NIRMA UNIVERSITY

Institute:	Institute of Technology
Name of Programme:	Integrated B.Tech.(CSE)-MBA
Course Code:	CSI0904
Course Title:	Big Data Technologies
Course Type:	(□ Core/ □ Value Added Course / **√ Department Elective** / □ Institute Elective/ □ University Elective/ □ Open Elective / □ Any other)
Year of Introduction:	2022-23

L	T	Practical Component				C
		LPW	PW	W	S	
3	-	2	-	-	-	4

Course Learning Outcomes (CLOs):

At the end of the course, the student will be able to –

1. outline the significance and challenges of big data (BL2)
2. apply big data techniques for useful business analytic applications (BL3)
3. compare different big data using tools and frameworks (BL4)
4. design algorithms for mining the data from large volumes (BL6)

Syllabus: **Total Teaching hours: 30**

Unit	Syllabus	Teaching hours
Unit-I	**Introduction to Big Data:** 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	**Apache Hadoop:** Introducing Hadoop, comparisons of RDBMS and Hadoop, Distributed Computing Challenges, Hadoop Overview, Hadoop Ecosystem, Business Value of Hadoop, Hadoop Distributed File System, Processing Data with Hadoop, Map Reduce programming fundamentals, Hadoop YARN, Hadoop in the Cloud, Limitations of Hadoop, 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 in MongoDB, Apache Cassandra, Features of Cassandra, CRUD in Cassandra	08
Unit-V	**Big data analytics Algorithm:** 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, Geert Wets, 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.	Title	Hours
1	Study and explore various applications of big data in different domains. Choose one of it and study in detail, also write down the report on different types of digital data generated in selected application.	02
2	Learning limitation of data analytics by applying Machine Learning Techniques on large amount of data. Write a 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. Students will learn the limitation of platform and algorithm.	02
3	Setup single node Hadoop cluster and apply HDFS commands on single node Hadoop Cluster.	02
4	Design MapReduce algorithms to take a very large file of integers and produce as output: a) The largest integer b) The average of all the integers.	02

5	Apply MapReduce algorithms to find phrase frequency from given dataset. Prepare a report to guide design of mapper and reducer.	02
6	Analyse impact of different number of mapper and reducer on same definition as practical 4. Prepare a conclusive report on analysis.	02
7	Setup MongoDB environment in your system. Import Restaurant Dataset and perform CRUD operation.	02
8	Setup Cassandra environment in your system and apply Create, Update, Read and Delete operations.	02
9	Implement any one of the analytic algorithms using MapReduce by handling larger datasets in main memory. • PCY/Multi-Hash/SON algorithm • Regression • K-means Clustering	02
10	Case study: Use 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-