

Hadoop Course Content

1. Introduction to Hadoop:

- What is cloud computing
- What is Grid computing
- What is Virtualization
- Interrelation between cloud computing, grid computing, virtualization
- What is Big data
- What is analytics
- Hadoop solutions
- Hadoop distributions
- Difference between Hadoop and other Traditional systems
- Volunteer Computing
- Data Retrieval
- Difference Between Random Access and Sequential Access
- Databases(No SQL)

2. Basic Concepts:

- What is Hadoop?
- Hadoop Distributed File System
- Working of MapReduce
- Anatomy of a Hadoop Cluster

3. Hadoop Demons:

- Namenode
- Datanode
- Secondary namenode
- Job tracker
- Task tracker

4. HDFS at details:

- Blocks and Splits
- Replication
- Data high availability
- Data Integrity
- Cluster architecture and block placement

5. Programming and Performance Tunning

- Developing MapReduce Programs in
 - a. Local Mode
 - b. Pseudo-distributed Mode
 - c. Fully distributed mode

6. MapReduce Programming

- Examining a Sample MapReduce Program
- Basic API Concepts
- The Driver Code
- The Mapper
- The Reducer
- Hadoop's Streaming API

7. Hadoop Cluster

- Install and configure Apache Hadoop
- Make a fully distributed Hadoop cluster on a single laptop/desktop
- Install and configure Cloudera Hadoop distribution in fully distributed mode
- Install and configure Horton Works Hadoop distribution in fully distributed mode
- Monitoring the cluster
- Getting used to management console of Cloudera and Horton Works

8. Hadoop Security

- Hadoop importance
- What Kerberos Is and How it Works
- Configuring Kerberos Security
- Integrating a Secure Cluster with Other Systems

9. Managing and Scheduling Jobs

- Managing Running Jobs
- Hands-On Exercise
- The FIFO Scheduler
- The Fair Scheduler
- Configuring the Fair Scheduler

10. Cluster Maintenance

- Checking HDFS Status
- Hands-On Exercise
- Copying Data Between Clusters
- Adding and Removing
- Cluster Nodes
- Rebalancing the Cluster
- Hands-On Exercise
- NameNode Metadata Backup

11. Cluster Monitoring and Troubleshooting

- General System Monitoring
- Managing Hadoop's Log Files
- Using the NameNode
- JobTracker Web UIs
- Hands-On Exercise
- Cluster Monitoring with Ganglia
- Common Troubleshooting Issues
- Benchmarking Your Cluster

