Session 8: ADVANCED HIVE

Assignment 8.1

Student Name: Subham Vishal

Course: Big Data Hadoop & Spark Training

Assignment 8.1–
Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit

List of all employees who draw higher salary than the average salary of that department

Contents

Introduction

In this assignment, I'm going to write HIVE QL to achieve the task,

Problem Statement

Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit

List of all employees who draw higher salary than the average salary of that department

Dataset

```
101,Amitabh,20000,1
102,Shahrukh,10000,2
103,Akshay,11000,3
104,Anubhav,5000,4
105,Pawan,2500,5
106,Aamir,25000,1
107,Salman,17500,2
108,Ranbir,14000,3
109,Katrina,1000,4
110,Priyanka,2000,5
111,Tushar,500,1
112,Ajay,5000,2
113,Jubeen,1000,1
114,Madhuri,2000,2
```

Prerequisite – Create Database and Table

Using existing database **emp_details,**

Table –

We are creating a table name called **emp** and we have columns as **emp_id, emp_name, sal** and **dept.**

HIVE QL

CREATE TABLE emp

(

Emp_id int,

Emp_name string,

Sal int,

Dept int

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA LOCAL INPATH '/home/acadgild/hadoop/employee_details_task1.txt'

INTO TABLE emp_details.emp;

```
hive> CREATE TABLE emp
    > (
    > Emp_id int,
    > Emp_name string,
    > Sal int,
    > Dept int
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.515 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/hadoop/employee_details_task1.txt'
    > INTO TABLE emp_details.emp;
Loading data to table emp_details.emp
OK
Time taken: 1.002 seconds
hive>
```

Select * From emp;

```
hive (emp_details)>
                  >
                  > select * from emp;
OK
emp.emp_id      emp.emp_name    emp.sal emp.dept
101     Amitabh 20000   1
102     Shahrukh        10000   2
103     Akshay  11000   3
104     Anubhav 5000    4
105     Pawan   2500    5
106     Aamir   25000   1
107     Salman  17500   2
108     Ranbir  14000   3
109     Katrina 1000    4
110     Priyanka        2000    5
111     Tushar  500     1
112     Ajay    5000    2
113     Jubeen  1000    1
114     Madhuri 2000    2
Time taken: 0.187 seconds, Fetched: 14 row(s)
hive (emp_details)>
```

Task 1

Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit

HIVE QL

with temp as(select emp_id,emp_name,sal,dept, LEAD(sal, 1) OVER(PARTITION BY dept ORDER BY sal) - sal as diff FROM emp) select emp_id, emp_name, dept, sal from temp where diff >100;

```
Time taken: 67.071 seconds, Fetched: 9 row(s)
hive (emp_details)> with temp as(select emp_id,emp_name,sal,dept, LEAD(sal, 1) OVER(PARTITION BY dept ORDER BY sal) - sal as diff FROM   emp) s
elect emp_id, emp_name, dept, sal from temp where diff >100;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e.
spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20171122220745_85fa7164-e2d1-4f2d-9886-fc5b2e96ee3d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1511328525537_0003, Tracking URL = http://localhost:8088/proxy/application_1511328525537_0003/
Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job  -kill job_1511328525537_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-22 22:08:10,075 Stage-1 map = 0%,  reduce = 0%
2017-11-22 22:08:31,358 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.6 sec
2017-11-22 22:08:49,847 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 7.68 sec
2017-11-22 22:08:51,290 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.61 sec
MapReduce Total cumulative CPU time: 8 seconds 610 msec
Ended Job = job_1511328525537_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.61 sec   HDFS Read: 10903 HDFS Write: 364 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 610 msec
```

Required Output

```
emp_id  emp_name        dept    sal
111     Tushar  1       500
113     Jubeen  1       1000
101     Amitabh 1       20000
114     Madhuri 2       2000
112     Ajay    2       5000
102     Shahrukh        2       10000
103     Akshay  3       11000
109     Katrina 4       1000
110     Priyanka        5       2000
Time taken: 68.359 seconds, Fetched: 9 row(s)
hive (emp_details)>
```

Task 2

List of all employees who draw higher salary than the average salary of that department

HIVE QL

SELECT temp.emp_name, temp.sal, temp.dept, temp.avg_salary FROM (SELECT avg(sal) OVER (PARTITION BY dept) AS avg_salary, emp_id, emp_name, sal, dept FROM emp) temp WHERE temp.sal > temp.avg_salary;

```
                >
                > SELECT temp.emp_name, temp.sal, temp.dept, temp.avg_salary FROM (SELECT avg(sal) OVER (PARTITION BY dept) AS avg_salary, em
p_id, emp_name, sal, dept FROM emp) temp WHERE temp.sal > temp.avg_salary;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e.
spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20171122221126_1f3691d1-9a2b-41aa-94fa-aeea3bd4b508
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1511328525537_0004, Tracking URL = http://localhost:8088/proxy/application_1511328525537_0004/
Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job  -kill job_1511328525537_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-22 22:11:48,641 Stage-1 map = 0%,  reduce = 0%
2017-11-22 22:12:04,464 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.43 sec
2017-11-22 22:12:21,231 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.19 sec
MapReduce Total cumulative CPU time: 7 seconds 190 msec
Ended Job = job_1511328525537_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.19 sec   HDFS Read: 10717 HDFS Write: 328 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 190 msec
```

Required Output

```
OK
temp.emp_name   temp.sal        temp.dept       temp.avg_salary
Aamir   25000   1       11625.0
Amitabh 20000   1       11625.0
Shahrukh        10000   2       8625.0
Salman  17500   2       8625.0
Ranbir  14000   3       12500.0
Anubhav 5000    4       3000.0
Pawan   2500    5       2250.0
Time taken: 55.741 seconds, Fetched: 7 row(s)
hive (emp details)>
```