Exercise-3: Perform Sort, Group, Join, Project and Filter Operation in PIG

Note: You need to submit the complete program
as part of your assignment.

Report Template

Name: SWETHA K
Roll Number : 235229143
Class : II M.Sc Data Science
Subject : Big Data Analytics Lab
Date : 17-08-2024

Step-1: **Download the** data **files "employee** data.txt" and "dept data"

Step-2: Share **the** data **files to the Cloudera platform**

hdfs dfs -mkdir /user/cloudera/employee_data

hdfs dfs -mkdir /user/cloudera/dept_data

Step-3: **Move** the datafiles into theHDFS

hdfs dfs -put employee_data.txt /user/cloudera/employee_data/

hdfs dfs -put dept_data.txt /user/cloudera/dept_data/

Step-4*:* Enter into the PIGcommand prompt Open
Terminal in Cloudera Type "pig" and press Enter

Step- 5: Load **Employee** Data

```
employee_data = LOAD
'employee_data.txt' USING
PigStorage(',') AS (emp_id:int,
name:chararray, dept_id:int);
```

Step-6: Display the Employee Data

```
DUMP employee_ data;
```

```
(1,Human Resources,)
(2,Finance,)
(3,Marketing,)
(4,IT,)
(5,Sales,)
(6,Customer Service,)
(7,Research & Dev,)
(8,Operations,)
(9,Legal,)
(10,Administration,)
(1,akkim,)
(2,anu,)
(3,ashraf,)
```

Step–7: **Load** Department Data

dept data = LOAD 'dept_data.txt' USING PigStorage(',') AS (dept_ id:int, dept_ name:chararray);

step-8: Display the Department data

DUMP dept_ data;

```
(1,Human Resources)
(2,Finance)
(3,Marketing)
(4,IT)
(5,Sales)
(6,Customer Service)
(7,Research & Dev)
(8,Operations)
(9,Legal)
(10,Administration)
```

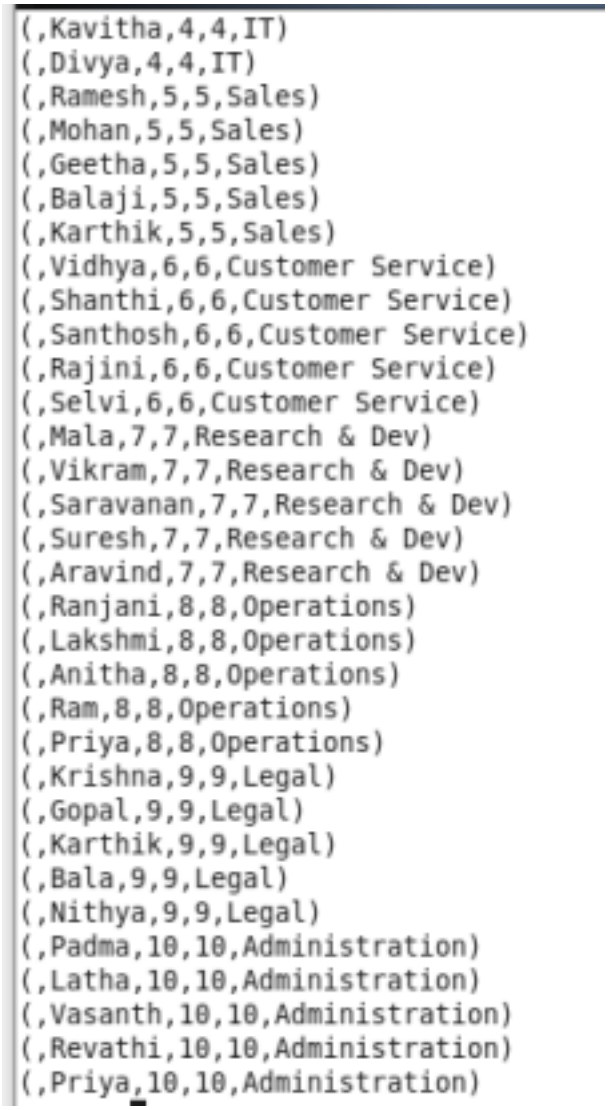
Step-9: **Join employee** data with dept data on dept id

joined_data = JOIN employee_data BY dept_id, dept_data BY

dept_id;

Step-10: Display the **Joined** data

DUMP joined_data;



Step-11:Project fields: **employee** name and department **name projected_data = FOREACH joined_data GENERATE employee_data::name AS emp_name, dept_data::dept_name AS dept_name;**

Step-12: Display the **projected** data

DUMP projected_data;

```
(Kavitha,IT)
(Divya,IT)
(Ramesh,Sales)
(Mohan,Sales)
(Geetha,Sales)
(Balaji,Sales)
(Karthik,Sales)
(Vidhya,Customer Service)
(Shanthi,Customer Service)
(Santhosh,Customer Service)
(Rajini,Customer Service)
(Selvi,Customer Service)
(Mala,Research & Dev)
(Vikram,Research & Dev)
(Saravanan,Research & Dev)
(Suresh,Research & Dev)
(Aravind,Research & Dev)
(Ranjani,Operations)
(Lakshmi,Operations)
(Anitha,Operations)
(Ram,Operations)
(Priya,Operations)
(Krishna,Legal)
(Gopal,Legal)
(Karthik,Legal)
(Bala,Legal)
(Nithya,Legal)
(Padma,Administration)
(Latha,Administration)
(Vasanth,Administration)
(Revathi,Administration)
(Priya,Administration)
```

Step-13:Group by department
grouped_by_dept = GROUP projected_data BY

dept_name;

Step-14*:* Display the grouped by dept

DUMP grouped_by_dept;

```
(IT,{(Divya,IT),(Kavitha,IT),(Aishwarya,IT),(Sindhu,IT),(Murali,IT)})
(Legal,{(Nithya,Legal),(Bala,Legal),(Karthik,Legal),(Gopal,Legal),(Krishna,Legal)})
(Sales,{(Ramesh,Sales),(Karthik,Sales),(Balaji,Sales),(Geetha,Sales),(Mohan,Sales)})
(Finance,{(Saranya,Finance),(Meena,Finance),(Radha,Finance),(Mani,Finance),(Suganya,Finance)})
(Marketing,{(Kumar,Marketing),(Vinoth,Marketing),(Prakash,Marketing),(Vijay,Marketing),(Sudha,Mar
keting)})
(Operations,{(Priya,Operations),(Ram,Operations),(Anitha,Operations),(Lakshmi,Operations),(Ranjan
i,Operations)})
(Administration,{(Priya,Administration),(Revathi,Administration),(Vasanth,Administration),(Latha,
Administration),(Padma,Administration)})
(Research & Dev,{(Mala,Research & Dev),(Aravind,Research & Dev),(Suresh,Research & Dev),(Saravana
n,Research & Dev),(Vikram,Research & Dev)})
(Human Resources,{(Kannan,Human Resources),(Rajesh,Human Resources),(Srinivasan,Human Resources),
(Ganesh,Human Resources),(Arun,Human Resources)})
(Customer Service,{(Vidhya,Customer Service),(Selvi,Customer Service),(Rajini,Customer Service),(
Santhosh,Customer Service),(Shanthi,Customer Service)})
```

Step-15: Count the **number of employees** per department

count_per_dept = FOREACH grouped_by_dept GENERATE group AS dept_name, COUNT(projected_data) AS employee_count;

Step-16: Order the **results** by **department** name **(ascending)**

ordered_by_dept_name = ORDER count_per_dept BY dept_name;

Step-17:Display the ordered by dept name

DUMP ordered_by_dept_name;

```
l - Total input paths to process : 1
(Administration,5)
(Customer Service,5)
(Finance,5)
(Human Resources,5)
(IT,5)
(Legal,5)
(Marketing,5)
(Operations,5)
(Research & Dev,5)
(Sales,5)
```

Step-18: Filter data to get employees in the Sales department

sales_employees = FILTER projected_data BY dept_name == 'Sales';

Step-19: Order Sales employees by name (ascending)

ordered_sales_employees = ORDER sales_employees BY emp_name;

Step-20: Display theOrdered sales employes

DUMP ordered_sales_employees;

```
(Balaji,Sales)
(Geetha,Sales)
(Karthik,Sales)
(Mohan,Sales)
(Ramesh,Sales)
```