U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

January / February 2025 Semester End Main Examinations

Programme: B.E. **Semester: VI**

Branch: Artificial Intelligence and Machine Learning **Duration: 3 hrs.**

Course Code: 24AM6PCBDA **Max Marks: 100**

Course: Big Data Analytics

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

			UNIT - I	*CO*	*PO*	Marks
1	a)		Explain why web data is considered the most popular form of Big Data, providing relevant examples to support the answer.	*CO1*	*PO2*	**5**
	b)		Describe the role of Big Data Analytics in the development and enhancement of smart cities.	*CO1*	*PO2*	**5**
	c)		George is a data analyst working for a global retail chain that operates both physical stores and an online platform. The company is looking to enhance its sales strategies and customer experiences by harnessing the power of big data analytics. i. Recommend suitable types of Big Data Analytics. ii. Identify specific tools and techniques associated with each type of analytics that George would propose for implementation.	*CO1*	*PO3*	**10**
			OR			
2	a)		Define Big Data Analytics. Explain the various sources of big data.	*CO1*	*PO1*	**8**
	b)		Describe the various characteristics of big data.	*CO1*	*PO1*	**8**
	c)		List and explain the merits and demerits of big data.	*CO1*	*PO1*	**4**
			UNIT - II			
3	a)		Peter is a data architect responsible for designing a comprehensive data storage and analytics solution for a multinational E-commerce corporation. The company operates in multiple regions and deals with a vast amount of data, including customer transactions, inventory updates, and website clickstream events. i. Suggest some of the practical uses Peter could implement for Orc, Parquet, and Avro file formats in his work. ii. Elaborate the unique features and advantages of each format.	*CO1*	*PO3*	**10**

		b)	Explain the different file compression techniques used for managing large files.	*CO1*	*PO2*	**4**
		c)	Outline the differences between row-based and column-based file formats.	*CO1*	*PO2*	**6**
			OR			
	4	a)	Differentiate between data and file format with suitable examples.	*CO1*	*PO1*	**6**
		b)	Distinguish between Lossless and Lossy data compression.	*CO1*	*PO1*	**7**
		c)	Elaborate on Columnar and Row Columnar file format with relevant examples.	*CO1*	*PO1*	**7**
			UNIT - III			
	5	a)	Is block replication necessary in data nodes for storing data in distributed systems like Hadoop? Support your explanation with an appropriate diagram.	*CO2*	*PO2*	**5**
		b)	A company needs to process large log files from a web server to generate a report on the most frequently accessed pages. i. Using the MapReduce framework in Hadoop, explain the data flow and the different phases involved in this process. ii. How does each phase contribute to transforming and aggregating the data to achieve the desired result?	*CO2*	*PO3*	**10**
		c)	Explain the components of YARN architecture and describe their roles in resource management and task scheduling within a Hadoop ecosystem.	*CO2*	*PO2*	**5**
			OR			
	6	a)	A Data Engineer at a financial services company needs to transfer large sets of transaction data between different stages of the Hadoop data pipeline efficiently. Elucidate how he would use data serialization and deserialization techniques to achieve this?	*CO2*	*PO3*	**10**
		b)	Describe the various modules in the Hadoop ecosystem.	*C02*	*PO1*	**5**
		c)	Illustrate the working of the Hadoop Distributed File System (HDFS) with a neat sketch.	*C02*	*PO2*	**5**
			UNIT - IV			
	7	a)	An E-commerce company deals with extensive customer transaction data, and uses Apache Hive to manage and analyze these transactions. The data is stored in the Hadoop Distributed File System. Draw a diagram illustrating the Hive architecture and its integration with Hadoop components for managing and analyzing transaction data. Write the query for the following tasks: i. Create a Hive table to store the customer transactions, assuming the transactions are in a structured format with fields like	*C03*	*PO3*	**10**

			transaction_id, customer_id, product_id, quantity, price, and transaction_date. ii. Compute the total spend for each customer_id and list the top 5 customers who have spent the most.			
		b)	Distinguish between HBase and Relational Database Management System (RDBMS).	*CO3*	*PO1*	**5**
		c)	Given an employee database in Hive with the following schema: CREATE TABLE employees (employee_id INT, employee_name STRING, department STRING, salary DOUBLE, hire_date DATE); Write the query for the following: i. List all employees with their details, sorted by their 'hire_date' in ascending order. ii. Calculate the total salary paid to employees in each department. Include only departments where the total salary exceeds $100,000.	*CO3*	*PO2*	**5**
			OR			
	8	a)	Given a banking database in HIVE containing a table named transactions with the following schema: transaction_id INT, account_number STRING, transaction_date DATE, transaction_type STRING, amount DECIMAL (10,2) Write a HIVE query to: i. Group the transactions by account_number and calculate the total transaction amount for each account. ii. Sort the results by the total transaction amount in descending order.	*CO3*	*PO2*	**6**
		b)	Consider two tables in a HIVE database for an employee management system: employees table: employee_id INT, name STRING, department_id INT, salary DECIMAL (10,2) departments table: department_id INT, department_name STRING Write a HIVE query to: 1. Join the employees table with the departments table on department_id. 2. Calculate the average salary for each department. 3. Display the department_name and the calculated average_salary.	*CO3*	*PO2*	**8**
		c)	Describe the various features of HIVE.	*CO3*	*PO1*	**6**

			UNIT - V			
	9	a)	Analyze the importance of using Zookeeper in managing and synchronizing distributed data processing tasks within a Hadoop cluster. Highlight its role in facilitating job scheduling and resource allocation, and support your analysis with suitable architecture.	*CO3*	*PO2*	**10**
		b)	Explain how Apache Hive transforms raw data into meaningful insights and outline the key features of Hive in the Hadoop Ecosystem.	*CO3*	*PO2*	**10**
			OR			
	10	a)	Describe the main features of Apache Pig that enhance the efficiency and flexibility of data processing with suitable example.	*CO3*	*PO2*	**10**
		b)	Write a Spark program that takes a file containing English sentences as input and outputs distinct words and their frequency of occurrence in the file.	*CO3*	*PO2*	**10**

****