

*Suggested Teaching Guidelines for*  
**Big Data Technologies**  
**PG-DBDA September 2022**

**Duration:** 66 Classroom hours + 84 Lab hours

**Objective:** To reinforce knowledge of BigData Technologies such as Hadoop, Map reduce, HBase, PIG, Spark (PySpark)

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

**Evaluation method:** Theory exam – 40% weightage  
Lab exam – 40% weightage  
Internal exam – 20% weightage

**List of Books / Other training material**

**Textbook:**

1. Hadoop: The Definitive Guide, SPD

**Reference:**

1. Big Data, Black Book by DreamTech
2. Programming Hive by O'Reilly (Author:- Edward Capriolo, Dean Wampler, and Jason Rutherglen)
1. Hadoop The Definitive Guide 4<sup>th</sup> Edition by O'Reilly (Author: - Tom White)
2. Hadoop In Practice by Manning (Author: - ALEX HOLMES)
3. Pro Hadoop by Aprss (Author:- Jason Venner)
4. Hadoop with python
5. Hadoop Real-World Solutions Cookbook by Packet publication (Author: Jonathan R. Owens, Jon Lentz, Brian Femiano)
6. Hadoop In Action by Manning Publications (Author: - CHUCK LAM)
7. Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault
8. Big Data Made Easy: A Working Guide to the Complete Hadoop Toolset
9. Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large-Scale Data Processing, Machine Learning, and Graph Analytics, and High-Velocity Data Stream Processing

**Note:** Each session having 2 Hours

**Introduction to Bigdata and Hadoop (Theory- 16 Hrs and Lab- 06 Hrs)**

**Session: 1, 2 & 3**

**Introduction to Big Data**

- Big Data - Beyond the Hype,
- Big Data Skills and Sources of Big Data,
- Big Data Adoption,
- Research and Changing Nature of Data Repositories,
- Data Sharing and Reuse Practices and Their Implications for Repository Data Curation,
- Overlooked and Overrated Data Sharing,

- Data Curation Services in Action,
- Open Exit: Reaching the End of The Data Life Cycle,
- The Current State of Meta-Repositories for Data
- Curation of Scientific Data at Risk of Loss: Data Rescue And Dissemination

**Introduction to Hadoop**

- A Brief History of Hadoop,
- Evolution of Hadoop,
- Introduction to Hadoop and its components
- Comparison with Other Systems,
- Hadoop Releases
- Hadoop Distributions and Vendors

**Hadoop Distributed File System (HDFS)****Session: 4 & 5****Hadoop Distributed File System (HDFS)**

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

**HDFS Architecture**

- HDFS Architecture,
- Scaling and Rebalancing,
- Replication,
- Rack Awareness,
- Data Pipelining,
- Node Failure Management.
- HDFS High Availability NameNode

**Lab-Assignment:**

- Run the HDFS commands, and add a one liner understanding for each of the command.
- Execute the provided code using HDFS, step run and understand

**Hadoop Installation and Cluster Configuration (Lab – 02 Hrs)****Session: 6****Getting Started: Hadoop Installation**

- Hadoop Operation modes
- Setting up a Hadoop Cluster,
- Cluster specification,
- Single and Multi-Node Cluster Setup on Virtual & Physical Machines,
- Remote Login using Putty/Mac Terminal/Ubuntu Terminal.
- Hadoop Configuration, Security in Hadoop, Administering Hadoop,
- HDFS – Monitoring & Maintenance, Hadoop benchmarks,
- Hadoop in the cloud.

**Session: 7****Hadoop Architecture**

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

**Session: 8**

**HDFS Data Storage Process**

- HDFS Data storage process,
- Anatomy of writing and reading file in HDFS,
- Handling Read/Write failures
- HDFS user and admin commands,
- HDFS Web Interface.

**Map Reduce (Theory – 06 Hrs & Lab – 12 Hrs)**

**Session: 9**

**Getting in touch with Map Reduce Framework**

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

**More Map Reduce Concepts**

- Partitioners and Combiners,
- Input Formats (Input Splits and Records, Text Input, Binary Input, Multiple Inputs),
- Output Formats (Text Output, Binary Output, Multiple Output).
- Distributed Cache

**Session: 10**

**Basics of Map Reduce Programming**

- Hadoop Data Types,
- Java and Map Reduce,
- Map Reduce program structure,
- Map-only program, Reduce-only program,
- Use of combiner and partitioner,
- Counters, Schedulers (Job Scheduling),
- Custom Writables, Compression

**Lab-Assignment:**

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

**Session: 11**

**Map Reduce Streaming**

- Complex Map Reduce programming,
- Map Reduce streaming,
- Python and Map Reduce,
- Map Reduce on image dataset

**Hadoop ETL**

**Session: 12**

- Hadoop ETL Development,
- ETL Process in Hadoop,
- Discussion of ETL functions,
- Data Extractions,
- Need of ETL tools,

- Advantages of ETL tools.

**Lab-Assignment:**

- Understand the file formats and read the provided links

**HBase (Theory – 06 Hrs & Lab – 06 Hrs)****Session: 13****Introduction to HBase**

- Overview of HBase
- HBase architecture
- Installation

**Session: 14 and 15****The HBaseAdmin and HBase Security**

- Various Operations on Tables
- HBase general command and shell,
- java client API for HBase
- Admin API
- CRUD operations
- Client API
- HBase – Scan, Count and Truncate
- HBase Security

**Lab-Assignment:**

- Run the Hbase shell commands
- Run the HBase using Java client

**Hive (Theory – 08 Hrs & Lab – 18 Hrs)****Session: 16****The Hive Data-ware House**

- Introduction to Hive,
- Hive architecture and Installation,
- Comparison with Traditional Database,
- Basics of Hive Query Language.

**Session: 17****Working with Hive QL**

- Datatypes,
- Operators and Functions,
- Hive Tables (Managed Tables and Extended Tables),
- Partitions and Buckets,
- Storage Formats,
- Importing data,
- Altering and Dropping Tables

**Lab-Assignment:**

- Create a hive DB and table ( internal and external )
- Load the data into hive table (using local inpath and HDFS inpath)

**Session:18****Querying with Hive QL**

- Querying Data-Sorting,

- Aggregating,
- Map Reduce Scripts,
- Joins and Sub queries,
- Views,
- Map and Reduce side joins to optimize query.

**Lab-Assignment:**

- Run all the types of joins in Hive
- Execute the data to be partitioned

**Session: 19****More on Hive QL**

- Data manipulation with Hive,
- UDFs,
- Appending data into existing Hive table,
- custom map/reduce in Hive
- Writing HQL scripts

**Apache Airflow (Theory – 06 Hrs & Lab – 06 Hrs)****Session: 20, 21 and 22**

- Introduction to Data Warehousing and Data Lakes
- Designing Data warehousing for an ETL Data Pipeline
- Designing Data Lakes for an ETL Data Pipeline
- ETL vs ELT
- Fundamentals of Airflow
- Work management with Airflow
- Automating an entire Data Pipeline with Airflow

**Lab-Assignment:**

- Create a airflow DAG for Extract -> Transform -> Load

**Introduction to Apache Spark& Kafka (Theory – 24 Hrs & Lab – 36 Hrs)****Session: 23, 24 and 25****Apache Spark APIs for large-scale data processing**

- Overview, Linking with Spark, Initializing Spark,
- Resilient Distributed Datasets (RDDs), External Datasets
- RDD v/s Data frames v/s Datasets
- Data frame operations
- Structured Spark Streaming
- Passing Functions to Spark, Working with Key-Value Pairs, Shuffle operations,
- RDD Persistence, Removing Data, Shared Variables, Deploying to a Cluster

**Lab-Assignment:**

- Run the provided Hadoop Streaming program using python

**Session: 26**

- Map Reduce with Spark
- Working with Spark with Hadoop
- Working with Spark without Hadoop and their Differences

**Lab Assignment**

- Execute all the provided code using step-runs for each and every codeline

- Setup the JDBC configuration and run the Spark JDBC Connectivity program
- Run the spark integrations using the provided code

**Session: 27**

- Data preprocessing
- EDA

**Session: 28 and 29**

- Introduction to Kafka
- Working with Kafka using Spark
- Spark streaming Architecture
- Spark Streaming APIs
- Building Stream Processing Application with Spark

**Lab Assignment**

- Execute the spark streaming with Kafka

**Session: 30**

- Setting up Kafka Producer and Consumer
- Kafka Connect API

**Session: 31**

- Spark SQL

**Lab Assignment**

- Run the sparkSQL programs using step-runs for each and every codeline
- Run all the SparkSQL programs
- Analyse the election data using spark and provide analysis

**Session: 32 and 33**

- Spark MLlib
- Predictive Analysis

**Lab Assignment:**

- Deep Learning with Spark
- Connecting DB's with Spark
- Accessing and manipulating the DB's
- Demo: Capstone Project
- Create a complex workflow using bash operator, a simple workflow using python
- Create Using python airflow operator to read data from your local drive, ingest the data into your HDFS, and perform a spark WC