database name :venkat_test

1. Using SQOOP move all those employees whose designation is consultant from "mysqooptable" of "sqoopex" database from MySQL into HDFS into directory "**/user/iaminusa19828041/sqoop_import_query4**" directory

Answer:

```
sqoop import --connect jdbc:mysql://ip-172-31-13-154/sqoopex --username
sqoopuser --password NHkkP876rp --table mysqooptable --where
"designation='consultant'"
 --target-dir /user/iaminusa19828041/sqoop_import_query4  --split-by salary
```

2. Use Pig script to replace the "consultant" role into "Big Data Consultant" and write the data into new HDFS directory "**/user/iaminusa19828041/bigdataconsultantdata2"**

Answer :

```
testA = LOAD '/user/iaminusa19828041/bigdataconsultantdata/emp_data.txt'
USING  PigStorage (',') as (id,name,desi
gnation,sal);

grunt> B = FOREACH testA generate id,name,designation,sal;


grunt> dump B;
```

```
nalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-07-17 18:17:44,814 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunc
her - Su
ccess!
2017-07-17 18:17:44,821 [main] INFO  org.apache.pig.data.SchemaTupleBackend -
Key [pig.schematuple] was not set... will
not generate code.
```

2017-07-17 18:17:44,839 [main] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to proces
s : 1
2017-07-17 18:17:44,840 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths t
o process : 1
(107,A,consultant,40000.0)
(107,A,consultant,40000.0)
(107,A,consultant,40000.0)
(101,peter,consultant,10000.0)
(103,craig,consultant,8000.0)
(104,hunt,consultant,5000.0)
(108,X,consultant,5000.0)
(105,katharin,consultant,10000.0)

rep_data = FOREACH B GENERATE
(id,name,designation,sal),REPLACE(designation,'consultant','Bigconsultant');

Total records proactively spilled: 0

Job DAG:
job_1498404677707_5647

2017-07-17 18:36:43,513 [main] INFO
org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address:
http://ip-172-31-13-154.ec2.internal
:8188/ws/v1/timeline/
2017-07-17 18:36:43,513 [main] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-172-31-53-48.ec2.internal/172.31.53.48:8050
2017-07-17 18:36:43,516 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirec
ting to job history server
2017-07-17 18:36:43,602 [main] INFO
org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address:
http://ip-172-31-13-154.ec2.internal
:8188/ws/v1/timeline/
2017-07-17 18:36:43,602 [main] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-172-31-53-48.ec2.internal/172.31.53.48:8050
2017-07-17 18:36:43,605 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirec
ting to job history server

2017-07-17 18:36:43,674 [main] INFO
org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address:
http://ip-172-31-13-154.ec2.internal
:8188/ws/v1/timeline/
2017-07-17 18:36:43,674 [main] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-172-31-53-48.ec2.internal/172.31.53.48:8050
2017-07-17 18:36:43,678 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirec
ting to job history server
2017-07-17 18:36:43,701 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunc
her - Success!
2017-07-17 18:36:43,702 [main] INFO org.apache.pig.data.SchemaTupleBackend -
Key [pig.schematuple] was not set... will not generate code.
2017-07-17 18:36:43,707 [main] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
process : 1
2017-07-17 18:36:43,707 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1
((107,A,consultant,40000.0),Bigconsultant)
((107,A,consultant,40000.0),Bigconsultant)
((107,A,consultant,40000.0),Bigconsultant)
((101,peter,consultant,10000.0),Bigconsultant)
((103,craig,consultant,8000.0),Bigconsultant)
((104,hunt,consultant,5000.0),Bigconsultant)
((108,X,consultant,5000.0),Bigconsultant)
((105,katharin,consultant,10000.0),Bigconsultant)

3. Create an external Hive table "Consultant_Table" representing this "Consultant_Data". This table will have 4 fields id,name,role and salary.

hive (venkat_test)> create external table consultant_table_data(id int,name string, role string,sal float) row format delimited fields terminated by ',' LO CATION '/user/iaminusa19828041/hve';
OK
Time taken: 0.062 seconds
hive (venkat_test)> show tables;
OK

```
consultant1_table
consultant_table
consultant_table_bucket
consultant_table_data
emp_dynamic_partition
emp_global
emp_orc
Time taken: 0.045 seconds, Fetched: 7 row(s)
hive (venkat_test)> LOAD data inpath '/user/iaminusa19828041/etab/
emp_data.txt' into table consultant_table_data;
Loading data to table venkat_test.consultant_table_data
Table venkat_test.consultant_table_data stats: [numFiles=1, numRows=0,
totalSize=215, rawDataSize=0]
OK
Time taken: 0.193 seconds
hive (venkat_test)> select * from consultant_table_data
                  > ;
OK
107     A       consultant      40000.0
107     A       consultant      40000.0
107     A       consultant      40000.0
101     peter   consultant      10000.0
103     craig   consultant      8000.0
104     hunt    consultant      5000.0
108     X       consultant      5000.0
105     katharin        consultant      10000.0
Time taken: 0.073 seconds, Fetched: 8 row(s)
```

————

4. create a new bucketed table "Consultant_Table_Bucket" having 4 buckets on the field salary.

```
hive (venkat_test)> create table consultant_table_bucket1(id int,name string, role
string,sal float) clustered by (sal) into   4 buckets row format delimi
ted fields terminated by ‘,';

OK
Time taken: 0.193 seconds
```

5. Insert all those employees whose salary is greater than 5000 into bucketed table "Consultant_Table_Bucket".

hive (venkat_test)> insert overwrite table consultant_table_bucket select * from consultant_table where (sal>5000);
Query ID = iaminusa19828041_20170717164922_e5f0851d-d773-4f56-b501-bd018733e8c4
Total jobs = 1
Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1498404677707_5621)

VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED

Map 1 ..... SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... SUCCEEDED 4 4 0 0 1 0

VERTICES: 02/02 [==========>>] 100% ELAPSED TIME: 7.51 s

Loading data to table venkat_test.consultant_table_bucket
[Warning] could not update stats.
OK
Time taken: 8.061 seconds

hive (venkat_test)> select * from consultant_table_bucket
> ;
OK
107 A consultant 40000.0
107 A consultant 40000.0
107 A consultant 40000.0
101 peter consultant 10000.0
103 craig consultant 8000.0
105 katharin consultant 10000.0

————

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

hive (venkat_test)> select max(sal),min(sal) from consultant_table_bucket;

```
Query ID =
iaminusa19828041_20170717175038_c80efcb3-6cb4-426c-8a7d-02b56c7025ed
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id
application_1498404677707_5638)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.04 s
--------------------------------------------------------------------------------
OK
40000.0 8000.0
Time taken: 4.412 seconds, Fetched: 1 row(s)
hive (venkat_test)>
```