Name: **Enrolment No:**	UPES UNIVERSITY OF TOMORROW

UPES

End Semester Examination, December 2024.

Course: Big Data and Large Scale Computing **Semester: III**
Program: MCA **Time : 03 hrs.**
Course Code: CSDA8003P **Max. Marks: 100**

Instructions: 1. Attempt all questions
2. For each question, provide concise, clear, and well-structured responses that address all parts of the question.
3. Where applicable, support your answers with relevant diagrams, examples, or case studies.

SECTION A
(5Qx4M=20Marks)

S. No.		Marks	CO
Q 1	Explain the characteristics of Big Data (Volume, Velocity, Variety, Value) and how they influence data management.	**4**	**CO1**
Q 2	Describe the concept of a Data Lake and list the different types of data sources that can be ingested into it.	**4**	**CO1**
Q 3	Outline the architecture of Hadoop and the role of HDFS in data storage.	**4**	**CO2**
Q 4	What are Resilient Distributed Datasets (RDDs) in Apache Spark? Discuss their significance in distributed computing.	**4**	**CO3**
Q 5	Discuss how distributed machine learning principles impact computation and data storage for large-scale datasets.	**4**	**CO2**

SECTION B
(4Qx10M= 40 Marks)

Q 6	Compare and contrast horizontal scaling and vertical scaling in terms of scalability and processing efficiency in big data systems.	**10**	**CO1**
Q 7	Explain the Hadoop ecosystem components Pig, Hive, and HBase, and their roles in big data processing.	**10**	**CO3**
Q 8	Describe the process of implementing a linear regression model using distributed machine learning principles. Provide relevant examples. OR Explain the significance of logistic regression and how it is applied in online advertising and click-through rate prediction.	**10**	**CO4**
Q 9	What are the storage and processing requirements of a Data Lake? How does it ensure scalability?	**10**	**CO2**

SECTION-C
(2Qx20M=40 Marks)

Q 10	Analyze the role of MapReduce in scalable data processing. Illustrate with an example how it works with distributed datasets.	**20**	**CO2, CO3**

	OR Explain the historical evolution and key architectural principles of Apache Spark. How does it differ from traditional Hadoop MapReduce?		
Q 11	Describe the process of performing Principal Component Analysis (PCA) on large datasets. What are its applications in neuroimaging and data dimensionality reduction? OR Explain the role of MLlib in Spark for building k-means clustering models. Discuss its significance in big data analytics.		**CO4**