BIG DATA & HADOOP

Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. It is at the center of a growing ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications. Hadoop can handle various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analyzing data than relational databases and data warehouses provide.

Resultantly, Big Data and Hadoop professionals are going to have a significant say on company policies and marketing strategies. One of the most important and motivating reasons to learn Big Data and Hadoop is the fact that it brings an array of opportunities to bolster your career to an unprecedented level.

Administration

1. Big Data & Hadoop

- Big data features and challenges
- 3Vs for Big Data
- Problems with traditional large-scale systems
- Why Hadoop? & Hadoop fundamental concepts
- Hadoop vs RDBMS vs No-SQL
- History of Hadoop with Hadoopable problems
- Hadoop distributed file system (HDFS)
- Limitation of Hadoop

2. Hadoop Architecture

- Hadoop Version - 2.x & 1.x
- Distributions of Hadoop (Cloudera, Hortonworks)
- Architecture of Hadoop
- Rack awareness and topology
- Cluster storage daemons Name Node
- Secondary Node
- Data Nodes
- YARN Responsibilities
- Resource Manager
- Job History Server
- Node Manager
- Application Manager
- Application Master

3. Hadoop High-Availability

- Name Node availability
- Architecture of HA
- Implementation of HA
- Apache Zookeeper service

1. Quorum Journal

- Active Name Node and Standby Name Node
- Zookeeper Fail Over controller
- Quorum Journal Manager
- Quorum Journal Node(s)

2. Namespace federation (NFS)

- Namespace information
- Zookeeper fail over controller

4. Linux Initials

- Installation of Linux (Red Hat)
- Basic Linux configurations
- Basic Linux commands
- Password less ssh
- IP address and hostname
- Firewall and selinux
- Yum and creating yum repository
- NTP configurations

5. Planning your Hadoop Cluster

- Installation Prerequisites
- General planning considerations
- Choosing the right hardware
- Network considerations
- Configuring nodes
- Planning for cluster management

6**. Installation & Deployment of Hadoop**

- Choosing deployment types
- Setting up multi-nodes
- Setting up Cloudera yum repository
- Installation for Cloudera Manager
- Installing multi-node Hadoop (Cloudera) environment
- Specifying the Hadoop configuration
- Performing Initial HDFS configuration
- Performing Initial YARN and Map Reduce configuration
- Hadoop logging & cluster monitoring

7. Accessing Hadoop

- Access HDFS using command line
- Hadoop fs
- Hadoop HDFS admin
- Access Cloudera Manager (Admin)
- Access HUE (Developer)

8. Configuration for Admin

- Add and remove services
- Configuring HDFS properties like Block size
- Setting up zookeeper on multi node
- Configuring Hadoop operating system
- (YARN) & Map-Reduce
- Configuring schedulers
- Hadoop logging & monitoring
- Advanced configuration parameters
- Configuring Hadoop ports
- Explicitly including and excluding hosts

9. Cluster Maintenance

- Checking HDFS status
- Copying Data between clusters
- Adding and removing cluster nodes
- Rebalancing the cluster
- Cluster upgrading

Big data Developer

1. Sandbox / Quick Start VMs

- Overview of sandBox
- Different flavours (Virtual Box / VMware) of sandBox
- Installation of sandbox

2. Hue or Hadoop UI

- Introduction of HUE
- Getting started with HUE
- Deployment of jobs
- Functional execution of Hive/HBase
- Design of work-flow using job designer
- Data transfer in Sqoop
- Start working with sandBox

3**. Hadoop Shell & Commands**

- Hadoop developer/admin commands using shell
- NameNode & Secondary Name Node commands
- HDFs dfsadmin and file system shell commands
- Hadoop Name Node / Data Node directory structure
- HDFS permissions model
- Map-Reduce job deployment
- Oozie workflow design
- Different components jobs design

4. Map-Reduce Concepts

- Introduction to map reduce
- Architecture of map reduce
- Understanding the concept of mappers & reducers
- Anatomy of map reduce program
- Phases of a map reduce progam
- Data-types in hadoop map reduce
- Driver, mapper and reducer classes
- Input split and record reader
- Input format and output format in hadoop
- Concepts of combiner and partitioner
- Running and monitoring mapreduce jobs
- Writing your own map reduce job using map reduce API
- Different interview questions raised for map reduce

5. Scala

- Scala Introduction
- Scala versus Java
- Scala basics
- Scala Data types
- Scala packages
- Variable Declarations
- Variable Type Inference
- Control Structures
- Interactive Scala - Scala shell
- Writing Scala Scripts – Compiling the Scala Programs
- Defining Functions in Scala
- Different IDEs for Scala

6. Spark

- Motivation for spark
- Spark vs Map Reduce Processing
- Architecture of Spark
- Spark Shell introduction
- Creating Spark Context
- File operations in Spark Shell
- Spark Project with SBT in Eclipse
- Caching in Spark
- Real time Examples of Spark
- Concepts of combiner and partitioner
- Running and monitoring mapreduce jobs
- Writing your own map reduce job using map reduce AP
- Resilient Distributed Dataset (RDDS)
- Introduction of RDDs
- Features of RDDs
- Creating RDDs
- Creating RDDs referencing an external dataset
- Creating RDDs using text files
- Creating RDDs using other hadoop input formats
- RDD operations & transformations
- Features of RDD persistence
- Storage levels Of RDD persistence
- Choosing the correct RDD persistence storage level
- Invoking the Spark shell
- Creating the Spark context
- basic operations on files in Spark shell RDD
- Demo-build a Spark Python project
- Build a Spark Java project
- Shared variables broadcast & accumulators
- Double RDD methods
- Pair RDD methods Join
- Pair RDD methods Others
- General RDD methods
- Transformations in RDD
- Actions in RDD
- Key Value pair RDD in Python
- Reading text & sequence file from HDFS
- Using groupby operation
- Python application performing group by operation

Spark SQL

- Introduction to Spark SQL
- The SQL Context
- Hive vs Spark SQL
- Spark SQL
- support for Text Files, Parquet and JSON files
- Data Frames
- Real time Examples of Spark SQL

Spark Streaming

- Introduction to Spark Streaming
- Architecture of Spark Streaming
- Spark Streaming vs Flume
- Introduction to Kafka
- Spark streaming Integration with Kafka overview
- Real time examples of Spark Streaming and Kafka

Spark Mlib

- Introduction to Machine Learning
- Vector Class in Mlib
- Spark Mlib Algorithms introduction
- Classification and Regression Algorithms
- Naïve Bayes Classification Algorithms
- Decision Trees Algorithm Overview

7. Hive

- Problems with No-SQL
- database
- Introduction & installation Hive
- Hive schema and data storage
- Data types & introduction to SQL
- Hive-SQL: DML & DDL
- Hive-SQL: views & indexes
- Explain and use the various Hive file formats
- Use Hive to run SQL-like queries to perform data analysis
- Use Hive to join data sets using a variety of techniques
- Map-side joins and Sort-Merge-Bucket joins

8. Sqoop

- Installation of Sqoop
- Ingesting data from external (DB) sources with Sqoop
- Ingesting data from/to relational databases with Sqoop
- Integration of Sqoop and Hive
- Best practices for importing data

9. Oozie

- Oozie introduction
- Oozie architecture
- Oozie Configuration Files
- Oozie Job Submission
- Workflow.xml
- Coordinator.xml
- job.coordinator.properties

10. Flume

- Flume introduction
- Flume Architecture
- Flume Master , Flume Collector and Flume Agent
- Flume Time Use Case using Apache Flume

11. Kafka

1: What is Kafka – An Introduction

- What is Apache Kafka
- Components and use cases of Kafka
- Implementing Kafka on a single node
- Installing spark as a standalone & cluster environment

2: Multi Broker Kafka Implementation

- Learning about the distributed Kafka terminology
- Deploying multi node Kafka with independent Zookeeper
- Adding replication in Kafka
- Working with Partitioning and Brokers
- Understanding Kafka consumers
- Kafka Writes terminology
- Various failure handling scenarios in Kafka

3: Multi Node Cluster Setup

- Multi node cluster setup in Kafka
- Various administration commands
- Leadership balancing and partition rebalancing
- Graceful shutdown of kafka Brokers and tasks
- Working with the Partition Reassignment Tool
- Cluster expending & Assigning Custom Partition
- Removing of a Broker and improving Replication Factor

4: Producers & Consumers

- Connecting Kafka using PyKafka
- Writing your own Kafka Producers and Consumers
- Writing a random JSON Producer
- Writing a Consumer to read the messages from a topic
- Writing and working with a File Reader Producer
- Writing a Consumer to store topics data into a file

Fee: 22,500 RS/-
Duration: 3 Months

FAQs

1. What is Big Data & Hadoop?

Ans. Hadoop is a software framework for storing and processing Big Data. It is an open-source tool build on java platform and focuses on improved performance in terms of data processing on clusters of commodity hardware. Hadoop comprises of multiple concepts and modules like HDFS, Map-Reduce, HBASE, PIG, HIVE, SQOOP and ZOOKEEPER to perform the easy and fast processing of huge data. Hadoop conceptually different from Relational databases and can process the high volume, high velocity and high variety of data to generate value

2. Who should do this course?

Ans. Software Engineers, who are into ETL/Programming and exploring for great job opportunities in Hadoop. Managers, who are looking for the latest technologies to be implemented in their organization, to meet the current &amp; upcoming challenges of data management. You can turn yourself into a certified professional via a professional Hadoop Institute in Noida, Gurgaon, Ghaziabad and Delhi NCR. Mapping Minds offers Hadoop Classes in Delhi, Gurgaon, Ghaziabad and Delhi NCR at most affordable prices.

3 what are prerequisites for big data Hadoop training?

Ans Prerequisites for learning Hadoop include hands-on experience in one programing language and good analytical skills to grasp and apply the concepts in Hadoop.

4. Who are the trainers?

Ans. The training's are delivered by highly qualified and certified instructors with relevant industry experience. And feel very proud while telling you, Mapping Minds is one of the well-established Hadoop training institute in Delhi that provides best Hadoop training in Delhi.

5. Do we do live project after completing the course?

Ans. We will provide 2 data sets to work on real live projects.

6. If I am not from a programming background but have a basic knowledge of programming can i still learn Hadoop?

Ans. Yes, you can learn Hadoop without being from a software background. We provide complimentary courses in Python and Linux so that you can brush up on your programming skills. This will help you in learning Hadoop technologies better and faster.

7. If i missed classes, so will get any back up?

Ans. Yes, if you missed any class you will get back with in 15 days.

8.How do I pay course fees?

Ans. We support multiple payment options online /offline. Choose an option that suits you the most.

9. Is the course theoretical or practical?

Ans. Course will be combination of theoretical and practical sections on each topic. We also provide exposure to our live projects.