Suggested Teaching Guidelines for

BigData Technologies – PG-DBDA Aug 19

Duration: 60 class room hours + 80 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

Text Book:

1. Hadoop: The Definitive Guide, SPD

Reference:

1. Big Data, Black Book by DreamTech
2. Programming Hive by O'Rellay (Author :- Edward Capriolo, Dean Wampler, and Jason RutherglenEdward Capriolo, Dean Wampler, and Jason Rutherglen)
1. Hadoop The Definitive Guide 4th 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 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
Session: 1
Lecture
Introduction to Big Data

- What is Big Data,
- Big Data Sources,
- Industries using Big Data,
- Big Data challenges
-
- Big Data Applications
- Various Big Data Technologies,
- Big Data/Hadoop Platforms,

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,

Suggested Teaching Guidelines for

BigData Technologies – PG-DBDA Aug 19

Hadoop Distributed File System (HDFS)

Session: 2 & 3

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

Hadoop Installation and Cluster Configuration (Lab – 02 Hrs)

Session: 4

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: 5 & 6

Hadoop Architecture

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

Session: 7

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: 8

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: 9

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

Session: 10

Map Reduce Streaming

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

Session: 11

Introduction to Hadoop ecosystem

- Hadoop Ecosystem
- Hadoop YARN
- Introduction to Hive, Pig, Sqoop, ZooKeeper, Flume, Oozie, Spark, HBase etc.

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.

HBase (Theory – 04 Hrs & Lab – 06 Hrs)

Session: 13

Introduction to HBase

- Overview of HBase
- HBase architecture
- Installation

Session: 14

The HBaseAdmin and HBase Security

- HBase general command and shell,
- java client API for HBase
- CRUD operations
- HBase Security

Suggested Teaching Guidelines for

BigData Technologies – PG-DBDA Aug 19

Hive (Theory – 08 Hrs & Lab – 16 Hrs)

Session: 15

The Hive Data-ware House

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

Session: 16

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.

Session: 17

Querying with Hive QL

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

Session: 18

More on Hive QL

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

PIG (Theory – 06 Hrs & Lab – 12 Hrs)

Session: 19

Introduction to PIG and PIG Latin

- Introduction to PIG,
- PIG vs Map Reduce,
- Pig Architecture and Installation
- Pig Execution Modes
- Running PIG,
- PIG Latin Statements.

Basics of PIG Latin Programming

- Conventions, Data Types,
- Arithmetic and Relational Operators,
- UDF Statements.
- PIG Latin Scripting,

Session: 20

PIG Built-In Functions

- Eval Functions, Load/Store Functions,
- Math Functions,
- String Functions,
- Date Time Functions,
- Tuple,
- Bag,
- Map Functions.

Session: 21
UDFs (user defined functions), Control Structures, Commands

- Writing a PIG UDF
- Piggy Bank
- Data Fu
- PIG Macros
- Parameter Substitution
- Shell and Utility Commands
- Combiner
- Use cases
- Real-Time Data Analytics using PIG

Introduction to Apache Spark & Kafka (Theory – 18 Hrs & Lab – 32 Hrs)

Session: 22, 23 and 24
Apache Spark APIs for large-scale data processing

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

Session: 25

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

Session: 26

- Introduction to Kafka
- Working with Kafka using Spark
- Spark streaming

Session: 27

- Spark MLlib

Session: 28

- Spark SQL

Session: 29

- Introduction to storm
- Comparison between Spark & Storm

Session: 30

- Using mongoDB with Spark
- Industrial Case studies