COURSE TITLE	**BIG DATA ANALYTICS**
COURSE CODE	**01CE0719**
COURSE CREDITS	**3**

Objective:

1 The objective of this syllabus is to equip students with the knowledge and skills necessary to effectively manage, analyze, and derive insights from large and complex datasets.

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

1 Understand the fundamental concepts, characteristics, and architecture of Big Data and its storage mechanisms.

2 Apply the Hadoop framework, including HDFS, YARN, and MapReduce to perform distributed data processing tasks.

3 Develop and execute HiveQL queries for data definition, manipulation, and analytical operations.

4 Design, analyse and create complex data by applying advanced techniques

5 Demonstrate integrated techniques like Spark for big data

Pre-requisite of course:NA

Teaching and Examination Scheme

Theory Hours	Tutorial Hours	Practical Hours	ESE	IA	CSE	Viva	Term Work
2	0	2	50	30	20	25	25

Contents : Unit	Topics	Contact Hours
1	**Understanding Big Data** Concepts and terminology, Big Data Characteristics,Different types of Data, Identifying Data Characteristics – Big Data Architecture – Big Data Storage: File System and Distributed File System, NoSQL, Sharding, Replication, ACID and BASE Properties	5
2	**Introduction Hadoop** Introduction Hadoop; comparisons of RDBMS and Hadoop, Distributed Computing Challenges, Hadoop Overview, Business Value of Hadoop,, Hadoop Distributed File System, NameNode, Secondary NameNode, and DataNode, Hadoop MapReduce paradigm, Map and Reduce tasks, Job, Task trackers Cluster Setup SSH & Hadoop Configuration - HDFS Administering Monitoring & Maintenance., Hadoop Yarn, Hadoop in the Cloud, Applications on Big Hadoop Ecosystem.	7

MR. HARDIK AJAYBHAI DOSHI
Digitally signed by (Name of HOD)

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

Contents : Unit	Topics	Contact Hours
3	**Introduction to Hive** Hive Architecture, Hive modules, Data types and file formats, Hive QL -Data Definition and Data Manipulation – Hive QL queries, Sorting and Aggregation, Hive QL views – reduce query complexity, Hive scripts, Hive QL Indexes – Aggregate functions – Bucketing vs Partitioning	5
4	**HBASE, PIG and Zookeeper** HBase concepts and models- Schema Design, Advance Indexing – PIG concepts – Data Types, Latin Concepts, Run Modes, Zookeeper – how it helps in monitoring a cluster, Benefits and Challenges of Zookeeper	5
5	**Spark** Overview of Spark – Hadoop Overview of Spark – Hadoop Vs Spark – Cluster Design – Cluster Management – performance, Application Programming Interface (API), Spark Context, Resilient Distributed Datasets, Creating RDD, RDD Operations, and Saving RDD, Lazy Operation – Spark Jobs.	6
Total Hours		**28**

Suggested List of Experiments:

Contents : Unit	Topics	Contact Hours
1	**Practical – 1** Installation of Hadoop on Windows	2
2	**Practical – 2** Implementing single node cluster on Hadoop Environment	2
3	**Practical – 3** Running Namenode, Datanode and working on different commands	2
4	**Practical – 4** Run HDFS commands in Hadoop environment	2
5	**Practical – 5** Run Counting Program using Mapreduce in Hadoop	2
6	**Practical – 6** Performing queries on MongoDB terminal	2
7	**Practical – 7** Performing Different Operations on MongoDB shell	2
8	**Practical – 8** Installation of Hive	2
9	**Practical – 9** Run all Hive Commands on a given data	2
10	**Practical – 10** Running HBase on framework	2
11	**Practical – 11** Executing a Script in Apache PIG	2

MR. HARDIK AJAYBHAI DOSHI
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	**Practical – 12** Comparing results of different dataset on Tableau	2
13	**Practical – 13** Implementing spark environment	2
14	**Practical – 14** An implementation of case study of detecting Medicare Fraud Detection using Big Data Analytic Tools	2
Total Hours		**28**

Textbook :

1 "Big Data Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses",, Michael Minelli, Michele Chambers, Ambiga Dhiraj, Wiley India, 2013

References:

1 "Understanding Big data", "Understanding Big data", Chris Eaton, Dirk derooset al, McGraw Hill,, 2012

2 "HADOOP: The Definitive Guide",, "HADOOP: The Definitive Guide",, Tom White, O Reilly , 2012

3 Learning Spark: Lightning-Fast Big Data Analysis Paperback, Learning Spark: Lightning-Fast Big Data Analysis Paperback, Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly Media, 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**
15.00	15.00	25.00	25.00	15.00	5.00

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 laboratory and class-room.

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

4 Students will use supplementary resources such as online videos, NPTEL videos, e-courses, Virtual Laborator

MR. HARDIK AJAYBHAI DOSHI
Digitally signed by (Name of HOD)

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

Supplementary Resources:

1 http://in.reuters.com/tools/rss
2 http://www.altova.com/xmlspy.html
3 https://www.w3.org/RDF/