25UCST416	BIG DATA ANALYTICS	Category	L	T	P	R	Credit	Year of Introduction
		PEC	3	0	0	0	3/5	2025

Preamble: This course introduces the fundamental concepts, architecture, and significance of Big Data in modern computing environments. It focuses on the challenges associated with handling and processing large volumes of data and emphasizes the need for scalable and efficient solutions. The course explores the impact of Big Data in areas such as business intelligence, scientific discovery, and everyday applications. Stu.

Prerequisite: Nil

Course Outcomes: After the completion of the course, the student will be able to:

CO	Course Outcomes	Knowledge Level
CO1	Explain fundamental concepts of Big Data systems.	K2 Understand
CO2	Apply Big Data tools and technologies for large-scale data processing and visualization	K3 Apply
CO3	Illustrate the architecture and working of Hadoop ecosystem components.	K2 Understand
CO4	Apply MapReduce algorithms and implement distributed data processing with resource management using YARN in a Hadoop ecosystem.	K3 Apply

Mapping of course outcomes with program outcomes and program specific outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PSO1	PSO2
CO1	3	1	1	2	1	-	-	-	-	-	1	2	1
CO2	3	2	2	2	3	-	-	-	-	-	1	2	2
CO3	3	2	1	2	2	-	-	-	-	-	1	3	1
CO4	3	3	2	2	3	-	-	-	-	-	1	3	2
Avg	3	2	1.5	2	2.25						1	2.5	1.5

Mapping of course outcomes with Knowledge and Attitude Profile (WK)

	WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9
CO1	*	*	*	*	*	*		*	
CO2	*	*	*	*	*	*		*	
CO3	*	*	*	*	*	*		*	
CO4	*	*	*	*	*	*		*	

WKs Mapped with the course

WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9
*	*	*	*	*	*		*	

Mark Distribution:

Total Marks	CIE Marks	ESE Marks	ESE Duration
100	40	60	2.5 hrs

Continuous Internal Evaluation Pattern:

Attendance : 5 marks
Continuous Assessment Test : 20 marks
Continuous Assessment Assignment : 15 marks

Criteria for Evaluation(Evaluate and Analyse): 20 marks Students must be asked to identify problems involving large datasets and identify the right solution from the concepts already learned. A comparison of the results with a similar approach also need to be performed to assess the Knowledge Level 5.

SYLLABUS

Module 1 (12 hrs)

Introduction to Big Data - Issues and Challenges in the traditional systems - Evolution of Big Data - Four V's of Big Data - Big Data Use Cases and characteristics - Intelligent Data Analysis - Data Analytic Tools - Big Data StorageStatistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error - Random Sampling .

Module 2 (10 hrs)

Big Data Applications using Pig and Hive - Fundamentals of HBase and ZooKeeper - IBM Infosphere Big Insights - Introduction to FLUME - KAFKA , Introduction to NoSQL - MongoDB - Spark - Cassandra - Cassandra Data Model - Data Design - Cassandra Architecture - Read and Write Data - Clients - Integrate with Hadoop Introduction - Importance of Effective Data Visualization - Introduction to Tableau - Choosing the Right Chart Type - Using the Color Effectively Reducing Clutter - Dashboard Creation and Formatting

Module 3 (11 hrs)

Introduction to HADOOP - Hadoop ecosystem components and uses, Hadoop Storage: HDFS, Concept of Hadoop Distributed file system, Design of HDFS, Configuration of HDFS. Hadoop Data Types, large-scale deep belief nets with mapreduce - Functional-Concept of Mappers, Functional-Concept of Reducers, MapReduce Execution Framework, Partitioners and Combiners, Hadoop Clusters component : NameNode, Secondary NameNode, and DataNode, Data flow (Anatomy of File Write and Read)

Module 4 (9 hrs)

Algorithms using map reduce - Matrix-Vector - Multiplication - Word Count - Understanding inputs and outputs of MapReduce, Data Serialization - Introduction to YARN - MapReduce Vs YARN - YARN Architecture - Scheduling in YARN - Fair Scheduler - Capacity Schedule

Completion of an industrial visit or industrial training is desirable for this course.

Text Books:

1. White, T. (2015). *Hadoop: The definitive guide* (4th ed.). O'Reilly Media.
2. Lam, C. (2018). *Hadoop in action*. Manning Publications.
3. Milligan, J. N. (2015). *Learning Tableau*. Packt Publishing.

Reference Books:

1. Hewitt, E. (2010). *Cassandra: The definitive guide*. O'Reilly Media.
2. Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets* (1st ed.). Cambridge University Press.
3. Lin, J., & Dyer, C. (2010). *Data-intensive text processing with MapReduce*. Morgan & Claypool Publishers.
4. Owens, J. R., Femiano, B., & Lentz, J. (2013). *Hadoop real world solutions cookbook*. Packt Publishing.

END SEMESTER EXAMINATION – MODEL QUESTION PAPER

Course Code	**25UCST416**
Course Name	**BIG DATA ANALYTICS**
Max. Marks: 60	**Duration: 2.5 Hours**

Part A *(Answer all questions. Each question carries 3 marks)*				
Qn No.	**Question**	**Marks**	**CO**	**BL**
1	Explain the Four V's of Big Data with relevant examples and discuss the challenges that traditional data systems face when dealing with Big Data.	3	1	L2
2	Describe sampling distributions and explain how re-sampling techniques are used in Big Data statistical inference.	3	1	L2
3	Apply HBase and ZooKeeper to design a distributed storage solution for a real-time e-commerce application. Justify your component choices.	3	2	L3
4	A healthcare organization needs to store millions of patient records with varying attributes, support real-time appointments booking, and run analytical queries on treatment history. Apply your knowledge of MongoDB, Cassandra, and HBase to select the most suitable NoSQL database for each requirement and justify your choice.	3	2	L3
5	Explain the HDFS architecture, including the roles of NameNode, Secondary NameNode, and DataNode, and describe the data flow during a file write operation.	3	3	L2
6	Describe the Hadoop ecosystem components and their uses. Explain how HDFS design enables fault tolerance in a distributed environment.	3	3	L2
7	Apply the MapReduce programming model to design a Word Count solution. Explain the role of Mapper, Reducer, Partitioner, and Combiner in the execution pipeline.	3	4	L3
8	Explain the Fair Scheduler and Capacity Scheduler in YARN with their working principles.	3	4	L2
Part B (*Answer any one full question from each Module. Each carry 9 marks*)				
Module I				
Qn No.	**Question**	**Marks**	**CO**	**BL**
9	Explain the evolution of Big Data from traditional data systems, covering the Four V's, Big Data use cases, and characteristics. Describe the role of Intelligent Data Analysis and the key data analytic tools used in the Big Data ecosystem.	9	1	L2

OR				
10	Describe the statistical concepts used in Big Data analytics including sampling distributions, re-sampling techniques, statistical inference, and prediction error. Explain how random sampling supports large-scale data analysis.	9	1	L2
Module II				
Qn No.	**Question**	**Marks**	**CO**	**BL**
11	Explain Big Data applications using Pig and Hive with suitable examples. Describe the fundamentals of HBase and ZooKeeper, and explain the role of FLUME and KAFKA in data ingestion pipelines.	9	2	L2
OR				
12	Describe the Cassandra data model, architecture, data design, read/write operations, its integration with Hadoop, and data visualization using Tableau with dashboard creation.	9	2	L2
Module III				
Qn No.	**Question**	**Marks**	**CO**	**BL**
13. a	Describe the HDFS architecture by explaining the roles of NameNode and DataNode, the replication strategy for fault tolerance, and the anatomy of file read and write operations.	5	3	L2
13. b	Explain the Hadoop cluster components and describe the Hadoop Data Types. How does the concept of large-scale deep belief nets integrate with MapReduce.	4	3	L2
OR				
14. a	Explain the HDFS design principles including cluster configuration, replication factors, and how data locality optimizes processing performance in a Hadoop environment.	5	3	L3
14. b	Explain the functional concepts of Mappers and Reducers in the MapReduce execution framework. Describe the role of Partitioners and Combiners with suitable examples.	4	3	L2
Module IV				
Qn No.	**Question**	**Marks**	**CO**	**BL**
15. a	Design a MapReduce solution for Matrix-Vector Multiplication: define the input/output formats, describe the Mapper and Reducer logic, and explain data serialization considerations for large matrices.	6	4	L3
15. b	Explain the YARN architecture, its components, and how it manages resource scheduling. Compare the Fair Scheduler and Capacity Scheduler in terms of resource allocation policies.	3	4	L2
OR				
16. a	A retail company needs to analyze transaction data across	6	4	

	10 million records daily using MapReduce on a YARN cluster. Apply MapReduce algorithms and YARN scheduling to design the processing pipeline, optimize resource allocation, and handle data serialization.			L3
16. b	Explain the introduction to YARN and compare MapReduce vs YARN in terms of job execution model, resource management, and scalability.	3	4	L2

******