Big Data Hadoop – Course Agenda

Lessons

1. Introduction to Big Data and Hadoop
   a. What is Big Data?
   b. Types of Data
   c. Need for Big Data
   d. Characteristics of Big Data
   e. Traditional IT Analytics Approach
   f. Big Data—Use Cases
   g. Handling Limitations of Big Data
   h. Introduction to Hadoop
   i. History and Milestones of Hadoop

2. Getting Started With Hadoop
   a. VMware Player—Introduction
   b. Installing VMware Player
   c. Setting up the Virtual Environment
   d. Oracle VirtualBox to Open a VM

3. Hadoop Architecture
   a. Hadoop Cluster in commodity hardware
   b. Hadoop core services and components
   c. Regular file system vs. Hadoop
   d. HDFS layer
   e. HDFS operation principle

4. Hadoop Deployment
   a. Introduction to Ubuntu Server
   b. Hadoop installation
   c. Single node and multi node configuration
   d. Hadoop Configuration in cluster environment
   e. Installing Hadoop 2.0

5. MapReduce
   a. Introduction to MapReduce
   b. Hadoop MapReduce example
   c. Hadoop MapReduce Characteristics
   d. Setting up your MapReduce Environment
   e. Building a MapReduce Program
   f. MapReduce Requirements and Features
   g. MapReduce Java Programming in Eclipse
   h. Checking Hadoop Environment for MapReduce
   i. MapReduce 2.0

6. Advanced HDFS & MapReduce
   a. HDFS Benchmarking
b. Setting up HDFS Blocks
c. Decommissioning a DataNode
d. Advanced MapReduce
e. Hadoop Data Types
f. InputFormats in MapReduce
g. OutputFormats in MapReduce
h. Distributed Cache
i. Joins in MapReduce

7. PIG
   a. Introduction to PIG
   b. Components of Pig
   c. Pig Data Model
   d. Pig Modes
   e. Pig Vs. SQL
   f. Installing Pig Engine
   g. Datasets for Pig Development
   h. Pig Latin
   i. Filtering and Transforming Data
   j. Grouping and Sorting
   k. Combining and Splitting
   l. Pig Commands

8. HIVE
   a. Why another data warehousing system
   b. What is HIVE
   c. Characteristics of Hive
   d. System Architecture and Components of Hive
   e. Hive Data Models
   f. Serialization/De-serialization
   g. Hive file formats
   h. Hive Query Language
   i. HIVE: Installing, running, and programming
   j. Hive Functions
   k. Difference between Hive and PIG

9. HBase
   a. HBase introduction
   b. Characteristics of HBase
   c. HBase Architecture
   d. Storage Model of HBase
   e. When to use HBase
   f. HBase Data Model
   g. HBase Families
   h. HBase Components
   i. Row Distribution between region servers
j. Data Storage
k. Installation of HBase
l. Configuration of HBase
m. HBase Shell Commands

10. Commercial Distribution of Hadoop
   a. Cloudera
   b. Downloading Cloudera Quickstart VM
   c. Starting the Cloudera VM
   d. Exploring the Welcome Page
   e. Understanding Hue
   f. Understanding Cloudera Manager
   g. Hortonworks Data Platform
   h. MapR Data Platform
   i. Pivotal HD
   j. IBM InfoSphere BigInsights

11. ZooKeeper, Sqoop and Flume
   a. Introduction to ZooKeeper
   b. Features of ZooKeeper
   c. Challenges faced in distributed applications
   d. Coordination
   e. ZooKeeper: Goals and Uses
   f. ZooKeeper: Entities, Data Model, Services
   g. Client APIs
   h. Recipes of Zookeeper
   i. Introduction to Sqoop (Why, what, processing, under the hood)
   j. Importing data into Hive
   k. Importing data into HBase
   l. Exporting data from Hadoop using Sqoop
   m. Sqoop Connectors
   n. Introduction to Flume
   o. Flume Use Cases
   p. Configuring and Running Flume Agents

12. Ecosystem and its Components
   a. Hadoop Ecosystem
   b. Components Overview
   c. Overview of Apache Oozie
   d. Overview of Mahout
   e. Overview of Apache Cassandra
   f. Apache Spark

13. Hadoop Administration and Troubleshooting
   a. Commands Used in Hadoop Programming
   b. Different configurations of Hadoop cluster
   c. Port Numbers for Individual Hadoop Services
d. Performance monitoring
e. Performance tuning
f. Troubleshooting and Log observation
g. Overview of Apache Ambari
h. Hadoop Security Using Kerberos