Marwadi University
Faculty of Diploma Studies
Information and Communication Technology

Subject Code: 09CT0606
Subject Name: Big Data Analytics
Diploma Year – III (Semester VI)

Objective:

Big data is an extremely useful area in the era of computing techniques as it aids in finding useful patterns from large datasets. Large datasets are so huge that they cannot be processed with traditional technologies. We require a special computing system which can handle large data and tandem it with other important aspects like parallel processing, data failure and data pre-processing.

Course Outcomes: After completion of this course, student will be able to

- Gain Understanding about Big Data Technology and its Tools. (Understand)
- Understand and apply extracting useful patterns from large datasets. (Apply)
- Implementation of Big data mining techniques using different software. (Create)
- Understand how data analytics and data science maps to current industry.(Analyze)
- Understanding and implementing Algorithms in an optimized way using various Big Data Tools. (Apply)

Prerequisite of course: Basic knowledge of C language

Teaching and Examination Scheme

Teaching Scheme (Hours)			Credits	Theory Marks			Tutorial/Practical Marks		Total Marks
Theory	Tutorial	Practical		ESE	IA	CSE	Viva	Term Work (TW)	
3	0	2	4	50	30	20	25	25	150

Contents:

S.NO	TOPIC	Teaching hours
1	**Introduction to Big Data** Introduction-Distributed file System, What is Big Data? Difference between traditional Distributed file system and Big Data Software, Big Data Analytics, Big data Applications.	8

2	**Introduction to Hadoop:** How Hadoop works? Hadoop Architecture, Explanation of Hadoop EcoSystem, Hadoop Basic commands	8
3	**Hadoop Input and Output:** Data Integrity in Hadoop, Data Compression and Data Serialization in Hadoop, Avro, How Avro works?	8
4	**Hadoop MapReduce:** Mapper, Reducer, MapReduce YARN, Job Scheduling, Sorting and Shuffling in MapReduce, MapReduce Input Formats, MapReduce Output Formats, How to code in MapReduce program , analyze data using MapReduce.	8
5	**Hadoop Ecosystem/Environment: Pig, Hive, Hbase, ZooKeeper** Pig Latin Structures, Statements, Functions, User-Defined Function in Pig, Loading, Storing and Sorting Data in Pig, HiveQL, Tables in Hive, Querying Data, User-Defined Function in Hive, Introduction to HBase, HBASE vs RDBMS, What is ZooKeeper, Zoopkeeper Services, Build Application with ZooKeeper.	8
6	**Apache Spark:** Introduction to Apache Spark, pySpark, RDD, Working with Key-value pair, Loading and saving data in spark, Learning about Machine Learning Library in Spark.	8
	Total Hours	**48**

References:

1. Tom White, “HADOOP: The definitive Guide”, O Reilly 2012.
2. BIG Data and Analytics , Sima Acharya, Subhashini Chhellappan, Willey
3. MongoDB in Action, Kyle Banker,Piter Bakkum , Shaun Verch, Dream tech Press
4. Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Karau

Suggested Theory distribution:

The suggested theory distribution as per Bloom’s taxonomy is as per follows. This distribution serves as guidelines for teachers and students to achieve an effective teaching learning process.

Distribution of Theory for course delivery and evaluation					
Remember	**Understand**	**Apply**	**Analysis**	**Evaluate**	**Create**
40%	40%	20%	20%	0%	0%

Suggested List of Experiments:

1. Installation and use of Hadoop in ubuntu.
2. Run HDFS commands in a hadoop environment.
3. Implementation of a MapReduce Algorithm.
4. Hive Installation.
5. Run Hive related commands on given data.
6. UDF creation in Hive to truncate blank space.
7. Install HBASE and Apply various table queries.
8. Install MongoDB and execute basic commands in MongoDB Shell.
9. Connect MongoDB with java.
10. Install Scala and program in interactive mode and script mode.
11. Run a job on Apache spark.