

Branch Name:	IMCA
Program Code:	CS301
Course Title:	Big Data Analytics
Course Code:	1CS3010801T
Pre-requisite Course:	Working knowledge of Programming (Java/Python), Database Management Systems (SQL), and basic understanding of Data Structures and Operating Systems.

Course Objective:

The objectives of the course are to:

1. To introduce the fundamental concepts of Big Data, its characteristics, architecture and real-world applications.
2. To understand distributed data storage and processing frameworks, particularly Hadoop and HDFS, for managing large-scale datasets.
3. To learn the MapReduce programming model and Hadoop ecosystem tools such as Hive, Pig and HBase for big data processing.
4. To explore modern big data processing frameworks such as Apache Spark for efficient large-scale data analytics.
5. To understand NoSQL databases and big data technologies used for handling structured and unstructured data in modern applications.

Teaching and Examination Scheme:

Teaching Scheme (Hours per week)				Evaluation Scheme (Marks)				
Lecture	Tutorial	Practical	Credit	Theory (Marks)		Practical (Marks)		Total (Marks)
				University Assessment	Continuous Assessment	University Assessment	Continuous Assessment	
4	-	-	4	60	40	-	-	100

Course Contents:

Unit No	Topic	Total Hours	Weightage (%)
1	Introduction to Big Data: Big Data Overview: concept and evolution , Need for Big Data (limitations of traditional systems) Characteristics of Big Data: 5Vs: Volume, Velocity, Variety, Veracity, Value Types of Data: structured, semi-structured, unstructured Sources of Big Data: social media, IoT, sensors, web logs, transactions Big Data Analytics Lifecycle: data collection, storage, pre-processing, analysis and visualization. Applications of Big Data: healthcare analytics, recommendation systems, fraud detection and social media analytics.	12	25

2	Hadoop Framework and HDFS: Apache Hadoop: open-source framework for distributed storage and processing of large datasets. Hadoop Architecture: master-slave architecture (Basic Concept) Hadoop Distributed File System (HDFS): concept of distributed storage for big data. HDFS Components: NameNode and DataNode roles and Secondary NameNode. Data Storage in HDFS: block storage and replication	12	25
3	Data Processing with MapReduce & Hadoop Ecosystem MapReduce Framework :concept, Map & Reduce functions, role MapReduce Execution Process:basic flow: input → map → shuffle → reduce MapReduce Applications: Basic examples such as Word Count Hadoop Ecosystem Tools: Apache Hive, Apache Pig Comparison with Traditional Databases:Differences between Hadoop and RDBMS, Advantages and limitations of both approaches	12	25
4	Modern Big Data Tools (Spark & NoSQL) Limitations of MapReduce: Issues with disk-based processing, performance limitations and lack of real-time processing Apache Spark: Introduction , Architecture (Driver, Workers, Executors) Spark Core Concepts: Resilient Distributed Datasets (RDD), Spark DataFrames, NoSQL Databases: Introduction to NoSQL, need for flexible schema, Types of NoSQL databases (key-value, document, column-family, graph) MongoDB: Overview of MongoDB, JSON-based data storage and basic CRUD operations	12	25

Text Books:

- Tom White, *Hadoop: The Definitive Guide*, 4th Edition, O'Reilly Media, USA, 2015..
- Raj Kamal and Preeti Saxena, *Big Data Analytics: Introduction to Hadoop, Spark and Machine Learning*, 1st Edition, McGraw Hill Education, 2018.
- Seema Acharya and Subhasini Chellappan, *Big Data Analytics*, Wiley Publications, 2015.
- Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis and Paul Zikopoulos, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill Education, 2012.

Reference Books:

- Bill Chambers and Matei Zaharia, *Spark: The Definitive Guide – Big Data Processing Made Simple*, O'Reilly Media, 2018..
- Holden Karau, Andy Konwinski, Patrick Wendell and Matei Zaharia, *Learning Spark: Lightning-Fast Data Analytics*, 2nd Edition, O'Reilly Media, 2020.
- Anand Rajaraman and Jeffrey David Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2012..

- Michael Minelli, Michele Chambers and Ambiga Dhiraj, *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Wiley Publications, 2013.
- Douglas Eadline, *Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop Ecosystem*, Pearson Education, 2016.

List of Open Source Software/learning website:

- <https://hadoop.apache.org>
- <https://spark.apache.org>
- <https://www.mongodb.com/docs>
- <https://www.kaggle.com>
- <https://www.databricks.com/learn>

Course Learning Outcomes (CLO): On completion of this course, the students will be able to:

CLO	Description	Bloom's Taxonomy Level
CLO1	Understand the fundamental concepts of Big Data, characteristics (5Vs), data sources, architecture and applications in modern data-driven systems.	1. Remembering, 2. Understanding
CLO2	Explain the Hadoop ecosystem, Hadoop architecture and Hadoop Distributed File System (HDFS) for distributed data storage and processing.	1. Remembering, 2. Understanding
CLO3	Apply the MapReduce programming model and Hadoop YARN for large-scale data processing and distributed computing.	3. Applying 4. Analyzing
CLO4	Use Hadoop ecosystem tools such as Hive, Pig and HBase for querying, analyzing and managing big data.	3. Applying 4. Analyzing
CLO5	Analyze large datasets using Apache Spark components such as RDD, DataFrames, Spark SQL and Spark Streaming.	3. Applying 4. Analyzing 6. Creating
CLO6	Evaluate the role of NoSQL databases and big data technologies in real-world applications such as social media analytics, recommendation systems and business intelligence.	3. Applying 5. Evaluating

Mapping of CLOs with POs & PSOs

Course Learning Outcomes	Program Outcomes (POs)												Program Specific Outcomes (PSOs)	
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO 10	PO11	PO12	PSO1	PSO2
CLO1	H	M	-	-	M	-	-	L	-	-	-	M	M	M
CLO2	H	M	M	-	H	-	-	-	-	-	-	M	H	M
CLO3	M	H	H	M	H	-	-	-	-	-	-	M	H	H
CLO4	M	H	H	M	H	-	-	-	-	-	-	M	H	H
CLO5	M	H	H	H	H	-	-	-	-	-	-	M	H	H
CLO6	M	M	H	H	H	-	-	M	M	L	L	H	M	H

H:High, M:Medium, L:Low