Big Data Midterm (Fall’23)

Q4: Hive SQL

Create two HIVE tables corresponding to employees.csv and departments.csv (in Q4 directory). Table names should be Net-ID\_tablename format, e.g., asp13\_employees and asp13\_departments. For emplyees table create partitions with department\_id and two buckets with job\_id attributes.

Created the tables as required:

Commands:

For ys5250\_employees:

CREATE TABLE ys5250\_employees (

employee\_id INT,

first\_name STRING,

last\_name STRING,

email STRING,

phone\_number STRING,

hire\_date DATE,

job\_id STRING,

salary DECIMAL(10, 2),

commission\_pct DECIMAL(5, 2),

manager\_id INT

)

PARTITIONED BY (department\_id INT)

CLUSTERED BY (job\_id) INTO 2 BUCKETS

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

For ys5250\_departments:

CREATE TABLE ys5250\_departments (

department\_id INT,

department\_name STRING,

location\_id INT,

tbl\_last\_date STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

A screenshot of a computer program

Description automatically generated

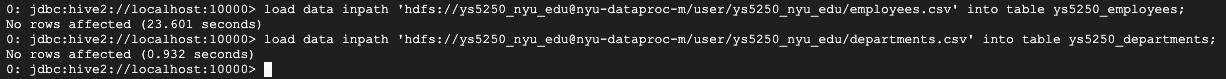
A screen shot of a computer

Description automatically generated

Uploading the data into tables:

load data inpath 'hdfs://ys5250\_nyu\_edu@nyu-dataproc-m/user/ys5250\_nyu\_edu/employees.csv' into table ys5250\_employees

load data inpath 'hdfs://ys5250\_nyu\_edu@nyu-dataproc-m/user/ys5250\_nyu\_edu/departments.csv' into table ys5250\_departments;



ys5250\_departments:

A screenshot of a computer

Description automatically generated

ys5250\_employees:

A screenshot of a computer screen

Description automatically generated

1. For each department find total salary. Your results should have department\_name and total salary of each department.

Query:

SELECT d.department\_name, SUM(e.salary) AS total\_salary

FROM ys5250\_employees e

JOIN ys5250\_departments d ON e.department\_id = d.department\_id

GROUP BY d.department\_name;

A screenshot of a computer

Description automatically generated

1. Find total number employees who works in ‘Sales’ department.

Query:

SELECT COUNT(\*) AS total\_sales\_employees

FROM ys5250\_employees e

JOIN ys5250\_departments d ON e.department\_id = d.department\_id

WHERE d.department\_name = 'Sales';

A computer screen with white text

Description automatically generated

1. Create a view with attributes: employee first name, last name, hire date, salary, department id, and department name. Include all departments even though there is no employee working for the department. Once view is created, query the view to list all details in ascending order of department id.

For creating a view:

Query:

CREATE VIEW employee\_details AS

SELECT e.first\_name, e.last\_name, e.hire\_date, e.salary, d.department\_id, d.department\_name

FROM ys5250\_departments d

LEFT JOIN ys5250\_employees e ON d.department\_id = e.department\_id;

A screen shot of a computer

Description automatically generated

For querying the view:

Query:

SELECT \*

FROM employee\_details

ORDER BY department\_id ASC;

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

And, so on. As we can see in the screenshots aboe that the results are in the ascending order of department id.

1. List employee number, first name, last name, salary for those employees whose salary is higher than average salary. Arrange the result in order of salary.

Average salary = 6805.5

A screen shot of a computer

Description automatically generated

Query:

WITH avg\_salary AS (

SELECT AVG(salary) AS avg\_sal

FROM ys5250\_employees

)

SELECT e.employee\_id, e.first\_name, e.last\_name, e.salary

FROM ys5250\_employees e

JOIN avg\_salary a ON e.salary > a.avg\_sal

ORDER BY e.salary;

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

1. Provide screenshot of HDFS directory for department\_id=20 and files under this directory (content of files are not requited, just the file names).

Commands:

hdfs dfs -ls /user/hive/warehouse/ys5250\_nyu\_edu.db/ys5250\_employees

hdfs dfs -ls /user/hive/warehouse/ys5250\_nyu\_edu.db/ys5250\_employees/department\_id=20

A black screen with white text

Description automatically generated