Tatyasaheb Kore Institute of Engineering and Technology, Warananagar

(An Autonomous Institute, Affiliated to Shivaji University, Kolhapur)

Department of Computer Science and Engineering

Problem Statements for Big Data Analytics (POE)
Academic Year: 2024-25

Problem Statements on Hive DDL

1. Write the Hive Commands for:

   i. Create a database named Employee_DB
   ii. Create a table employee with columns (empid, ename, designation, department) partitioned by department
   iii. Show the structure of employee table
   iv. Modify the name of column ename to emp_name
   v. Add the column salary
   vi. Rename table to employee_new
   vii. Delete a table employee_new

2. Write the Hive Commands for:

   i. Create a database named Student_DB
   ii. Create a table student with columns (class, division, roll_no, sname) partitioned by class
   iii. Show the structure of student table
   iv. Modify the name of column sname to student_name
   v. Add the column percentage
   vi. Rename table to student_new
   vii. Delete a table student_new

Problem Statements on Hive DML

3. Write the Hive queries for the following:

   i. Create a database named Employee_DB
   i. Create a table employee with columns (empid, ename, designation, department, salary)
   ii. Insert the five records into employee table.
   iii. Update the salary of employee to 40000 whose empid is 101
   iv. Delete the record of employee whose empid is 101

4. Write the Hive queries for the following:

   i. Create a database named Student_DB
   ii. Create a table student with columns (class, division, roll_no, sname, percentage) partitioned by class
   v. Insert the five records into student table.
   vi. Update the percentage of student to 80 whose roll_no is 4
   vii. Delete the record of student whose roll_no is 4

Problem Statements on Hive Data Retrieval

5. Write the Hive queries for the following:
   i. Create a database named Employee_DB
   ii. Create a table employee with columns (empid, ename, designation, department, salary)
   iii. Insert the ten records into employee table.
   iv. Retrieve the employee records that have salary greater than 20000 and from “sales” department
   v. Calculate the average salary obtained by employees from all department. Result should display the department and average salary.
   vi. Retrieve the name of department having average salary greater than 20000.

6. Write the Hive queries for the following:
   i. Create a database named Student_DB
   ii. Create a table student with columns (class, division, roll_no, sname, percentage) partitioned by class
   iii. Insert the ten records into student table.
   iv. Retrieve the student records that have percentage greater than 60 and from “TY” class
   v. Calculate the average percentage obtained by students from “TY” Class. Result should display the class and average percentage.
   vi. Retrieve the name of class having average percentage greater than 60.

Problem Statements on Pig Operators

7. Write the Pig scripts for the following:
   i. Create a database named Employee_DB
   ii. Create a table employee with columns (empid, ename, department, salary)
   iii. Insert the ten records into employee table.
   iv. Load the employee data from the HDFS. Utilize the FOREACH operator to transform the employee table to extract the fields ‘empid’, ‘ename’,’salary’ and display the output.
   v. Filter the employee data to retrieve records of employees earning a salary greater than 20,000.

8. Write the Pig scripts for the following:
   i. Create a database named Employee_DB
   ii. Create a table employee with columns (empid, ename, department, salary)
   iii. Insert the ten records into employee table.
   iv. Group the employee dataset by department and display the output
   v. Sort the employee dataset based on the salary field in descending order to identify the highest-paid employees first.

9. Write the Pig scripts for the following:
   i. Create a database named Student_DB
   ii. Create a table student_info with columns (exam_no, name, class, department)
   iii. Insert the five records into student table.
   iv. Create a table result_info with columns (exam_no, percentage, grade)
   v. Insert the five records into result_info table.
   vi. Perform an inner join between the student_info and result_info and display output.

10. Write the Pig scripts for the following:
   i. Create a database named Student_DB
   ii. Create a table student_info with columns (exam_no, name, department, class, division)
   iii. Insert the five records into student table.
   iv. Create a table result_info with columns (exam_no, percentage, grade)
   v. Insert the five records into student table.
   vi. Perform the left, right and full outer join between the student_info and result_info and display output.

11. Write the Pig scripts for the following:
   i. Create a database named Student_DB
   ii. Create a table result_info with columns (exam_no, name, percentage)
   iii. Insert the ten records into result_info table.
   iv. Find the top three rankers and display their details.

12. Write the Pig scripts for the following:
   i. Create a database named Student_DB
   ii. Create a table result_info with columns (exam_no, name, percentage)
   iii. Insert the ten records into result_info table.
   iv. Randomly sample 20% of records from the result_info dataset to conduct statistical analysis and display the output
   v. Divide the result_info dataset into two separate datasets based on a condition on their percentage being greater than equal to 60% or below the 60%.

Problem Statements on Exploring R

13. Complete the following tasks in Rstudio using R:
   i. Demonstrate the use of following functions to handle data in R workspace
      a) ls() b) rm() c) getwd() d) save() e) load()
   ii. Write a script to take the student information such as roll no, name, marks of five subjects. Calculate total marks and percentage and display all student information.
   iii. Display the list of Built-in Datasets in R. Choose any dataset from it, summarize it and plot the graph of that dataset.

Problem Statements on Reading Datasets and Exporting Data from R

14. Complete the following tasks in Rstudio using R:
    1. Create the CSV file containing student information such as roll no, name and percentage. Read this file in R console and display the data.
    2. Create the data for student information such as roll no, name and percentage in R Console. Store this data into CSV file.

Problem Statements on Manipulating and Processing Data in R

15. Complete the following tasks in Rstudio using R:
    1. Create the data frame for Student Information containing roll no, student name and percentage. Display the list students who have failed, pass, second class, first class and distinction using subset function.
    2. Create the data frame for Employee Information containing emp id, emp name, designation and salary. Display the list of employees having salary greater than 30,000 using subset function and result should not contain the designation column.

16. Complete the following tasks in Rstudio using R:
    1. Create two data frames containing Student Information, first containing (exam no, name) and second containing exam no and marks of 5 subjects (exam no, S1, S2, S3, S4, S5). Merge these two data frames on the basis of exam no and display the result.
    2. Create two data frames containing Student Information, first containing (exam no, name, class) and second containing exam no, name and marks of 5 subjects (exam no, name, S1, S2, S3, S4, S5). Merge these two data frames on the basis of (exam no, name) and display the result.

17. Complete the following tasks in Rstudio using R:
    1. Create two data frames containing Student Information first Details(rollno, name, class) and second Marks(rollno,name, total_marks, percentage). Display the data of these two data frames and combine the first, second and third columns of Details data frame and third and fourth columns of the Marks data frame using cbind() function.
    2. Create two data frames containing employee information, first is sales_dept (empid, name) and second is finance_dept(empid, name) . Combine the rows of these two data frames using rbind() and display the result.

18. Complete the following tasks in Rstudio using R:
    i. Create data frame containing employee information Employee (empid, name, salary) and display the data in data frame. Then sort the data in data frame on the basis of salary in increasing order and display the result.
    ii. Create data frame containing student information Student (rollno, name, division, percentage) and display the data in data frame. Then sort the data in data frame first on the basis of division in increasing order and then on the basis of percentage in decreasing the order and display the result.

19. Complete the following tasks in Rstudio using R:
    i. Create data frame containing student information such as roll no , name and marks of three subjects Student (rollno, name, S1, S2, S3) and display the data in data frame.
    ii. Apply melt() function on this student data frame to create new data frame Student_long with roll no, name as ID variables and S1, S2, S3 as measurement variables and display the result.
    iii. Apply dcast() function on the Student_long to create new data frame Student_wide (rollno, name, Subject1, Subject2, Subject3) and display the result.

20. Create data frame containing employee information Employee (empid, name, salary) containing ten records. Then display the following details:
    i. All the records in data frame
    ii. First 3 rows in data frame
    iii. Last 3 rows in data frame
    iv. Only second column of data frame.

Prof. Kanishka .N. Kamble

Subject Mentor