LNMIIT, Jaipur
Department of Computer Science & Engineering

Programme: B.Tech (CSE), (CCE), and Integrated Course	Course Title: Big Data Computing			Course Code: CSE-----
Type of Course: Program Elective	Prerequisites: Advanced Programming, Introduction to Data Science			Total Contact Hours: 40
Year/Semester: 4/Odd	Lecture Hrs./Week: 3	Tutorial Hrs./Week: 0	Practical Hrs./Week:0	Credits: 3

Learning Objective:

Extreme data volume, velocity, and variety challenge conventional data-processing platforms and practices. Big data discipline trades some advantages of the established approaches to surmount the limitations of conventional storage infrastructures, data structures, databases and algorithms. The course provides an understanding of the needs, purposes, and characteristics of the Big Data domain.

The students will gain an understanding of the platforms for executing big data applications, algorithms, and analytical libraries. Hadoop and Spark frameworks will guide the students in learning about the execution platforms that grow linearly with the problem size. The students will also learn how these systems stay resilient and tolerant against failures. The programming language Scala will be introduced as it provides the base for building Apache Spark Analytical libraries. The libraries contain algorithms and techniques for solving big data problems.

Course outcomes (COs):

On completion of this course, the students will have the ability to:		**Bloom's Level**
S. No.	**Course Outcomes**	
CO1	Understand the fundamental of Big Data and issues in Big Data.	**L2**
CO2	Explore and apply Hadoop's architecture and ecosystem components, through practical hands-on activities.	**L3**
CO3	Explore and apply Spark using Scala for large-scale data processing, focusing on core functionalities like RDDs and job optimization and applying for real life problems.	**L3**
CO4	Apply Spark to solve and analyses real life problems in big data environment.	**L4**

Course Topics	Lectures
UNIT – I Introduction	**2**
1.1 Introduction to big data, Three Vs: Volume, Velocity, Variety, other properties of big data. Big Data Enabling Technologies.	2
UNIT – II Apache Hadoop	**14**
2.1 Introduction of Big Data Programming-Hadoop, Components of Hadoop, The Hadoop Distributed File System (HDFS), Design of HDFS, Java interfaces to HDFS, Hands-on experience with HDFS, Fault tolerance in HDFS.	5
2.2 Hadoop cluster setup and configuration, MapReduce introduction, and working, developing Map Reduce Applications (word count problem) and limitations of Map Reduce.	6
2.3 Data placement strategies, need of YARN, features, components of YARN, application work flow in YARN, Fault tolerance in YARN.	3
UNIT – III Apache Spark	**12**
3.1 Scala for Spark, basic operations, Scala essentials, OOPs, and FP, Practice some Scala code.	6
3.2 Spark-overview, ecosystem, Spark execution model, Hands-on: Installation, Programs in Command line Interface & IDE, Processing of Local, and HDFS files.	6
UNIT – IV RDD Fundamentals and Spark Application	**12**
4.1 RDD fundamentals, purpose, and structure, Transformations, Actions and DAG	3
4.2 Key-Value Pair in RDDs, Spark RDD Fault tolerance. Hands-On: Creating RDDs from Data Files, Interactive Queries Using RDDs.	4
4.2 Spark SQL, architecture, SQLContext in Spark SQL, working with DataFrames and DataSets, Hands-on: Creating (CSV, JSON) DataFrames, Querying with DataFrame API and SQL.	5

Textbook References (IEEE format) :

1. [TW] Tom White, Hadoop: The Definitive Guide, 4th Edition, O'Reilly, 2015.
2. [HADOOP] http://hadoop.apache.org/
3. [SPARK] https://spark.apache.org/
4. [SPK] Bill Chambers and Matei Zaharia, SPARK: The Definitive Guide, O'Reilly Media, Inc, 2017.
5. [JP] Jean-Georges Perrin, Spark IN ACTION, 2020.
6. [AS] Abdulhamit Subasi, Practical Machine Learning for Data Analysis Using Python, Academic Press, 2020.

Evaluation Method	

Item	Weightage (%)	Associated CO
Mid-Term	20	CO1,CO2
Continuous evaluation (Quizzes, Assignments)	20	CO1,CO2,CO3,CO4
Projects	20	CO2,CO3,CO4
End-term	40	CO1,CO2,CO3,CO4

*Please note, as per the existing institute's attendance policy the student should have a minimum of 75% attendance. Students who fail to attend a minimum of 75% lectures will be debarred from the End Term/Final/Comprehensive examination.

CO and PO Correlation Matrix

For CSE Students

CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1	3	3	3	2	3	1	1		3	3	3	3	3	3	3
CO2	3	3	3	2	3	1	1		3	3	3	3	3	3	3
CO3	3	3	3	2	3	1	1		3	3	3	3	3	3	3
CO4	2	2	2	1	3	1	1		2	2	2	3	2	2	2

For CCE Students

CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1	3	3	3	2	3	1	1		3	3	3	3	1	2	3
CO2	3	3	3	2	3	1	1		3	3	3	3	1	2	3
CO3	3	3	3	2	3	1	1		3	3	3	3	2	2	3
CO4	2	2	2	1	3	1	1		2	2	2	3	2	2	2