राष्ट्रीय प्रौद्योगिकी संस्थान पटना / NATIONAL INSTITUE OF TECHNOLOGY PATNA
संगणक विज्ञान एंव अभियांत्रिकी विभाग / DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
अशोक राजपथ, पटना-८००००५, बिहार / ASHOK RAJPATH, PATNA-800005, BIHAR
Phone No.: 0612-2372715, 2370419, 2370843, 2371929 Ext- 200, 202 Fax-0612-2670631 Website: www.nitp.ac.in

No:- Date:

CSX4152: Big Data Analytics

L-T-P-Cr: 2-0-2-3

Pre-requisites: Fundamental knowledge of Database, linear algebra, probability & statistics, and algorithms

Objectives/Overview:

- To help students learn, understand, and practice big data analytics and machine learning approaches, which include the study of modern computing big data technologies and scaling up machine learning techniques focusing on industry applications.
- Conceptualization and summarization of big data and machine learning, trivial data versus big data, big data computing technologies, machine learning techniques, and scaling up machine learning approaches using MapReduce and Spark.

Course Outcomes:
At the end of the course, a student should:

Sl. No.	**Outcomes**	**Mapping to POs**
1.	Identify the characteristics of datasets and compare the trivial data and big data for various applications.	PO1, PO2, PO5
2.	Understand the concept and challenge of big data and why existing technology is inadequate to analyze the big data.	PO1, PO2, PO4, PO5
3.	Solve problems associated with batch learning and online learning, and the big data characteristics such as high dimensionality, dynamically growing data and in particular scalability issues.	PO1, PO2, PO4, PO12
4.	Understand the various new technologies like MapReduce and Spark to process Big datasets.	PO2, PO4, PO5, PO6, PO12
5.	Understand and write the various analytics algorithms using MapReduce and Spark.	PO2, PO4, PO5, PO6, PO12

6.	Integrate machine learning libraries and mathematical and statistical tools with modern technologies like Hadoop, MapReduce, and Spark.	PO2, PO4, PO5, PO12

UNIT I: **Lectures: 6**

Understanding Big Data and Hadoop: Introduction to Big Data & Big Data Challenges, Limitations & Solutions of Big Data Architecture, Hadoop & its Features, Hadoop Ecosystem, Hadoop 2.x Core, Components, Hadoop Storage: HDFS (Hadoop Distributed File System), Hadoop Processing: MapReduce Framework, Different Hadoop Distributions.

UNIT II: **Lectures: 12**

Hadoop Architecture and HDFS: Hadoop 2.x Cluster Architecture, Federation and High Availability Architecture, Typical Production Hadoop Cluster, Hadoop Cluster Modes, Common Hadoop Shell Commands, Hadoop 2.x Configuration Files, Single Node Cluster & Multi-Node Cluster set up, Basic Hadoop Administration, MapReduce, Implementation of various analytics algorithms using MapReduce approach.

UNIT III: **Lectures: 12**

Apache Spark: Introduction to Data Analysis with Spark, Programming with RDDs, Working with

Key/Value Pairs, Loading and Saving Your Data, Advanced Spark Programming, running on a Cluster, Tuning and Debugging Spark, Applications and Use cases.

UNIT IV: **Lectures: 12**

Spark SQL, DataFrames, DataSets, Spark Streaming, Graph analysis using Spark, Lambda Architecture in depth, Machine Learning with MLlib, Applications and Use cases.

Text/References Books:

1. Tom White, Hadoop Definite Guide, 3rd Edition, O'Rielly Publication
2. Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, "Learning Spark: Lightning-Fast Big Data Analysis", O'Rielly Publication
3. Nathan Marz, James Warren, "Big Data: Principles and best practices of scalable realtime data systems"
4. Mining Massive Data Sets, A. Rajaraman and J. Ullman, Cambridge University Press, 2012
5. MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems, Donald Miner, Adam Shook, O'Reilly, 2014