NIRMA UNIVERSITY

Institute:	Institute of Technology
Name of Programme:	B.Tech. Computer Science and Engineering
Course Code:	2CS702
Course Title:	Big Data Analytics
Course Type:	Core
Year of Introduction:	2021-22

Credit Scheme

L	T	Practical Component				C
		LPW	PW	W	S	
2	0	2	-	-	-	3

Course Learning Outcomes (CLO):

At the end of the course, students will be able to -

1. outline the significance and challenges of big data
2. model big data using different tools and frameworks
3. apply big data techniques for useful business analytic applications
4. design algorithms for mining the data from large volumes

Syllabus: **Total Teaching hours:**
30

Unit	Syllabus	Teaching hours
Unit-I	**Introduction:** Evolution of Big Data, Types of Digital Data, Classification of Digital Data, Structured Data, Semi-Structured Data, Unstructured Data, Definition of Big Data, Challenges of Conventional Systems, Big data platforms and data storage	04
Unit-II	**Big Data Analytics:** Importance of Big data analytics, Classification of Analytics, Top Challenges Facing Big Data, Technologies to meet the Challenges Posed by Big Data, Terminologies Used in Big Data Environment	04
Unit-III	**Hadoop:** Introducing Hadoop, comparisons of RDBMS and Hadoop, Distributed Computing Challenges, Hadoop Overview, Business Value of Hadoop, Hadoop Distributed File System, Processing Data with Hadoop, working with Map Reduce, Hadoop YARN, Hadoop in the Cloud, Applications on Big Hadoop Ecosystem, Fundamentals of Pig, Hive, HBase and ZooKeeper, Basic concepts of Apache Spark	08
Unit-IV	**The Big data technology landscape:** CAP Theorem - BASE Concept, NoSQL, Types of No SQL databases, Introduction to MongoDB, Data Types in MongoDB, CRUD, Apache Cassandra, Features of Cassandra, CRUD	08
Unit-V	**Big data analytics Algorithm:** Applying Linear Regression, Clustering, Association rule mining, Decision tree on Big Data.	06

Self-Study: The self-study contents will be declared at the commencement of semester. Around 10% of the questions will be asked from self-study contents

Suggested Readings/ References:

1. Michael Berthold, David J. Hand, Intelligent Data Analysis, Springer
2. Tom White, Hadoop: The Definitive Guide, O'reilly Media
3. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill Publishing
4. Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets, Cambridge University Press
5. Bill Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, John Wiley & sons
6. Glenn J. Myatt, Making Sense of Data, John Wiley & Sons
7. Da Ruan, Guoquing Chen, Etienne E.Kerre, GeertWets, Intelligent Data Mining, Springer
8. Paul Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, James Giles, David Corrigan, Harness the Power of Big Data the IBM Big Data Platform, Tata McGraw Hill Publications
9. Michael Minelli, Michele Chambers, Ambiga Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, Wiley Publications
10. Zikopoulos, Paul, Chris Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Tata McGraw Hill Publications
11. Seema Acharya and Subhashini C, Big Data and Analytics, Wiley India

Suggested List of Experiments:

Sr. No.	**Practical Title**	**Hours**
1	Learning limitation of data analytics by applying Machine Learning Techniques on large amount of data. Write R/Python program to Read data set from any online website, excel file and CSV file and to perform a) Linear regression and logistic regression on iris dataset. b) K-means clustering.	02
2	Setup single node Hadoop cluster and apply HDFS commands on single node Hadoop Cluster. (*students can setup multimode cluster in laboratory)	04

3.	Apply MapReduce algorithms to perform analytics on single node cluster: a) Analyze phrase frequency from given dataset b) Search Records with matching criteria c) Aggregate inputs and search records based on aggregation	04
4	Analyze impact of different number of mapper and reducer on same definition as practical 3.	02
5	Implement PCY/Multi-Hash/SON algorithm for identification of frequent item set by handling larger datasets in main memory.	02
6	Setup the MongoDB environment in your system. Import Restaurant Dataset and perform CRUD operation.	02
7	Extend MongoDB functionality for MapReduce on document collection	02
8	SPark SQL and MLLib: (i) PYspark shell exploration and reading and writing in HDFS (ii) Clustering using MLlib , compare results of clustering with Hadoop MR and with Spark	02
9	Identify a case study to perform analytics on different platforms (like NoSQLs, Spark, Zookeeper and analyse differences.	04
10	Case study: Use the following platforms for solving any big data analytic problem of your choice. (1) Amazon web services, (2) Microsoft Azure, (3) Google App engine	02

Suggested Case List: -NA-