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: V**
Branch: AI & DS / CSE(DS) **Duration: 3 hrs.**
Course Code: 23DS5PCBDA **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)	Consider the E-commerce application such as Myntra. Highlight the characteristics of Big Data.	CO2	PO2	**06**
	b)	With a neat diagram, explain the Berkely Data Analytics stack.	CO1	PO1	**08**
	c)	Inspect the application domain and apply: - i. Phases of analytics for fraud detection in the financial sector. ii. Characteristics of big data analytics with a healthcare management system.	CO2	PO2	**06**
		OR			
2	a)	Discuss the evolution of big data. Identify the types for data for social media.	CO2	PO2	**10**
	b)	Design a Berkeley Data Analytics Stack (BDAS) for a healthcare system that needs to handle both real-time patient data and historical records using the Big Data stack.	CO1	PO1	**10**
		UNIT - II			
3	a)	Consider a distributed database used by a large bank to maintain customer balances. The database is replicated across multiple data centers. If a network partition occurs between two data centers, how would the CAP theorem influence the decision to either maintain consistency or availability of transactions during the partition?	CO2	PO2	**10**

		b)	Write MongoDB commands for the following: i. Create a collection employee with fields EMP_ID, NAME, DEPARTMENT, SALARY, and GRADE, and set EMP_ID as the primary key. ii. Insert 3 documents at once into the employee collection with fields EMP_ID, NAME, DEPARTMENT, SALARY, and GRADE. iii. Find a document where EMP_ID = 101. iv. Retrieve employees who are in the DEPARTMENT = 'HR' or have EMP_ID = 102. v. Group employees based on the GRADE field and count the number of employees in each grade	*CO3*	*PO3*	**10**
			OR			
	4	a)	Write CQL commands for the following: i. Create a keyspace named RetailDB with a replication factor of 3. ii. Create a table `OrderInfo` in the keyspace `retail` with a primary key field `OrderID`. iii. Display the details of `OrderInfo`. Add a column `OrderDate` to the `OrderInfo` table. iv. Delete a row from the `OrderInfo` table where `OrderID = 101`. v. Insert a row into the `OrderInfo` table with a TTL (Time-To-Live) of 300 seconds. vi. Add a column product_name to table `OrderInfo`.	*CO3*	*PO3*	**12**
		b)	Identify the type of consistency level for the following i. A write must be written to commitlog and memtable of at least one replica node. ii. A write must be written for at least one replica node in the local data center. iii. A write must be written to commitlog and memtable on quorum of replica nodes in all data centers. iv. A write must be written to commitlog and memtable on quorum of replica nodes in the same center.	*CO2*	*PO2*	**08**
			UNIT - III			
	5	a)	A company generates 500 MB of sensor data every hour, which needs to be stored and processed for analytics. Explain the roles of NameNode and DataNode in storing this data using HDFS. Additionally, describe the shuffle and sort phases in MapReduce and their importance in processing large datasets	*CO2*	*PO2*	**10**
		b)	Identify and explain the following ecosystem tools- i. Build-in Machine learning algorithm ii. Monitors workflow scheduling iii. Co-ordinates distributed applications	*CO2*	*PO2*	**10**

			iv. Distributed service aggregating large amount of log data v. Transfer data between Hadoop and External data sources.			
			OR			
	6	a)	Explain the anatomy of a MapReduce job execution. Illustrate how the map phase and reduce phase can be used to count the frequency of characters (excluding spaces) in the following input data: 'Big Data', 'Hello Data', 'Welcome Data'. Also, describe how the spill buffer manages intermediate data when memory overflow occurs during the map phase.	*CO3*	*PO3*	**10**
		b)	Illustrate YARN based execution model and its functions with a neat diagram.	*CO2*	*PO2*	**10**
			UNIT - IV			
	7	a)	With the help of a diagram, Write the anatomy of a file write operation by HDFS client. Highlight packet write and acknowledgements.	*CO1*	*PO1*	**10**
		b)	Construct a Directed Acyclic Graph (DAG) workflow in Oozie for executing a word count program. What is the role of each component in the DAG.	CO3	PO3	**10**
			OR			
	8	a)	With the help of a diagram, explain the anatomy of a file read in in HDFS, highlighting the interaction between the client, NameNode, and DataNodes	*CO3*	*PO3*	**10**
		b)	Discuss how Flume is used to collect and transfer large volumes of log data into HDFS. Describe its key components such as source, channel, and sink.	*CO1*	*PO1*	**10**
			UNIT - V			
	9	a)	Outline the steps to perform sum of odd numbers over RDDs using structured API execution.	*CO1*	*PO1*	**10**
		b)	A CSV dataset contains Branch, Student_name, USN, I_SEM_percentage,II_SEM_percentage Write the Spark queries to perform the following operations i. Create a dataframe by reading the contents from a CSV file in a local directory ii. Add a new column SEM_percentage to calculate the aggregate the I_SEM_percentage,II_SEM_percentage iii. Rename the Branch as Branch_code iv. Use filter to extract the students I_SEM_percentage greater than 60 v. Remove I_SEM_percentage	*CO3*	*PO3*	**10**
			OR			

	10	a)	Write a Spark SQL query for the following: i. Boolean operation to return a) True b) false ii. Date and Timestamp to return addition of 5 days and subtraction of 5 days iii. Person correlation co-efficient iv. Read a JSON file.	*CO3*	*PO3*	**10**
		b)	Write the Spark SQL query the given a raw CSV file containing sales data for a retail store. The file, sales_data.csv, is located at /data/sales_data.csv and contains the following columns: invoice_id,product_name,quantity,unit_price,customer_age,city 1001,Smartphone,2,600,25,Bengaluru 1002,Laptop,1,1200,30,Chennai 1003,Headphones,5,50,22,Hyderabad 1004,Smartwatch,3,200,27,Bengaluru 1005,Laptop,2,1100,35,Chennai 1. Create a DataFrame from the provided CSV file. 2. Add a new column called total_value by multiplying the quantity and unit_price for each sale. 3. Filter the DataFrame to show only sales where the total_sale_value is greater than 1000. 4. Use selectExpr to create a new column discounted_price, where the price is reduced by 15% for all products in the DataFrame.	*CO3*	*PO3*	**10**

******