Session 22: Interview Preparation-II

Assignment – Case Study II- Hive and HBase Integration

Task 1 –

Let us take up the CUSTOMER and TRANSACTIONS table we have created in the Let's Do Together section. Let us solve the following use cases using these tables :-

Find out the number of transaction done by each customer

```
hive> select fname,count(txnno) from CUSTOMER a join TRANSACTIONS b on a.custid = b.custno group by fname;
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_20180709032802_e9bcebfe-52a4-4c01-a203-b6b23d4d7cf9
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2018-07-09 03:28:17     Starting to launch local task to process map join;      maximum memory = 518979584
2018-07-09 03:28:22     Dump the side-table for tag: 0 with group count: 7 into file: file:/tmp/acadgild/f6e76ffa-175f-4d89-859a-419deb41
1f5d/hive_2018-07-09_03-28-02_068_8310031496676592178-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile140--.hashtable
2018-07-09 03:28:22     Uploaded 1 File to: file:/tmp/acadgild/f6e76ffa-175f-4d89-859a-419deb411f5d/hive_2018-07-09_03-28-02_068_83100314
96676592178-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile140--.hashtable (442 bytes)
2018-07-09 03:28:22     End of local task; Time Taken: 4.79 sec.
Execution completed successfully
MapredLocal task succeeded
```

```
MapReduce Total cumulative CPU time: 8 seconds 100 msec
Ended Job = job_1531036500831_0014
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 8.1 sec   HDFS Read: 12967 HDFS Write: 213 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 100 msec
OK
Aamir   1
Anubahv 1
Pawan   1
Ranbir  1
Salman  1
Sharukh 1
Time taken: 86.616 seconds, Fetched: 6 row(s)
```

Task 2 -

Create a new table called TRANSACTIONS_COUNT. This table should have 3 fields - custid, fname and count.


```
hive> CREATE TABLE TRANSACTIONS_COUNT (custid int,name string,count int) row format delimited fields terminated by ',';
OK
Time taken: 0.139 seconds
hive> describe transactions_count
    > ;
OK
custid              int
name                string
count               int
Time taken: 0.125 seconds, Fetched: 3 row(s)
```

Task 3 –

Now write a hive query in such a way that the query populates the data
obtained in Step 1 above and populate the table
in step 2 above.

```
hive> insert into table transactions_count select custid,fname,count(txnno) from CUSTOMER a join TRANSACTIONS b on a.custid = b.custno gr
oup by custid,fname;
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_20180709034349_b27dd864-ea36-4103-8b1d-9e44391baf56
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2018-07-09 03:44:11     Starting to launch local task to process map join;     maximum memory = 518979584
2018-07-09 03:44:21     Dump the side-table for tag: 0 with group count: 7 into file: file:/tmp/acadgild/f6e76ffa-175f-4d89-859a-419deb41
1f5d/hive_2018-07-09_03-43-49_593_3367704929699502689-1/-local-10003/HashTable-Stage-2/MapJoin-mapfile150--.hashtable
2018-07-09 03:44:23     Uploaded 1 File to: file:/tmp/acadgild/f6e76ffa-175f-4d89-859a-419deb411f5d/hive_2018-07-09_03-43-49_593_33677049
```

```
hive> select * from transactions_count;
OK
102     Sharukh 1
104     Anubahv 1
105     Pawan   1
106     Aamir   1
107     Salman  1
108     Ranbir  1
Time taken: 0.256 seconds, Fetched: 6 row(s)
```

Task 4 –

Now lets make the TRANSACTIONS_COUNT table Hbase complaint. In the sense, use Ser Des And Storage handler features of hive to change the TRANSACTIONS_COUNT table to be able to create a TRANSACTIONS table in Hbase.

```
CREATE TABLE TRANSACTIONS_COUNT(custid int, fname string, count int)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,cf1:fname,cf1:count")
TBLPROPERTIES ("hbase.table.name" = "TRANSACTIONS");
```

```
hive> CREATE TABLE transactions_count(custid int,fname string, count int)
    > STORED BY "org.apache.hadoop.hive.hbase.HBaseStorageHandler"
    > WITH SERDEPROPERTIES("hbase.columns.mapping"=":key,cf1:fname,cf1:count")
    > TBLPROPERTIES("hbase.table.name"="TRANSACTIONS");
OK
Time taken: 12.603 seconds
```

```
hbase(main):013:0> list
TABLE
TRANSACTIONS
1 row(s) in 0.0190 seconds

=> ["TRANSACTIONS"]
hbase(main):014:0> describe "TRANSACTIONS"
Table TRANSACTIONS is ENABLED
TRANSACTIONS
COLUMN FAMILIES DESCRIPTION
{NAME => 'cf1', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODI
NG => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLI
CATION_SCOPE => '0'}
1 row(s) in 0.0840 seconds

hbase(main):015:0>
```

Task 5 –

Now insert the data in TRANSACTIONS_COUNT table using the query in step
3 again, this should populate the Hbase TRANSACTIONS table automatically

```
hive> insert into table transactions_count select custid,fname,count(txnno) from CUSTOMER a join TRANSACTIONS b on a.custid =
 b.custno group by custid,fname;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180709202801_2607f494-a44d-48b1-8141-e9aa63f825bf
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j
/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
```

```
hbase(main):019:0> list
TABLE
TRANSACTIONS
1 row(s) in 0.0270 seconds

=> ["TRANSACTIONS"]
hbase(main):020:0> scan "TRANSACTIONS"
ROW                           COLUMN+CELL
 102                          column=cf1:count, timestamp=1531148452232, value=1
 102                          column=cf1:fname, timestamp=1531148452232, value=Sharukh
 104                          column=cf1:count, timestamp=1531148452232, value=1
 104                          column=cf1:fname, timestamp=1531148452232, value=Anubahv
 105                          column=cf1:count, timestamp=1531148452232, value=1
 105                          column=cf1:fname, timestamp=1531148452232, value=Pawan
 106                          column=cf1:count, timestamp=1531148452232, value=1
 106                          column=cf1:fname, timestamp=1531148452232, value=Aamir
 107                          column=cf1:count, timestamp=1531148452232, value=1
 107                          column=cf1:fname, timestamp=1531148452232, value=Salman
 108                          column=cf1:count, timestamp=1531148452232, value=1
 108                          column=cf1:fname, timestamp=1531148452232, value=Ranbir
6 row(s) in 0.7220 seconds
```