RV College of Engineering®

Go, Change the World

Academic Year 2023-24 (EVEN Semester)

USN | 1 | R | V | 2 | 1 | A | I | 0 | 3 | 2 |

Department of Artificial Intelligence and Machine Learning

Course Code:	:	**21AI62**	**Date**	:	**22 .07.2024**
Semester	:	**VI**	**Time**	:	**02:00 – 04:00pm**
Max Marks	:	**50 Marks**	**Duration**	:	**120 mins**

Big Data Technologies

CIE -2

Note: Answer all the Questions

SL. No		Questions	M	BT	CO
		PART - A			
1		The Map task is created for each split, the number of splits can be determined by _____ package	01	2	2
2		A map reduce job is said to be a small job, if it has _____ mappers and _____ reducers	01	2	1
3		The _____ memory and _____ virtual core is allocated for each map and reduce task	01	1	2
4		The blacklisting of the node manager is done by _____	01	2	2
5		The Black listing of Node manager is done if more than _____ tasks fails.	01	1	1
6		Which configuration parameter controls the maximum number of MapReduce tasks in Hive?	01	2	2
7		What is the purpose of the ANALYZE command in Hive?	01	1	2
8		The manageability and security of the database which is stored in hive is best achieved by _____	01	1	1
9		The port number which is used when Metastore run as a Hive service is _____	01	2	1
10		The format of the table which is stored in Hive is _____	01	1	1
		PART - B			
1	a	Discuss the Job Submission and Job initialization steps in Anatomy of a Map Reduce Job Run with a neat diagram?	06	2	1
	b	Explain Progress and Status Updates when a map reduce job is submitted to Hadoop cluster?	04	2	1
2	a	Discuss how Task Failure and Application Master Failure are handled in Hadoop Environment?	06	2	1
	b	The map function starts producing output, it is not simply written to disk. Justify the given statement	04	4	1

3	a	Considering the scenario. A company has many employees, where each employee's id, name, job name, manager_id, hire_date, salary, commission is recorded. All employees' works in department. The department details department id, department name and department location is stored. The salary information of all the employee are also stored through grade. The minimum and maximum salary of the employee are also recorded. 1. Write the relevant table information in HiveQL 2. Load the data into the tables 3. Write the following queries using HiveQL and also write the **sample output** (a) Write a Hive query to find the employee ID, salary, and commission of all the employees. (b) Write query to find those employees who joined before 1991. Return complete information about the employees. (c) Write a query to find the managers. Return complete information about the managers (d) Write query to find the employees of grade 2 and 3.Return all the information of employees and salary details. (e) Write a query to compute the experience of all the managers. Return employee ID, employee name, job name, joining date, and experience. (f) Write a SQL query to find those employees who are senior to ADELYN. Return complete information about the employees.	10	4	2
4	a	A data engineer at a retail company uses Hive for data warehousing. The company wants to analyze sales data to identify trends and make strategic decisions. The sales data is stored in a Hive table called sales_data. This table contains the following columns: transaction_id , transaction_date , customer_id , product_id , quantity ,price The company has recently expanded its product line and introduced a new category of products. As a result, they want to partition the sales_data table by the product_category column. This column will have values such as 'Electronics', 'Clothing', 'Home', etc. The data engineer task is to create a new partitioned table in Hive and migrating the existing data from sales_data to the new table. Give the answers to the following questions? 1. How would she/he creates a new partitioned table sales_data_partitioned in Hive with the same structure as the original sales_data table but partitioned by the product_category column? 2. Write the HiveQL query to insert the existing data from the sales_data table into the sales_data_partitioned table. Assume that the product_category column can be derived from the product_id using a lookup table called product_lookup that contains product_id and product_category. 3. After the data migration, how would you verify that the data has been correctly partitioned in the new sales_data_partitioned table? 4. Describe the benefits of partitioning the sales_data table by product_category for query performance and data management. 5. What potential challenges might you encounter when partitioning a large existing table, and how can you mitigate these challenges?	10	4	2
5	a	Differentiate between Inserts and Multitable Insert in Hive with an example?	04	1	2

	b	You are working with two tables in Hive: **customers and orders**. The customers table contains information about your company's customers, and the orders table contains information about the orders placed by these customers. The structure of the tables is as follows: **Table: customers** • customer_id , customer_name ,customer_email . **Table: orders** • order_id , customer_id , order_date, order_amount , Write Hive queries using Joins 1. Write a HiveQL query to retrieve a list of customers along with their orders. Include the customer_id, customer_name, order_id, and order_date in the result using Joins 2. Write a HiveQL query to retrieve a list of all customers along with their orders. Include all customers even if they have not placed any orders. Include the customer_id, customer_name, order_id, and order_date in the result. 3. Write a HiveQL query to retrieve a list of all orders along with customer information. Include all orders even if the customer information is missing. Include the order_id, order_date, customer_id, and customer_name in the result.	06	4	2

M-Marks, BT-Blooms Taxonomy Levels, CO-Course Outcomes

	Particulars	CO1	CO2	CO3	CO4	CO5	L1	L2	L3	L4	L5	L6
Marks Distribution	Max Marks CIE	25	35	--	--	--	09	21	30	--	--	--

Course Outcomes: After completing the course, the students will be able to:-	
CO1	Understand and apply the different building blocks of Big Data Technologies to a given problem
CO2	Articulate the programming aspect of Big Data Technologies to obtain solution to the problem through lifelong learning
CO3	Exhibit effective communication to represent the analytical aspects of Big Data Technologies for obtaining solution to the problems
CO4	Demonstrate solutions for societal and environmental concern problems using modern engineering tools through writing effective reports
CO5	Appraise the knowledge of Big Data Technologies as an Individual /as a team member