

*Suggested Teaching Guidelines for*

**Hadoop Administration – PG-DHPCSA September 2022**

**Duration:** 40 class room hours + 40 Lab hours

**Objective:** To reinforce knowledge of BigData Technologies such as Grid Computing, Hadoop Administration

**Prerequisites:** Knowledge of Linux command, SQL and Core Java

**Evaluation method:** CCEE Theory exam– 40% weightage  
Lab exam (Case Study based) – 40% weightage  
Internal exam – 20% weightage

**List of Books / Other training material**

**Text Book:**

1. Hadoop Operations , Eric Sammer by O'reilly

**Reference:**

1. Hadoop The Definitive Guide 3rd Edition by O'Rellay ( Author :- Tom White)
2. Hadoop In Practice by Manning (Author:- ALEX HOLMES)
3. Pro Hadoop by Aprss(Author:-Jason Venner)
4. Hadoop In Action by Manning Publications (Author:- CHUCK LAM)

**Note:** Each session having 2 Hours

**Session 1**

**Introduction to Big Data**

- What is Big Data,
- Big Deal about Big Data,
- Big Data Sources,
- Industries using Big Data,
- Big Data challenges

**Big Data Technologies and Hadoop**

- Solution to Big Data problems,
- Various Big Data Technologies,
- Big Data/Hadoop Platforms,
- Hadoop Distributions and Vendors,
- Big Data Suites.

**Introduction to Hadoop**

- A Brief History of Hadoop,
- Evolution of Hadoop,
- Comparison with Other Systems,
- Hadoop Releases

*Suggested Teaching Guidelines for*

***Hadoop Administration – PG-DHPCSA September 2022***

**Session: 2 & 3**

**Hadoop Architecture**

- Hadoop Architecture,
- Core components of Hadoop,

**Getting Started: Hadoop Installation**

- Setting up a Hadoop Cluster,
- Logging configuration
- Cluster specification,
- Cluster Setup and Installation,
- Common Hadoop Shell commands
- Clustering Monitoring
- Single and Multi-Node Cluster Setup on Virtual Machine,
- Hadoop Configuration, Security in Hadoop, Administering Hadoop,
- HDFS – Monitoring & Maintenance, Hadoop benchmarks
- Hadoop in the cloud.

**Session: 4**

**Hadoop Distributed File System (HDFS)**

- Distributed File System,
- What is HDFS,
- Major goals of HDFS Design
- Where does HDFS fit in,
- Core components of HDFS,
- Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node

**Lab-Assignment:**

- Run the HDFS commands, and add a one liner understanding for each of the command.

**Session: 5, 6 & 7**

**HDFS Architecture**

- HDFS Architecture,
- Scaling and Rebalancing,
- Big Deal about HDFS,
- Replication,
- Rack Awareness,
- Data Pipelining,
- Node Failure Management.
- HDFS NameNode High Availability
- Components and daemon of an HDFS HA-Quorum cluster
- HDFS Federation use case
- Kerberos: Role of HDFS security

**HDFS Data Storage Process**

- HDFS Data storage process,
- Anatomy of writing and reading file in HDFS,
- HDFS user and admin commands,
- HDFS Web Interface.

**Lab-Assignment:**

- Execute the provided code using HDFS, step run and understand

**Session: 8 & 9**

**Getting in touch with Map Reduce Framework**

- Hadoop Map Reduce paradigm,
- Stages of MapReduce
- Map and Reduce tasks,
- Map Reduce Execution Framework,

*Suggested Teaching Guidelines for*

***Hadoop Administration – PG-DHPCSA September 2022***

- Anatomy of a Map Reduce Job run

**Lab-Assignment:**

- Execute the train data example.
- Execute the train data example using chained methods

**Session: 10 & 11**

**YARN**

- YARN Architecture
- YARN Resource Management
- Hadoop Schedulers
- Upgrading cluster from Hadoop1 to Hadoop2
- MapReduce job workflow on YARN
- Migration from MRv1 to MRv2 on YARN : Configuration changes in files

**Session: 12**

**Security in Hadoop**

- HDFS Security Model
- LDAP and Hadoop
- LDAP support in Hadoop

**Lab-Assignment:**

- Configure LDAP in Linux.
- Integrate LDAP with Hadoop.

**Session: 13, 14, & 15**

**Hadoop Cluster Planning**

- Choosing hardware and operating systems,
- OS comparison based on features like kernel tuning, disk swapping & etc.
- Based on scenario and workload identify hardware, cluster size
- Based on scenario identify eco-system components
- Identify key network components, Network topology/design based on network usage in Hadoop

**Cluster Maintenance**

- Managing Hadoop Process both with script and manually
- HDFS Maintenance tasks - Adding, decommissioning data node & etc.
- MapReduce Maintenance tasks - Adding, decommissioning Taskt
- racker, killing job/task & etc.
- Backup & Recovery