NIRMA UNIVERSITY

Institute:	**Institute of Technology, School of Technology**
Name of Programme:	**BTech CSE**
Course Code:	4CS101ME25
Course Title:	**Big Data Systems**
Course Type:	**Department Elective-III**
Year of Introduction:	**2025-26**

L	**T**	**Practical Component**				**C**
		LPW	**PW**	**W**	**S**	
3	0	2	-	-	-	4

Course Learning Outcomes (CLO):

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

1. outline the significance and challenges of big data (BL2)
2. model big data applications using various platforms (BL3)
3. utilise big data systems for practical business analytics (BL3)
4. design data analytics algorithms for various datasets. (BL6)

Unit	**Contents**	**Teaching Hours (Total 45)**
Unit-I	**Introduction to Big Data and Big Data Storage Platforms**: Evolution of Big Data, Types of Big Data, Definition of Big Data, Importance of Big data analytics, Challenges of Conventional Systems, Big data platforms and data storage	06
Unit-II	**Hadoop and HDFS:** Hadoop Ecosystem, Comparisons of RDBMS and Hadoop, Distributed Computing Challenges, Hadoop Overview, Processing Data with Hadoop, Hadoop YARN, Hadoop Ecosystem	06
Unit-III	**MapReduce:** working with Map Reduce, Anatomy of a Map Reduce Job Run, Failures, Job Scheduling, Shuffle, and Sort, Task Execution, Map Reduce Types and Formats, Map Reduce Features	08
Unit-IV	**Big data Machine Learning with Spark:** Basic concepts of Apache Spark, Spark - RDDs, DataFrames, PySpark, NumPy, SciPy, and Spark ML library, big data algorithms for Linear Regression, Clustering, Association rule mining, Decision tree	15
Unit-V	**NoSQL Database:** CAP Theorem - BASE Concept, NoSQL, Types of No SQL databases, Introduction to MongoDB, Data Types in MongoDB, CRUD operations	10

Self-Study:

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

Suggested Readings/ References:

1. Bill Chambers and Matei Zaharis, *Spark: The Definitive Guide: Big Data Processing Made Simple*, O'Reilly
2. Michael Berthold, David J. Hand, *Intelligent Data Analysis*, Springer
3. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill
4. Anand Rajaraman and Jeffrey David Ullman, *Mining of Massive Datasets*, Cambridge University Press
5. Seema Acharya and Subhashini C, *Big Data and Analytics*, Wiley

Suggested List of Experiments:

Sr.	Name of Experiments/Exercises	Hours
1	Study and explore various applications of big data in different domains. Choose one of them and study in detail. Also, write down the report on different types of digital data generated in selected applications. For example.: • Big Data in Retail • Big Data in Healthcare • Big Data in Education • Big Data in E-commerce • Big Data in Media and Entertainment • Big Data in Finance • Big Data in Travel Industry • Big Data in Telecom • Big Data in Automobile	04
2	Learning limitations of data analytics by applying Machine Learning Techniques on large amounts of data. Write a program to read data sets from any online website, excel file, and CSV file and to perform a) Linear regression and logistic regression on the iris dataset. b) K-means clustering.	02
3	Set up a single-node Hadoop cluster and apply HDFS commands to the single-node Hadoop Cluster.	04
4	Design MapReduce algorithms to take a very large file of integers and produce as output: • The largest integer • The average of all the integers. • The same set of integers, but with each integer appearing only once. • The count of the number of distinct integers in the input.	04
5	Apply MapReduce algorithms to find phrase frequency from a given dataset. Prepare a report to guide the design of the mapper and reducer.	02
6	Analyze the impact of different numbers of mappers and reducers on the same definition as practical 4. Prepare a conclusive report on the analysis.	02
7	Implement regression or classification analytic algorithms using MapReduce by handling given datasets using PySpark.	04
8	Implement any one of the analytic algorithms of the clustering application using PySpark.	02
9	Set up the MongoDB environment in your system. Import restaurant dataset and perform CRUD operation.	04
10	Case study: Use open-source platforms to solve any big data analytic problem of your choice.	02