1. Using Hadoop command move all those employees data into HDFS directory "/user/your_user_name/employees_data" directory

---

hadoop fs -mkdir /user/your_user_name/employees_data hadoop fs -mv <source file path> /user/your_user_name/employees_data hadoop fs -ls /user/your_user_name/employees_data/

---

2. Create an external Hive table "employees_Table" representing this "employees_data". This table will have 5 fields id,age,gender,role and salary.

---

create external table employees_Table(id string, age int, gender string, role string, salary int)

COMMENT ‘Employee details’

row format delimited

fields terminated by ',' //use “\t” if file is tab separated.

lines terminated by '\n'

STORED AS TEXTFILE;

load data local inpath "/user/your_user_name/employees_data/" into table employees_Table;

load data inpath "hdfs://localhost:8020/user/your_user_name/employees_data/" into table employees_Table;

//select * from employees_Table limit 1000;

---

3. Create a new bucketed table "Consultant_Table_Bucket" having 4 buckets on the field salary. This table should store the data into columnar format ORC.

---

set hive.enforce.bucketing = true; // (Note: Not needed in Hive 2.x onward)

CREATE TABLE Consultant_Table_Bucket(id string, age int, gender string, role string, salary int)

COMMENT 'bucketed salary field into 4 buckets'

CLUSTERED BY(salary) INTO 4 BUCKETS

stored as orc tblproperties ("orc.compress"="ZLIB"); //compression

is not needed just storing as ORC is enough. describe extended

table Consultant_Table_Bucket;

// it will show the in-depth details of Consultant_Table_Bucket

---

4. Insert all those employees whose salary is greater than 5000 into bucketed table "Consultant_Table_Bucket". While inserting into "Consultant_Table_Bucket" table you need to convert "consultant" role into "BigData Consultant" role.

---

insert into Consultant_Table_Bucket as select id, age, gender, regexp_replace(role,"**consultant**","**BigData Consultant**") as role, salary from where employees_Table salary > 5000;

or insert into Consultant_Table_Bucket as select id, age, gender, CASE when role is "**consultant** " THEN " **BigData Consultant** " ELSE role END AS role, salary from where employees_Table salary > 5000;

select distinct role from Consultant_Table_Bucket; // should return only distinct roles

select role, salary from Consultant_Table_Bucket where salary < 5000; // should return empty

---

5.Write a Hive query to find out Max, min salary of "BigData Consultant" from the "Consultant_Table_Bucket" table.

---

select max(salary) as max_salary, min(salary) as min_salary from Consultant_Table_Bucket where role="BigData Consultant";
