Big Data - Hadoop Training Course Outline

4 Roles in Hadoop Technology : This course targets the first 3 roles.

1 Hadoop Developer
2 Hadoop Analyst
3 Hadoop Tester
4 Hadoop Admin

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 Covered

2 Introducing Hadoop as a Solution to Big Data Problem

2.1 Introduction to Hadoop
2.2 History and Milestones of Hadoop
2.3 Organizations Using Hadoop
2.4 Hadoop Cluster Using Commodity Hardware

3 Architecture of Hadoop and Installation

3.1 Hadoop Architecture
3.2 Hadoop Cluster Using Commodity Hardware
3.3 Introduction to Hadoop Release-1
3.4 Hadoop Daemons / Services in Hadoop Release-1
3.5 Hadoop Cluster and Racks
3.6 Hadoop installation
3.7 How to start and stop the services
3.8 Hadoop Architecture Breakdown and various components Overview

4 HDFS - Hadoop Distributed File System

4.1 What is HDFS
4.2 HDFS Characteristics
4.3 HDFS Key Features
4.4 HDFS Architecture
4.5 Regular File System vs. HDFS
4.6 How to read and write files

4.7 Basic Unix commands for Hadoop
4.8 Hadoop FS shell
4.9 HDFS Daemons Practical Demo
4.10 NameNode Operation
4.11 Data Block Split
4.12 Benefits of Data Block Approach
4.13 HDFS Block Replication Architecture
4.14 Replication Method
4.15 Data Replication Topology
4.16 HDFS Access
4.17 Case Study - Demo - HDFS
4.18 Setting Up HDFS Block Size

5 Map Reduce Framework

5.1 How Map Reduce works as Processing Framework
5.2 End to End execution flow of Map Reduce job
5.3 Different tasks in Map Reduce job
5.4 Combiner and Partitioner
5.5 Characteristics of MapReduce
5.6 Real-time Uses of MapReduce
5.7 Build MapReduce Program, WordCount Demo in Eclipse, Ubuntu Machine
5.8 Realtime work with MapReduce
5.9 Map Reduce complex scenarios

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 Pig

7.1 Pig Introduction
7.2 Brief History and Reason for Naming this component as Pig
7.3 Pig Real Life Use Cases Introduction
7.4 How Pig Works
7.5 Pig Execution Modes
7.6 Pig Features
7.7 Why Pig if MR Already there?
7.8 Data Model in Pig
7.9 Pig Data Types
7.10 Pig Latin Language Introduction
7.11 Pig Latin Manual with commands, Functions
7.12 Wordcount Demo in Pig - Code Demo
7.13 Installing Pig
7.14 Pig UDFs
7.15 Processing Structured Data using Pig
7.16 Processing Semiu Structured Data using Pig
7.17 Pig Libraries
7.18 Pig Complex Data Types
7.19 Finding the Number of Occurrences of a particular Word Code Demo in Pig
7.20 Getting Datasets for Pig Development
7.21 When to use Pig and when not to ?
7.22 Pig Assignment

8 Hive

8.1 What is Hive and Why we have Hive when Pig and MR Already there?
8.2 Brief History of Hive
8.3 Hive Architecture and its components
8.4 Hive Metastore and its configuration types
8.5 Installing Hive
8.6 Hive Thrift Server
8.7 Hive Query Language (HQL)
8.8 Hive vs SQL
8.9 Types of Tables in Hive
8.10 HQL Syntax and Practical Demo side by side
8.11 Data Types In Hive
8.12 Run and Execute Hive queries
8.13 Hive query writing and execution modes
8.14 Programming using Hive
8.1.5 Hive Functions - Builtin and UDFs
8.16 UDAFs using Hive
8.17 Hive Versions and the features included in various versions
8.18 Partitioning and Bucketing in Hive

8.19 POC on Hive Data sets provided in class which includes all the Hive concepts
8.20 HQL Assignment
8.21 When to use Hive and when not to ?

9 SQOOP

9.1 What is Sqoop
9.2 Introducing Data Import / Export Tools
9.3 Sqoop and Its Uses
9.4 Benefits of Sqoop
9.5 Sqoop Processing
9.6 Sqoop Execution Process
9.7 Installing 2 RDBMS (mysql and Oracle) for Sqoop Demo Purposes - demo would be covered in class side by side
9.8 Installing Sqoop
9.9 Importing Data Using Sqoop - Practical Demo Side by Side using mysql relational db
9.10 Installing Sqoop Connectors
9.11 Importing Data to Hive Directly
9.12 Exporting Data from Hadoop Using Sqoop
9.13 Sqoop Practical Demo - Using mysql and Oracle to Import Data from RDBMS to Hadoop and vice versa.

10 FLUME

10.1 Why Flume is used when Sqoop is already there?
10.2 Apache Flume Introduction
10.3 Flume Model
10.4 Flume Features - Scalability
10.5 Hands-on on Flume (Includes Installation followed by code demo and assignments)
10.6 Loading Twitter Data from your account to Hadoop using Flume (Includes Handson example of Twitter Developer Account)

11 Oozie

11.1 Introduction to Oozie
11.2 How to schedule jobs using Oozie
11.3 What kind of jobs can be scheduled using Oozie
11.4 How to schedule jobs which are time based
11.5 Installing and Executing Oozie Schedulers (Includes Handson Practice and demo)
11.6 Apache Oozie Workflow and Coordinator (With Demo Side by Side)
11.7 Oozie Bundles

12 NoSQL DB Overview - HBase along with Zookeeper

12.1 Introduction to NoSQL DB
12.2 NoSQL DB Vs RDBMS
12.3 Schemaless Approach explained

12.4 CAP Theorem with Real time example
12.5 ACID Vs. CAP
12.6 Which NoSQL to use in different situations?
12.7 Types of NoSQL DBs
12.8 Hbase Architecture of Column Families
12.9 HBase table and column family structure
12.10 HBase versioning concept
12.11 HBase flexible schema
12.12 Hbase in Cloudera HUE
12.13 Introducing Zookeeper and where it fits in Hadoop Ecosystem to help and maintain servers

13 Major Hadoop Distributions in Market Currently

13.1 Major Players in the market who offer Hadoop as a Product and Service
13.2 Cloudera Distribution In Depth Study
13.3 Clooudera Machine Live Demo
13.4 Cloudera HUE and its practical use
13.5 Practicing on the Latest Cloudera Machine Access - 5.7.2 (All above tools would be practice over cloudera HUE Machine)

14 Spark and Scala Introduction

14.1 Introduction to Spark
14.2 Introduction to scala
14.3 Features of Spark
14.4 Spark Architecture and its components
14.5 RDD - Resilient Distributed Datasets Overview
14.6 Wordcount Practical Demo using Spark

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

15.1 Overview of Hadoop Ecosystem and its components which are covered in depth above but at a higher level of classification
15.2 Apache Hadoop Ecosystem
15.3 File System Component
15.4 Data Store Components
15.5 Serialization Components
15.6 Job Execution Components
15.7 Work Management, Operations, and Development Components
15.8 Security Components
15.9 Data Transfer Components
15.10 Analytics and Intelligence Components
15.11 Graph-Processing Framework Components
15.12 Search Frameworks Components

16 Hadoop Admin Basics

16.1 Different Configuration Files of Hadoop Cluster
16.2 Properties of hadoop-default.xml
16.3 Port Numbers for Individual Hadoop Services
16.4 Hadoop Security Kerberos

17 Case Study and Real Time Projects

17.1 8 Case Study with combination of multiple ecosystem components (Is covered time by time)
17.2 2 Real Time Projects with Data sets (At the end of training as this requires all the components knowledge thoroughly)
17.3 Interview Questions Discussions
17.4 Mock Interviews
17.5 Tips to crack Interviews on Hadoop
17.6 Hadoop Certifications and Guidelines

Training Highlights :

- 70 Hours (5 Hours on Weekend, 14 Weekends)
- 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
- Practice on Cloudera Machines (Provided as a part of the training only!)
- Latest Cloudera 5.7 Demo and given to students for further practice at Home
- Trainer has worked with Industry Leaders and working with a giant MNC as a Warehouse Developer
- Trainer Experience : 5 Years in IT Industry
- Guidance for Interview and Certifications