

Branch Name:	M.Sc. (Artificial Intelligence & Data Science)
Program Code:	ET214
Course Title:	Big Data Analytics
Course Code:	1ET2140205
Pre-requisite Course:	Database Management Systems, Programming in Python/Java, Basic Machine Learning

Course Objectives:

1. To understand concepts, characteristics, and challenges of big data and modern data warehouse environments.
2. To learn Hadoop ecosystem components and distributed data processing using MapReduce.
3. To explore big data storage technologies such as HDFS, Hive, HBase, and NoSQL databases.
4. To analyze large-scale datasets using Apache Spark and machine learning libraries.
5. To design scalable big data analytics solutions for modern web and enterprise applications.

Teaching and Examination Scheme:

Teaching Scheme (Hours per week)				Evaluation Scheme (Marks)				
Lecture (L)	Tutorial (T)	Practical (P)	Credit	Theory (Marks)		Practical (Marks)		Total (Marks)
				University Assessment	Continuous Assessment	University Assessment	Continuous Assessment	
03	-	02	04	50	50	25	25	150

Course Contents:

Unit No	Topic	Total Hours	Weightage (%)
1	Introduction to Big Data: Types of digital data, structured/semi-structured/unstructured data, definition and challenges of big data, characteristics (Volume, Velocity, Variety), traditional BI vs Big Data, modern data warehouse architecture	08	15%
2	Hadoop and MapReduce Framework: Overview of Apache Hadoop ecosystem, Hadoop architecture, MapReduce paradigm, input-output processing, data serialization and distributed computing workflow	08	15%
3	Big Data Storage Technologies: HDFS architecture and operations, Hive architecture and HiveQL querying, joins and aggregations, MapReduce scripts, HBase concepts and schema design, Pig, Zookeeper cluster coordination	10	20%
4	Apache Spark and Big Data Processing: Introduction to Spark ecosystem, RDD programming, Spark SQL, DataFrame operations, distributed data analysis, machine learning using MLlib	08	15%
5	NoSQL Databases: Overview and need of NoSQL, types (Key-value, Document, Column-family, Graph), SQL vs NoSQL vs NewSQL, industry use cases and performance considerations	08	15%
6	Databases for Modern Web & MongoDB: MongoDB architecture and key features, document-oriented schema design, CRUD operations, indexing, query language, collections and documents, building scalable web data applications	08	15%

Text Books:

1. Tom White – *Hadoop: The Definitive Guide*, O'Reilly Media.
2. Holden Karau, Andy Konwinski, Patrick Wendell – *Learning Spark*, O'Reilly Media.
3. Seema Acharya & Subhasini Chellappan – *Big Data Analytics*, Wiley.

Reference Books:

1. Nathan Marz & James Warren – *Big Data: Principles and Best Practices*, Manning.
2. Bill Franks – *Taming the Big Data Tidal Wave*, Wiley.
3. MongoDB Official Documentation – MongoDB Manual.
4. Kleppmann – *Designing Data-Intensive Applications*, O'Reilly.

LIST OF EXPERIMENTS / PRACTICALS

1. Understanding Hadoop architecture and MapReduce programming workflow.
2. Implementation of basic MapReduce job for word count analysis.
3. Performing CRUD operations in MongoDB using Python/Java driver.
4. Querying and retrieving documents from MongoDB collections.
5. Development of MapReduce-based data processing application.
6. Creating HDFS directories, uploading datasets, and performing file operations.
7. Creating Hive tables, loading data, and executing HiveQL queries with joins and aggregations.
8. Implementation of Pig scripts for big data processing tasks.
9. Programming with Spark RDDs and DataFrame operations.
10. Applying MLlib algorithms for large-scale data analytics.
11. Comparative analysis of SQL vs NoSQL database performance.
12. Mini Project: End-to-end big data analytics pipeline using Hadoop/Spark and MongoDB.

Course Learning Outcomes (CLO):

CO	Statement	Bloom's Level
CO1	Understand big data concepts, characteristics, and distributed processing architecture	L2 – Understand
CO2	Apply Hadoop ecosystem tools and MapReduce for large-scale data processing	L3 – Apply
CO3	Analyze big data storage solutions including HDFS, Hive, HBase, and NoSQL databases	L4 – Analyze
CO4	Evaluate Spark-based distributed data analytics and machine learning workflows	L5 – Evaluate
CO5	Design scalable big data analytics applications for enterprise and web-based systems	L6 – Create

Mapping of CLO with PO and PSO

Course	Program Outcomes (POs)												Program Specific Outcomes (PSOs)			
Learning Outcomes	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3	PSO4
CLO1	H	M	–	–	M	–	–	–	–	L	–	M	M	M	H	L
CLO2	M	H	M	M	H	–	–	–	L	M	–	M	H	H	M	M
CLO3	M	H	H	H	H	–	–	–	L	M	–	M	H	H	H	M
CLO4	M	M	H	H	H	–	–	–	M	H	M	M	H	H	H	M
CLO5	M	H	H	H	H	M	L	M	M	H	H	H	H	H	H	H

H:High, M:Medium, L:Low