams - Types of Sources - Transformations on Streams - Types of operations - Check pointing - Tuning

Day24

- - Spark Structured Streaming Execution model - Structured Streaming - Demo

Day 25

- - Kafka Kafka basics - Architecture - Integration with Spark

Day 26

- - Spark with Cassandra

Day 27

- - SPARK Debugging & Performance Tuning Configuration Management - Spark UI - Logs - Key Performance Considerations

Day 28

- - SPARK Mlib Basics Machine Learning Basics - Data Types - Working with Vectors - Algorithms

Day29

- - SPARK Mlib Programming Vectors - PipeLines - Examples Using Several ML/Mlib algorithms

Day30 & 31 & 32

- - SPARK Real world use cases

Day33

- - Source Code Management using GIT Build management with Jenkins

Day34

- - Resume preperation

Day35

- - Mock interviews