Roll No												

PRESIDENCY UNIVERSITY
BENGALURU

School Of Computer Science and Engineering & Information Science

End-Term Examinations, Aug 2024

Odd Semester: 2023 - 24
Course Code: CSE3002
Course Name: Big Data Technologies
Department: SOCSE/IS

Date: 06/08/24
Time: 1:00 pm – 4:00 pm
Max Marks: 100
Weightage: 50%

Instructions:

(i) Read the all questions carefully and answer accordingly.
(ii) Do not write any matter on the question paper other than roll number.

Q. No	Questions	Marks	CO	RBT
1	a. What are the "Four Vs" of Big Data and how do they contribute to its importance?	4	CO 1	L1
	b. What are the primary drivers for the adoption of Big Data technologies in modern applications?	6	CO 1	L2
	c. What are the main types of No-SQL databases and how do they differ from relational databases?	10	CO 1	L3
	OR			
2	a. Identify and explain three major challenges associated with Big Data.	4	CO 1	L1
	b. What are the key components of the Hadoop ecosystem and their functions?	6	CO 1	L2
	c. Discuss the process of file writing and reading in HDFS.	10	CO 1	L3
3	a. What are the key features introduced in Hadoop 2.0, and how do they improve upon the original Hadoop architecture?	4	CO2	L1
	b. What are the key features of Sqoop?	6	CO2	L2
	c. Discuss the steps and options for exporting data from HDFS to a relational database using Sqoop.	10	CO2	L3
	OR			
4	a. Explain Hive table partitioning and its benefits for query performance.	4	CO 2	L1
	b. List and describe the common Hive DDL (Data Definition Language) commands.	6	CO 2	L2
	c. What is Hive bucketing, and how does it affect data organization and query performance?	10	CO 2	L3

5	a. Describe the architecture of HBase and its key components.	4	CO 3	L 1
	b. Explain the commands for disabling and enabling tables in HBase.	6	CO 3	L 2
	c. Describe the commands for scanning, counting, and truncating tables in HBase.	10	CO 3	L 3

OR

6	a. What is Apache Spark, and how does it unify different data processing tasks?	4	CO 3	L 1
	b. What is an RDD (Resilient Distributed Dataset) in Apache Spark, and what are its key characteristics?	6	CO 3	L 2
	c. Describe the process of creating and manipulating RDDs in Spark.	10	CO 3	L 3

7	a. Evaluate the use of Hive for real-time data processing versus batch processing. What are the limitations of Hive in handling real-time data, and how can they be addressed?	4	CO 4	L 1
	b. Discuss the impact of Hive's table partitioning and bucketing strategies on query performance. In what scenarios are these features most beneficial?	6	CO 4	L 2
	c. How does HBase handle scalability and high availability, and what are the trade-offs involved in its architecture?	10	CO 4	L 3

OR

8	a. List the commands used for creating and listing tables in HBase.	4	CO 4	L 1
	b. How do you describe and drop tables in HBase?	6	CO 4	L 2
	c. Describe the commands for scanning, counting, and truncating tables in HBase.	10	CO 4	L 3

9	a. How do data security and privacy concerns impact Big Data implementations?	4	CO 1	L 1
	b. Explain the architecture of HDFS and the roles of NameNode and DataNode.	6	CO 1	L 2
	c. How do Combiners and Partitioners improve the efficiency of MapReduce jobs?	10	CO 1	L 3

OR

10	a. Explain the difference between structured, unstructured, semi-structured, and quasi-structured data.	4	CO 2	L 1
	b. Compare and contrast traditional data management approaches with Big Data approaches	6	CO 2	L 2
	c. Discuss the advantages and limitations of No-SQL databases in handling Big Data.	10	CO 2	L 3