Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – December 2025

Date: 15-12-2025 **Time:** 1.00pm to 04.00pm

School: SOCSE	**Program:** B.Tech (CAI)	
Course Code: CAI3401	**Course Name:** Big Data Analytics for AI	
Semester: V	**Max Marks**: 100	**Weightage**: 50%

CO - Levels	CO1	CO2	CO3	CO4	CO5
Marks	26	24	24	26	NA

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	What are the key characteristics of Big Data	**2 Marks**	**L2**	**CO1**
2.	Describe big data analytics	**2 Marks**	**L2**	**CO1**
3.	Define “Volume” in Big Data.	**2 Marks**	**L1**	**CO1**
4.	Describe key applications of NoSQL	**2 Marks**	**L2**	**CO2**
5.	What is the role of a DataNode in HDFS?	**2 Marks**	**L2**	**CO2**
6.	What is data preprocessing?	**2 Marks**	**L2**	**CO3**
7.	What is Apache NiFi used for?	**2 Marks**	**L2**	**CO3**
8.	What is RDD lineage?	**2 Marks**	**L2**	**CO4**
9.	Give one example of a DataFrame transformation.	**2 Marks**	**L2**	**CO4**
10.	What is a Pipeline in Spark MLlib?	**2 Marks**	**L2**	**CO4**

Part B

Answer the Questions. **Total Marks 80M**

11.	**a.**	Discuss the Data Processing Lambda Architecture	**10 Marks**	**L3**	**CO1**
	b.	Describe key components of the big data ecosystem with examples	**10 Marks**	**L2**	**CO1**
		Or			
12.	**a.**	Define Big Data. Explain the need for Big Data analytics in modern applications.	**10 Marks**	**L2**	**CO1**
	b.	What are the major modules in Apache Spark ecosystem?	**10 Marks**	**L2**	**CO1**

13.	**a.**	Compare MongoDB with traditional RDBMS. Why is MongoDB considered better for Big Data applications	**10 Marks**	**L3**	**CO2**
	b.	Describe the Hadoop Distributed File System used for Big Data	**10 Marks**	**L2**	**CO2**
		Or			
14.	**a.**	Explain how Cassandra handles read and write operations.	**10 Marks**	**L2**	**CO2**
	b.	What is replication in Cassandra? Describe replication strategies	**10 Marks**	**L2**	**CO2**

15.	**a.**	Describe the data preprocessing steps in a data pipeline	**10 Marks**	**L2**	**CO3**
	b.	Compare Apache NiFi, Sqoop, and Flume.	**10 Marks**	**L2**	**CO3**
		Or			
16.	**a.**	Discuss the importance and advantages of ETL in big data systems.	**10 Marks**	**L3**	**CO3**
	b.	What is data cleaning? Explain common techniques used.	**10 Marks**	**L2**	**CO3**

17.	**a.**	Describe the lineage and fault-tolerance mechanism in RDD	**10 Marks**	**L2**	**CO4**
	b.	Discuss the working of Logistic Regression in Spark MLlib.	**10 Marks**	**L3**	**CO4**
		Or			
18.	**a.**	Define DataFrame. Explain its advantages over RDD.	**10 Marks**	**L2**	**CO4**
	b.	Explain Dataset operations with an example	**10 Marks**	**L2**	**CO4**