CERTIFICATE COURSE IN BIG DATA

(6 Months)

STATE BOARD OF TECHNICAL EDUCATION AND TRAINING

SANKETHIKA VIDYA BHAVAN, MASAB TANK, TELANGANA:HYDERABAD

Certificate Course in Big Data

Duration of the Course	**:**	**6 Months**
Eligibility	**:**	**Intermediate or its equivalent**
Total Teaching Hrs	**:**	**250 Hrs**

Scheme of Instruction and Examination

Sub Code	Subject Name	Instruction Period/Week		Total Periods	Scheme of Examination			
		Theory	Practical		Duration	Internal Marks	End Exam Marks	Total Marks
THEORY								
BD-101	Big Data Hadoop Programming	03	-	50	3Hrs	0	100	100
BD-102	Big Data - II	03	-	50	3Hrs	0	100	100
PRACTICALS								
BD-103	Big Data Hadoop Programing Lab-I	-	04	75	3Hrs	40	60	100
BD-104	Big Data Hadoop Programing Lab-II	-	04	75	3Hrs	40	60	100
	TOTAL	06	08	250		80	320	400

Certificate Course in Big Data

Subject Code	**:**	BD - 101
Subject Name	:	Big Data Hadoop Programming
Periods/Week	:	03 Hrs
Total Periods	:	50 Hrs

About Course

Hadoop is an Apache project (i.e. open source software) to store & process Big Data. Hadoop stores Big Data in a distributed & fault tolerant manner over commodity hardware. Afterwards, Hadoop tools are used to perform parallel data processing over HDFS (Hadoop Distributed File System).

Course Objectives: This course enables the students to

- Learn the concepts of Big Data and Hadoop including HDFS (Hadoop Distributed File System), YARN (Yet Another Resource Negotiator)
- Learn Hadoop Map Reduce Programming
- Configure Single Node Cluster in Hadoop Environment
- Learn Hadoop pre requisites

Course Outcomes: On completion of this course the students are able to

- Hadoop Cluster Architecture
- Write essential Java Programs
- Write Basic Linux commands
- Setup Single Node and Multi-Node Hadoop Cluster.
- Write Advanced Map Reduce concepts such as Counters, Distributed Cache

UNIT-1: Big Data Pre-requisites

Linux Fundamentals, SQL Essentials, Java Essentials for Big Data

UNIT-2: Introduction to Big Data

Introduction to Big Data & Big Data Challenges Preview, Limitations & Solutions of Big Data Architecture, Hadoop & its Features, Hadoop Ecosystem, Hadoop 2.x Core Components Preview

UNIT-3: Hadoop Distributed File System (HDFS)

Hadoop Storage: HDFS (Hadoop Distributed File System), HDFS user commands- ls, mkdir, touchz, copyFromLocal, put, cat, copyToLocal, get, moveFromLocal, cp, mv, rmr, du, dus, stat

UNIT-4: Hadoop Architecture

Hadoop 2.x Cluster Architecture Preview, Federation and High Availability Architecture Preview, Typical Production Hadoop Cluster, Hadoop Cluster Modes, Common Hadoop Shell Commands Preview , Hadoop 2.x Configuration Files ,Single Node Cluster & Multi-Node Cluster set up ,Basic Hadoop Administration

UNIT-5: Hadoop Map Reduce Framework

Traditional way vs. Map Reduce way, Why Map Reduce Preview, YARN Components, YARN Architecture, YARN Map Reduce Application Execution Flow, YARN Workflow Anatomy of Map Reduce Program Preview, Input Splits, Relation between Input Splits and HDFS Blocks, Map Reduce: Combiner & Partitioner

UNIT-6: Advanced Map Reduce Programming

Counters, Distributed Cache, MR unit, Reduce Join Preview, Custom Input Format Preview Sequence Input Format, XML file parsing using Map Reduce

Certificate Course in Big Data

Subject Code	**:**	BD - 102
Subject Name	:	Big Data - II
Periods/Week	:	03 Hrs
Total Periods	:	50 Hrs

About Course

As organizations have realized the benefits of Big Data Analytics, so there is a huge demand for Big Data & Hadoop professionals. Companies are looking for Big data & Hadoop experts with the knowledge of Hadoop Ecosystem and best practices about HDFS, Map Reduce, Spark, HBase, Hive, Pig, Oozie, Sqoop & Flume.

Course Objectives: This course enables the students to

- Acquire the knowledge of various tools that fall in Hadoop Ecosystem like Pig, Hive, Sqoop, Flume, Oozie, and HBase
- Ingest data in HDFS using Sqoop & Flume, and analyze those large datasets stored in the HDFS
- Implement the Projects which are diverse in nature covering various data sets from multiple domains such as banking, telecommunication, social media, insurance, and e-commerce

Course Outcomes: On completion of this course the students are able to

- Apply the Data Loading Techniques using Sqoop & Flume
- Write PIG Scripts
- Write HIVE Queries, Table Partition, bucketing and UDF
- Understand the Hive concepts, Hive Data types
- Import and export data with SQOOP
- Write NO SQL queries with HBASE
- Configure and Implement Apache Spark

UNIT-1: Apache PIG

Introduction to Apache Pig, Map Reduce vs Pig, Pig Components & Pig Execution, Pig Data Types & Data Models in Pig, Pig Latin Programs, Shell and Utility Commands, Pig UDF & Pig Streaming

UNIT-2: Apache HIVE

Introduction to Apache Hive, Hive vs Pig, Hive Architecture and Components, Hive Meta store, Limitations of Hive, Comparison with Traditional Database, Hive Data Types and Data Models, Hive Partition, Hive Bucketing, Hive Tables (Managed Tables and External Tables) Importing Data, Querying Data & Managing Outputs, Hive Script & Hive UDF

UNIT-3: Apache HBASE

Apache HBase: Introduction to No SQL Databases and HBase Preview, HBase v/s RDBMS HBase Components, HBase Architecture, HBase Run Modes, HBase Configuration, HBase Cluster Deployment

UNIT-4: Apache Sqoop

Sqoop Introduction, Why Sqoop, Sqoop Features, Flume vs Sqoop, Sqoop Architecture & Working, Sqoop Commands

UNIT-5: Processing Distributed data with Apache Spark

What is Spark, Spark Ecosystem, Spark Components, What is Scala, Why Scala, Spark Context, Spark RDD.

UNIT-6: CASE STUDY

- Analysis of Online Book Store
- Click stream Analysis

Certificate Course in Big Data

Subject Code	**:**	BD - 103
Subject Name	:	Big Data Hadoop Programming LAB-1
Periods/Week	:	04 Hrs
Total Periods	:	75 Hrs

Course Objectives: This course enables the students to

- Learn the concepts of Big Data and Hadoop including HDFS (Hadoop Distributed File System), YARN (Yet Another Resource Negotiator)
- Learn Hadoop Map Reduce Programming
- Configure Single Node Cluster in Hadoop Environment
- Learn Hadoop pre requisites

List of Experiments

1. Installation of Hadoop Single Node Cluster Setup with all pre requisites.
2. Do the following File Management tasks in HDFS
   a. Create a directory in HDFS at given path
   b. List the content of a directory
   c. Upload and Download content a file in HDFS
3. Do the following File Management tasks in HDFS
   a. Copy a file from source to destination
   b. Copy a file from/to local file system to HDFS
   c. Move files from source to destination
   d. Remove a file or directory in HDFS
4. Write a Map Reduce program to find word count with three following functions
   a. Mapper Code
   b. Reducer Code
   c. Driver Code
5. Weather Report POC-Map Reduce Program to analyse time-temperature statistics and generate report with max/min temperature.
6. Implementing Matrix Multiplication with Hadoop Map Reduce.

Certificate Course in Big Data

Subject Code	**:**	BD - 104
Subject Name	:	Big Data Hadoop Programming LAB-II
Periods/Week	:	04 Hrs
Total Periods	:	75 Hrs

Course Objectives: This course enables the students to

- Acquire the knowledge of various tools that fall in Hadoop Ecosystem like Pig, Hive, Sqoop, Flume, Oozie, and HBase
- Ingest data in HDFS using Sqoop & Flume, and analyze those large datasets stored in the HDFS
- Implement the Projects which are diverse in nature covering various data sets from multiple domains such as banking, telecommunication, social media, insurance, and e-commerce

List of Experiments

1. a) Write a Pig Latin scripts to sort, group, join, project, and filter your data.

   b) Write a Pig Latin Scripts to find Word Count.

2. Run the Pig Latin Scripts to find a max temp for each and every year.

3. Create Hive Databases, Tables, Views, Functions and Indexes.

4. Create External Partition Tables in HIVE

5. Adding, Modifying, and Dropping a Table Partition using HIVE Query Language

6. a) Create user defined functions (UDF) in HIVE.

   b) Implement importing and exporting using SQOOP.