Big Data Analytics using Apache Hadoop and Spark with Scala

Training Highlights :

- 80% of the training is with Practical Demo (On Custom Cloudera and Ubuntu Machines)
- 20% Theory Portion will be important to understand the basics of Hadoop and will help in cracking the Interview Rounds
- 8 POCs, Countless Assignments and Real Time examples
- 2 Real Time Big Data Projects Implemented using Hadoop and Spark
- Practice on Cloudera Machines (Provided as a part of the training only!)
- Latest Cloudera Demo and given to students for further practice at Home
- Trainer has worked with Industry Leaders and working with a giant MNC as a Enterprise Data Warehouse (EDW) Developer
- Trainer Experience : 6+ Years in IT Industry
- Mock Interviews Preparations and Guidelines
- Guidance for Interview and Certifications

Many Roles in Big Data Analytics Technolgy : This course targets below roles.

1 Big Data Developer
2 Hadoop Developer
3 Spark Developer
4 Big Data Analysts with Hadoop and Spark
5 Hadoop / Spark QA - Tester

1 BigData Introduction

1.1 What is Big Data? Definition of BigData
1.2 Why Big Data?
1.3 Evolution of Big Data
1.4 Market Trends
1.5 Types of Data
1.6 Big Data and Its Sources
1.7 Big Data Use Cases
1.8 Why Hadoop is leading tool in current It Industry
1.9 Java Essentials for Hadoop Guidelines

2 Introducing Hadoop as a Solution to Big Data Analytics Problem

2.1 Introduction to Hadoop
2.2 History and Milestones of Hadoop

6 YARN Framework And Advanced MapReduce Framework

6.1 Introduction to YARN
6.2 Why YARN If MapReduce Already there?
6.3 In Depth YARN Architecture
6.4 Role of each and every component of YARN Architecture with Practical Demo
6.5 Image Data Analysis using Advanced MapReduce APIs - Code Demo
6.6 Distributed Cache
6.7 Input Formats in MapReduce
6.8 Output Formats in MapReduce
6.9 Data Types in Hadoop
6.10 Joins in MapReduce
6.11 Reduce Side Join
6.12 Map Side Join
6.13 Skewed Join
6.14 Replicated Join
6.15 Composite Join
6.16 Cartesian Product
6.17 MapReduce program for Writable classes Demo

7 Introducing Hadoop's new release - Hadoop 3 and Major Players in the market who offer Hadoop as a Product and Service

7.1 Hadoop 3 Introduction and its new features explained
7.2 Diff between various versions of Hadoop
7.3 Cloudera Distribution In Depth Study
7.4 Clooudera Machine Live Demo
7.5 Cloudera HUE and its practical use
7.6 Practicing on the Latest Cloudera Machine Access (All tools would be practice over cloudera HUE Machine)

8 Pig

8.1 Pig Introduction
8.2 Brief History and Reason for Naming this component as Pig
8.3 Pig Real Life Use Cases Introduction
8.4 How Pig Works
8.5 Pig Execution Modes
8.6 Pig Features
8.8 Why Pig if MR Already there?
8.8 Data Model in Pig
8.9 Pig Data Types
8.10 Pig Latin Language Introduction
8.11 Pig Latin Manual with commands, Functions
8.12 Wordcount Demo in Pig - Code Demo

8.13 Installing Pig
8.14 Pig UDFs
8.15 Processing Structured Data using Pig
8.16 Procdessing Semiu Structured Data using Pig
8.18 Pig Libraries
8.18 Pig Complex Data Types
8.19 Finding the Number of Occurrences of a particular Word Code Demo in Pig
8.20 Getting Datasets for Pig Development
8.21 When to use Pig and when not to ?
8.22 Pig Assignment - POCs

9 Hive

9.1 What is Hive and Why we have Hive when Pig and MR Already there?
9.2 Brief History of Hive
9.3 Hive Architecture and its components
9.4 Hive Metastore and its configuration types
9.5 Installing Hive
9.6 Hive Thrift Server
9.7 Hive Query Language (HQL)
9.8 Hive vs SQL
9.9 Types of Tables in Hive
9.10 HQL Syntax and Practical Demo side by side
9.11 Data Types In Hive
9.12 Run and Execute Hive queries
9.13 Hive query writing and execution modes
9.14 Programmming using Hive
9.15 Hive Functions - Builtin and UDFs
9.16 UDAFs using Hive
9.17 Hive Versions and the features included in various versions
9.18 Partitioning and Bucketing in Hive
9.19 POC on Hive Data sets provided in class which includes all the Hive concepts
9.20 HQL Assignment - POCs
9.21 When to use Hive and when not to ?

10 SQOOP

10.1 What is Sqoop
10.2 Introducing Data Import / Export Tools
10.3 Sqoop and Its Uses
10.4 Benefits of Sqoop
10.5 Sqoop Processing
10.6 Sqoop Execution Process
10.7 Installing 2 RDBMS (mysql and Oracle) for Sqoop Demo Purposes - demo would be covered

in class side by side
10.8 Installing Sqoop
10.9 Importing Data Using Sqoop - Practical Demo Side by Side using mysql relational db
10.10 Installing Sqoop Connectors
10.11 Importing Data to Hive Directly
10.12 Exporting Data from Hadoop Using Sqoop
10.13 Sqoop Practical Demo - Using mysql and Oracle to Import Data from RDBMS to Hadoop and vice versa.

11 Introducing Scala Language as a Support to Implement Spark Framework

11.1 Why Scala?
11.2 What is Scala?
11.3 Scala Basics
11.4 Scala Basic Types
11.5 Defining Functions
11.6 Control and Loop Statements (If, While, Do While, etc.)
11.7 Operators, Precedence Rules, Conditional Operators, Enumerations
11.8 Method Declarations, Literals, Lists, Tuples, Options, Maps, Reserved Words
11.9 IDE for Scala
11.10 Traits Intro – Traits as Mixins, Stackable Traits, Creating Traits Basic OOPS – Class and Object Basics, Scala Constructors, Nested Classes, Visibility Rules
11.11 Functional Programming in Scala, Functional Data Structures, Implicit Function Parameters
11.12 Recursion, Tail Calls, Functional Literals and Closures

12 Introducing Apache Spark - A new Dimension to the Big Data World

12.1 Batch Vs Real Time Big Data Analytics
12.2 In Memory Data - Spark
12.3 What is Spark
12.4 Spark Architecture and its components
12.5 Features of Spark
12.6 Spark vs Hadoop
12.7 Challenges Spark is addressing and know how it is faster than Hadoop

13 Hadoop / Spark Admin Basics (Hadoop & Spark Installation and Cluster Configuration)

13.1 Different Configuration Files of Hadoop Cluster
13.2 Properties of hadoop-default.xml
13.3 Installing and Configuring Hadoop Cluster with Spark and Scala Tools
13.4 Port Numbers for Individual Hadoop and related Spark Services
13.5 Hadoop Security Kerberos

14 Spark Component /Tools and In Depth Study Part 1

RDDs, Spark SQL

14.1 Spark Context Significance
14.2 RDD - Resilient Distributed Data set
14.3 Need for RDD, what it is and what it is not
14.4 Playing with RDDs with various Transformations and Actions on them
14.5 Wordcount Practical Demo using Spark
14.6 Behind the scene execution of a Spark Program
14.7 Dependencies in Spark Program, Job, Stage and Task classification
14.8 Spark Memory Management and Fault Tolerance
14.9 Caching overview and Distributed Persistence
14.10 Spark SQL Overview
14.11 Accessing Hive using Spark
14.12 Shared Variables: Broadcast Variables, Accumulators
14.13 Data Frame, Dataset and various operations on Dataset

15 Spark Component /Tools and In Depth Study Part 2

Spark Streaming, Spark MLLib, GraphX

15.1 Spark Streaming Architecture
15.2 First Spark Streaming Program
15.3 Transformations in Spark Streaming
15.4 Structured Streaming
15.5 Intro to Spark MLLib
15.6 Intro to Spark GraphX

16 NoSQL DB - HBase , MongoDb and Cassandra along with Zookeeper

16.4 Introduction to NoSQL DB
16.2 NoSQL DB Vs RDBMS
16.3 Schemaless Approach explained
16.4 CAP Theorom with Real time example
16.5 ACID Vs. CAP
16.6 Which NoSQL to use in different situations?
16.7 Types of NoSQL DBs
16.8 Hbase Architecture of Column Families
16.9 HBase table and column family structure
16.10 HBase versioning concept
16.11 HBase flexible schema
16.12 Hbase in Cloudera HUE
16.13 Introducing Zookeeper and where it fits in Hadoop Ecosystem to help and maintain servers
16.14 Cassandra Overview

16.15 MongoDB Overview and Introducing Mongod Shell

17 FLUME - Log Data Collection Tool for Big Data Analytics

17.1 Why Flume is used when Sqoop is already there?
17.2 Apache Flume Introduction
17.3 Flume Model
17.4 Flume Features - Scalability
17.5 Hands-on on Flume (Includes Installation followed by code demo and assignments)
17.6 Loading Twitter Data from your account to Hadoop using Flume (Includes Handson example of Twitter Developer Account)

18 Oozie

18.1 Introduction to Oozie
18.2 How to schedule jobs using Oozie
18.3 What kind of jobs can be scheduled using Oozie
18.4 How to schedule jobs which are time based
18.5 Installing and Executing Oozie Schedulers (Includes Handson Practice and demo)
18.6 Apache Oozie Workflow and Coordinator (With Demo Side by Side)
18.7 Oozie Bundles

19 Hadoop EcoSystem at High Level Classification (Useful to know the future trends in Hadoop)

19.1 Overview of Hadoop Ecosystem and its components which are covered in depth above but at a higher level of classification
19.2 Apache Hadoop Ecosystem
19.3 File System Component
19.4 Data Store Components
19.5 Serialization Components
19.6 Job Execution Components
19.7 Work Management, Operations, and Development Components
19.8 Security Components
19.9 Data Transfer Components
19.10 Analytics and Intelligence Components
19.11 Graph-Processing Framework Components
19.12 Search Frameworks Components

20 Case Study and Real Time Projects

20.1 8 Case Study with combination of multiple ecosystem components (Is covered time by time)
20.2 2 Real Time Projects with Data sets (At the end of training as this requires all the components knowledge thoroughly)

20.3 Interview Questions Discussions
20.4 Mock Interviews
20.5 Tips to crack Interviews on Big Data Analytics on Hadoop and Spark
20.6 Hadoop and Spark Developer Certifications and Guidelines