

Branch Name:	Artificial Intelligence and Data Science
Program Code	ET108
Semester/Year:	Semester-VII/Fourth Year
Course Name:	Distributed Computing and Big Data Analytics
Course Code:	2ET1080702T
Pre-requisite:	Knowledge of Data Structures, DBMS, Operating Systems, and Programming (Java/Python)

Course Objective:

1. To understand concepts of distributed computing, scalability, and fault tolerance.
2. To introduce distributed file systems and computing frameworks such as Hadoop and Spark.
3. To learn various techniques for Big Data storage, processing, and analytics.
4. To develop practical skills in managing and analyzing large datasets using distributed systems.

Teaching and Examination Scheme:

Teaching and Examination Scheme:													
Credit	Teaching Scheme Hours per week				Evaluation Scheme (Marks)								Total (Marks)
					Theory (Marks)				Practical (Marks)				
	L	T	P	Total	SEE		CCA		SEE		CCA		
					Min.	Max.	Min	Max.	Mn	Max.	Min.	Max.	
3	3	-	-	3	24	60	16	40	-	-	-	-	100

Subject Contents:

Sr. No	Topic	Total Hours	Weightage (%)
1	Introduction to Distributed Computing: Distributed systems: goals, characteristics, and challenges, Models of distributed systems: client-server, peer-to-peer, cluster computing, grid computing, and cloud computing, Communication in distributed systems: message passing, RPC, RMI, Synchronization, coordination, and consistency models.	9	20
2	Distributed File Systems and Resource Management: Concepts of distributed file systems (DFS), Google File System (GFS), Hadoop Distributed File System (HDFS) – architecture and design, Resource management and scheduling in distributed systems, Fault tolerance, replication, and consistency handling.	9	20
3	Big Data Fundamentals: Characteristics of Big Data (Volume, Velocity, Variety, Veracity, Value), Big Data ecosystem overview, Data storage: NoSQL databases (HBase, MongoDB, Cassandra), Data ingestion tools: Sqoop, Flume, Kafka, Data preprocessing and cleaning in distributed environments.	7	15
4	Hadoop Ecosystem and MapReduce: Hadoop architecture and components, HDFS and YARN – detailed understanding, MapReduce programming model: concepts, operations, and examples, Advanced MapReduce patterns: joins, sorting, combiners, partitioners, Introduction to Hive, Pig, and HBase.	7	15

5	Apache Spark and Real-Time Analytics: Introduction to Apache Spark: architecture, RDDs, DataFrames, and Datasets, Spark SQL and Spark Streaming, Machine learning with Spark MLlib, Real-time data analytics using Spark Streaming and Kafka integration, Performance comparison: MapReduce vs Spark.	7	15
6	Case Studies and Applications: Big Data in e-commerce, healthcare, finance, IoT, and social networks, Cloud-based Big Data solutions (AWS EMR, Google BigQuery, Azure HDInsight), Ethical issues and privacy in Big Data, Future trends: edge analytics, federated learning.	6	15

List of References:

1. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, “Professional Hadoop Solutions”, Wiley, 2015.
2. Chris Eaton, Dirk deroos et al., “Understanding Big data”, McGraw Hill, 2012.
3. Tom White, “HADOOP: The definitive Guide”, O Reilly 2012

E-Resources:

1. <http://bigdatauniversity.com/courses/?cat=124>
2. <http://www.coreservlets.com/hadoop-tutorial/>

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

CLO	Description	Bloom's Taxonomy Level
CLO1	Explain the principles and architecture of distributed computing systems	2 Understanding
CLO2	Implement distributed data storage and processing using Hadoop ecosystem.	2 Understanding 3 Applying
CLO3	Apply Big Data analytics tools such as Spark for large-scale data analysis.	3 Applying
CLO4	Demonstrate proficiency in handling streaming and real-time data analytics.	4 Analyze
CLO5	Evaluate performance and fault tolerance aspects in distributed environments.	2 Understanding 3 Applying

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	PO10	PO11	PO12	PSO1	PSO2	PSO3	PSO4
CLO1		M	M	M	M	L						M	H	M	M	
CLO2	L	H	H	M	H					L		M		H	L	
CLO3	L	H	M	M	H	L								M		M
CLO4	M	M	H	M	M			L					H		M	H
CLO5	H	M	H		H				L			M	M		M	H

H: High, M: Medium, L: Low