COURSE TITLE	**BIG DATA ANALYTICS**
COURSE CODE	**01CT1622**
COURSE CREDITS	**4**

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 Gain Understanding about Big Data Technology and its Tools.

2 Understand and 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 Understanding and 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	2	50	30	20	25	25

Contents : Unit	Topics	Contact Hours
1	**Introduction to Big Data** Introduction-Distributed file System, What is Big Data?, Introducti Difference between traditional Distributed file system and Big Data Software, Big Data Analytics, Big data Applications	3
2	**Introduction to Hadoop** How Hadoop works? , Hadoop Architecture, Explanation of Hadoop EcoSystem, Hadoop Basic commands.	5
3	**Hadoop Input and Output** Data Integrity in Hadoop, Data Compression and Data Serialization in Hadoop, Avro, How Avro works?	6
4	**Hadoop MapReduce** Mapper, Reduce, MapReduce YARN, Sorting and Shuffling in MapReduce, MapReduce Input Formats, MapReduce Output, How to code in MapReduce program, analyze data using MapReduce	7

MR. CHANDRASINH DHIRUBHA PARMAR
Digitally signed by (Name of HOD)

DR. RAJENDRASINH BAHADURSINH JADEJA
Digitally signed by (Name of Dean/ Principal)

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, 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, Working with Key-value pair, Loading and saving data in spark, Learning about Machine Learning Library in Spark.	7
7	**NoSql** Introduction to NoSql, NoSql vs SQL, Introduction to MongoDB, MongoDB Create-Drop Databases, Create-Drop Collection, CRUD operation in documents, MongoDB indexing, Aggregation, replication, Connect Java Application with MongoDB.	6
Total Hours		**42**

Suggested List of Experiments:

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

MR. CHANDRASINH DHIRUBHA PARMAR
Digitally signed by (Name of HOD)

DR. RAJENDRASINH BAHADURSINH JADEJA
Digitally signed by (Name of Dean/ Principal)

Suggested List of Experiments:

Contents : Unit	Topics	Contact Hours
12	**Experiment-12** Create and run SQL and NOSQL queries	2
Total Hours		**24**

Textbook :

1 HADOOP: The definitive Guide, Tom White, O Reilly, 2012

2 BIG Data and Analytics, Sima Acharya, Subhashini Chhellappan, Willey, 2015

References:

1 MongoDB in Action, MongoDB in Action, Kyle Banker,Piter Bakkum , Shaun Verch, Dream tech Press, 2016

2 Learning Spark: Lightning-Fast Big Data Analysis, Learning Spark: Lightning-Fast Big Data Analysis, Holden Karau, Shroff/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	25.00	25.00	10.00	10.00

Instructional Method:

1 The internal evaluation will be done on the basis of continuous evaluation of students in the laboratory and class-room.

2 Practical examination will be conducted at the end of the semester for evaluation of performance of students in laboratory.

3 Students may use supplementary resources such as online videos, NPTEL videos, e-courses, Virtual Laboratory, etc.

4 The course delivery method will depend upon the requirement of content and need of the students. The teacher in addition to conventional teaching method (Chalk and Talk) may use any of the tools such as demonstration, role play, Quiz, brainstorming, Flipped class, Project based learning, Collaborative learning, MOOCs etc. for effective teaching.

Supplementary Resources:

1 http://in.reuters.com/tools/rss

2 http://www.altova.com/xmlspy.html

3 https://www.w3.org/RDF/