“Subjective Quiz”

1. Using SQOOP move all those employees whose designation is consultant from "mysqooptable" of "sqoopex" database from MySQL into HDFS into directory "/user/your_user_name/sqoop_import_query" directory
2. Use Pig script to replace the "consultant" role into "Big Data Consultant" and write the data into new HDFS directory "/user/your_user_name/BigData_Consultant_Data"
3. Create an external Hive table "Consultant_Table" representing this "Consultant_Data". This table will have 4 fields id,name,role and salary.
4. create a new bucketed table "Consultant_Table_Bucket" having 4 buckets on the field salary.
5. Insert all those employees whose salary is greater than 5000 into bucketed table "Consultant_Table_Bucket".
6. Write a Hive query to find out Max, min salary of Consultant from the "Consultant_Table_Bucket" table

Answer:

Step1:

Importing data from database table based on a condition into hdfs local directory using sqoop

sqoop import --connect jdbc:mysql://ip-172-31-13-154/sqoopex --username sqoopuser --password NHkkP876rp --table mysqooptable --where "designation = 'consultant'" --target-dir /user/dalonlobo2857/sqoop_import_query --fetch-size 10 --split-by salary

Verify the output using following command
hdfs dfs -cat /user/dalonlobo2857/sqoop_import_query/*

OUTPUT:
101,peter,consultant,10000.0
103,craig,consultant,8000.0
104,hunt,consultant,5000.0

Step 2:

sqoopxData = LOAD '/user/dalonlobo2857/sqoop_import_query/' using PigStorage(',') AS (empid:int, empname:chararray, designation:chararray, salary:float);

grunt> describe sqoopxData;
sqoopxData: {empid: int,empname: chararray,designation: chararray,salary: float}

Replacing consultant with ‘Big Data Consultant’
replacedData = FOREACH sqoopxData GENERATE
empid,empname,REPLACE(designation,'consultant','Big Data Consultant'), salary;

Storing it in hdfs directory
store replacedData into '/user/dalonlobo2857/BigData_Consultant_Data' using PigStorage(',');

Verify the output using following command
hdfs dfs -cat /user/dalonlobo2857/BigData_Consultant_Data/*

OUTPUT:
101,peter,Big Data Consultant,10000.0
103,craig,Big Data Consultant,8000.0
104,hunt,Big Data Consultant,5000.0
108,X,Big Data Consultant,5000.0

Step 3:Creating an external Hive table "Consultant_Table" and loading data

---

use dalon_test;

Setting hive execution engine to mapreduce
SET hive.execution.engine=mr;

create external table Consultant_Table(id int,name string,role string,salary float) row format delimited fields terminated by ',' stored as textfile location '/user/dalonlobo2857/BigData_Consultant_Data';

SELECT * FROM consultant_table;

OUTPUT:
101 peter Big Data Consultant 10000.0
103 craig Big Data Consultant 8000.0
104 hunt Big Data Consultant 5000.0
108 X Big Data Consultant 5000.0

Step 4

Creating bucketed table on field salary having 4 buckets

create external table Consultant_Table_Bucket(id int,name string,role string,salary float) clustered by (salary) into 4 buckets row format delimited fields terminated by ',' stored as textfile;

Step 5

Inserting employees whose salary is greater than 5000 into bucketed table "Consultant_Table_Bucket".

from Consultant_Table insert into table Consultant_Table_Bucket select id,name,role,salary where salary > 5000;

OUTPUT:

select * from Consultant_Table_Bucket;

OK
101 peter Big Data Consultant 10000.0
103 craig Big Data Consultant 8000.0
105 katharin Big Data Consultant 10000.0
107 A Big Data Consultant 40000.0
107 A Big Data Consultant 40000.0
107 A Big Data Consultant 40000.0
Time taken: 0.057 seconds,
Fetched: 6 row(s)

Step 6

Finding the maximum and minimum salary:

select MAX(salary) as maximum, MIN(salary) as minimum from Consultant_Table_Bucket;

OUTPUT:
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 1 Cumulative CPU: 6.28 sec HDFS Read: 988 HDFS Write: 15 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 280 msec
OK
40000.0 8000.0
Time taken: 20.056 seconds, Fetched: 1 row(s)

Maximum salary = 40000.0
Minimum salary = 8000.0