

COURSE TITLE	BIG DATA AND ANALYSIS
COURSE CODE	01CE1632
COURSE CREDITS	3

Objective:

- 1 Big data is an extremely useful area in the era of computing techniques as it aids in finding useful pattern from large datasets. Large datasets are so huge that they cannot be processed with traditional technologies. We require 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:

- 1 Understanding about Big Data Technology and its Tools.
- 2 Apply extracting useful pattern from large datasets.
- 3 Implementation of Big data mining techniques using different software.
- 4 Understand how data analytics and data science maps to current industry.
- 5 Implementing Algorithms in an optimized way using various Big Data Tools.

Pre-requisite of course: Basic Programming Knowledge, Data Mining.

Teaching and Examination Scheme

Theory Hours	Tutorial Hours	Practical Hours	ESE	IA	CSE	Viva	Term Work
3	0	0	50	30	20	0	0

Contents : Unit	Topics	Contact 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	4
2	Introduction to Hadoop How Hadoop works?, Hadoop Architecture, Explanation of Hadoop EcoSystem, Hadoop Basic commands	7
3	Hadoop Input and Output Data Integrity in Hadoop, Data Compression and Data Serialization in Hadoop, Avro & How Avro works?	7
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	12

Contents : Unit	Topics	Contact Hours
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, Zookeeper Services, Build Application with ZooKeeper	12
Total Hours		42

Textbook :

- 1 HADOOP: The definitive Guide, Tom White, O Reilly, 2012
- 2 BIG Data and Analytics, Sima Acharya, Subhashini Chhellappan, Wiley, 2019
- 3 MongoDB in Action, Kyle Banker, Peter Bakkum, Shaun Verch, Douglas Garrett, and Tim Hawkins, Dreamtech Press, 2016
- 4 Learning Spark: Lightning-Fast Big Data Analysis, Holden Karau, O Reilly, 2015

Suggested Theory Distribution:

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

Distribution of Theory for course delivery					
Remember / Knowledge	Understand	Apply	Analyze	Evaluate	Higher order Thinking / Creative
10.00	20.00	40.00	20.00	10.00	0.00

Instructional Method:

- 1 The course delivery method will depend upon the requirement of content and need of students. The teacher in addition to conventional teaching method by black board, may also use any of tools such as demonstration, role play, Quiz, brainstorming, MOOCs etc.
- 2 The internal evaluation will be done on the basis of continuous evaluation of students in the class-room
- 3 Students will use supplementary resources such as online videos, NPTEL videos, e-courses, Virtual Laboratory

Supplementary Resources:

- 1 <http://www.bigdatauniversity.com/>
- 2 <http://hadoop.apache.org/>
- 3 <https://spark.apache.org/documentation.html>
- 4 <http://www.prajval.in/edudetail/18>
- 5 <https://www.tutorialspoint.com/mongodb/index.htm>