

Suggested Teaching Guidelines for

Hadoop Administration – PG-DHPCSA August 2019

Duration: 44 class room hours + 56 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 & 2:

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

Getting Started: Hadoop Installation

- Setting up a Hadoop Cluster,
- Logging configuration
- Cluster specification,
- Cluster Setup and Installation,
- 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.

Suggested Teaching Guidelines for

Hadoop Administration – PG-DHPCSA August 2019

Session: 3 & 4

Hadoop Architecture

- Hadoop Architecture,
- Core components of Hadoop,
- Common Hadoop Shell commands

Session: 5, 6 & 7

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

Session: 8, & 9

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

Session: 10

HDFS Data Storage Process

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

Session: 11 & 12

Getting in touch with Map Reduce Framework

- Hadoop Map Reduce paradigm,
- Stages of MapReduce
- Map and Reduce tasks,
- Map Reduce Execution Framework,
- Anatomy of a Map Reduce Job run

Session: 13 & 14

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: 15 & 16

Security in Hadoop

- HDFS Security Model

Suggested Teaching Guidelines for

Hadoop Administration – PG-DHPCSA August 2019

- Introduction to LDAP tool
- LDAP and Hadoop
- LDAP support in Hadoop

Session: 17, 18, 19 & 20

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

Session: 21 & 22

Cluster Maintenance

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