*Suggested Teaching Guidelines for*

# *Big Data Technologies*

# *PG-DBDA August 2025*

**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. Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization, DT Editorial Services , Wiley India,Latest.

**Reference:**

1. Big Data, Black Book by DreamTech
2. Programming Hive by O'Rellay
3. Hadoop The Definitive Guide 4$^{th}$Edition by O'Reilly
4. Hadoop with python
5. Hadoop Real-World Solutions Cookbook by Packet publication
6. Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault
7. 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 of Theory & 2 Hours of Lab unless mentioned otherwise.***

**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

*Suggested Teaching Guidelines for*

***Big Data Technologies***

***PG-DBDA August 2025***

- 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**
**Hadoop Distributed File System (HDFS)**

- Distributed File System
- Introduction to HDFS
- Core components of HDFS
- HDFS Daemons
- Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node

**Session: 5**
**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 thecommand.
- Execute the provided code using HDFS, step run and understand

*Suggested Teaching Guidelines for*

# *Big Data Technologies*

# *PG-DBDA August 2025*

**Session: 6 (2T + 4L)**
**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

**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, MultipleInputs),
- 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 (2T + 4L)**
**Map Reduce Streaming**

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

- **Hadoop ETL**

**Session: 12 (2T + 4L)**

- 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

**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

**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:**

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

**Session: 18 (2T + 4L)**
**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

*Suggested Teaching Guidelines for*

# *Big Data Technologies*

# *PG-DBDA August 2025*

**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

**Session: 20, 21 & 22 (6T + 10L)**

- 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/Informatica
- Work management with Airflow/ Informatica
- Automating an entire Data Pipeline with Airflow/Informatica

**Lab-Assignment:**

- Create an airflow DAG/ Informatica for Extract -> Transform -> Load

**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 (2T + 4L)**

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

*Suggested Teaching Guidelines for*

# *Big Data Technologies*

# *PG-DBDA August 2025*

**Lab Assignment**

- Execute all the provided code using step-runs for each 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(4T + 8L)**

- Spark MLlib
- Predictive Analysis

*Suggested Teaching Guidelines for*

## *Big Data Technologies*

## *PG-DBDA August 2025*

**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 thedata into your HDFS, and perform a spark WC