CASE STUDY 2

We have created two tables i.e. **customers** and **transactions** during the session.

1. **Find the count of transactions for each customer.**

Sol: **Command used:**

> ➔ select fname, custno, count(custno) from transactions, customer where custid=custno
> group by fname, custno;

```
acadgild@localhost:~                                                        −   □   ×
hive>
hive> select fname, custno, count(custno) from transactions, customer where custid=custno group by fname, custno;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different ex
ecution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180515052459_b08561d0-0ed9-4217-9de4-256445a6f6e2
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/sl
f4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.9.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.
jar!/org/slf4j/impl/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]
2018-05-15 05:25:19     Starting to launch local task to process map join;     maximum memory = 518979584
2018-05-15 05:25:25     Dump the side-table for tag: 1 with group count: 8 into file: file:/tmp/acadgild/bee967af-dd67-411
e-9f97-a0fe33120558/hive_2018-05-15_05-24-59_297_1321412203390748796-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile41--.
hashtable
2018-05-15 05:25:25     Uploaded 1 File to: file:/tmp/acadgild/bee967af-dd67-411e-9f97-a0fe33120558/hive_2018-05-15_05-24-
59_297_1321412203390748796-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile41--.hashtable (469 bytes)
2018-05-15 05:25:25     End of local task; Time Taken: 5.532 sec.
Execution completed successfully
MapredLocal task succeeded
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_1526339074855_0006, Tracking URL = http://localhost:8088/proxy/application_1526339074855_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill job_1526339074855_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-05-15 05:25:47,359 Stage-2 map = 0%,   reduce = 0%
2018-05-15 05:26:04,287 Stage-2 map = 100%,   reduce = 0%, Cumulative CPU 4.19 sec
```

```
acadgild@localhost:~                                                        −   □   ×
hashtable
2018-05-15 05:25:25     Uploaded 1 File to: file:/tmp/acadgild/bee967af-dd67-411e-9f97-a0fe33120558/hive_2018-05-15_05-24-
59_297_1321412203390748796-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile41--.hashtable (469 bytes)
2018-05-15 05:25:25     End of local task; Time Taken: 5.532 sec.
Execution completed successfully
MapredLocal task succeeded
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_1526339074855_0006, Tracking URL = http://localhost:8088/proxy/application_1526339074855_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill job_1526339074855_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-05-15 05:25:47,359 Stage-2 map = 0%,   reduce = 0%
2018-05-15 05:26:04,287 Stage-2 map = 100%,   reduce = 0%, Cumulative CPU 4.19 sec
2018-05-15 05:26:18,397 Stage-2 map = 100%,   reduce = 100%, Cumulative CPU 7.52 sec
MapReduce Total cumulative CPU time: 7 seconds 520 msec
Ended Job = job_1526339074855_0006
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1   Cumulative CPU: 7.52 sec   HDFS Read: 13388 HDFS Write: 263 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 520 msec
OK
Amitabh 101     2
Sharukh 102     1
Anubahv 104     1
Pawan   105     1
Aamir   106     1
Salman  107     1
Ranbir  108     1
Time taken: 80.313 seconds, Fetched: 7 row(s)
hive>
```

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

Command used:

➔ create table transactions_count_1(
   fname string, custid int,
   count int) row format delimited fields terminated by ' ';

```
hive> create table TRANSACTIONS_COUNT_1(
    > fname string,
    > custid int,
    > count int) row format delimited fields terminated by ' ';
OK
Time taken: 0.739 seconds
hive> insert into TRANSACTIONS_COUNT_1 select fname, custno, count(custno) from
transactions, customer where custid=custno group by fname, custno;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the futu
re versions. Consider using a different execution engine (i.e. spark, tez) or us
ing Hive 1.X releases.
Query ID = acadgild_20180516044338_b25f6504-7b2f-4c1e-ada6-9718847070ef
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-2.9.0/sha
re/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/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]
2018-05-16 04:44:00     Starting to launch local task to process map join;      m
aximum memory = 518979584
2018-05-16 04:44:05     Dump the side-table for tag: 1 with group count: 8 into
file: file:/tmp/acadgild/e9d40923-62d9-4f81-a7b4-65831edba9b5/hive_2018-05-16_04
-43-38_492_3292043870929906869-1/-local-10003/HashTable-Stage-2/MapJoin-mapfile0
1--.hashtable
2018-05-16 04:44:05     Uploaded 1 File to: file:/tmp/acadgild/e9d40923-62d9-4f8
1-a7b4-65831edba9b5/hive_2018-05-16_04-43-38_492_3292043870929906869-1/-local-10
003/HashTable-Stage-2/MapJoin-mapfile01--.hashtable (469 bytes)
2018-05-16 04:44:05     End of local task; Time Taken: 5.05 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
```

3. To populate the data into recently created table:

Command used:

➔ insert into Transactions_count_1 select fname, custno, count(custno) from transactions, customer where custid = custno group by fname, custno;

```
Time taken: 0.739 seconds
hive> insert into TRANSACTIONS_COUNT_1 select fname, custno, count(custno) from
transactions, customer where custid=custno group by fname, custno;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the futu
re versions. Consider using a different execution engine (i.e. spark, tez) or us
ing Hive 1.X releases.
Query ID = acadgild_20180516044338_b25f6504-7b2f-4c1e-ada6-9718847070ef
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-2.9.0/sha
re/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/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]
2018-05-16 04:44:00     Starting to launch local task to process map join;      m
aximum memory = 518979584
2018-05-16 04:44:05     Dump the side-table for tag: 1 with group count: 8 into
file: file:/tmp/acadgild/e9d40923-62d9-4f81-a7b4-65831edba9b5/hive_2018-05-16_04
-43-38_492_3292043870929906869-1/-local-10003/HashTable-Stage-2/MapJoin-mapfile0
1--.hashtable
2018-05-16 04:44:05     Uploaded 1 File to: file:/tmp/acadgild/e9d40923-62d9-4f8
1-a7b4-65831edba9b5/hive_2018-05-16_04-43-38_492_3292043870929906869-1/-local-10
003/HashTable-Stage-2/MapJoin-mapfile01--.hashtable (469 bytes)
2018-05-16 04:44:05     End of local task; Time Taken: 5.05 sec.
Execution completed successfully
MapredLocal task succeeded
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:
```

```
Time taken: 0.072 seconds
hive> show tables;
OK
customer
transactions
transactions_count
transactions_count_1
transactions_hbase
Time taken: 0.15 seconds, Fetched: 5 row(s)
hive> select * from transactions_count_1;
OK
Amitabh 101     2
Sharukh 102     1
Anubahv 104     1
Pawan   105     1
Aamir   106     1
Salman  107     1
Ranbir  108     1
Time taken: 4.843 seconds, Fetched: 7 row(s)
hive>
    >
    >
    >
    >
    >
    >
    >
    >
    >
    >
    >
    >
    >
```

4. Make the transactions_count table Hbase complaint and use serde and storage handler property to put the data into it

Command used :

➔ create table transactions_hbase1(username string, id int, count int) stored by 'org.apache.hadoop.hive.hbase.HbaseStorageHandler' with serdeproperties ('hbase.columns.mapping' = ':key, ID:id, stats:count');


```
)
hive>
    > CREATE TABLE TRANSACTIONS_HBase1(username STRING, id INT, count INT)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ('hbase.columns.mapping' = ':key, ID:id ,stats:count');
OK
Time taken: 4.296 seconds
hive> show tables;
OK
customer
transactions
transactions_count
transactions_count_1
transactions_hbase
transactions_hbase1
Time taken: 0.099 seconds, Fetched: 6 row(s)
hive> insert into TRANSACTIONS_HBase1
    > select * from TRANSACTIONS_COUNT_1;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Con
sider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180517053222_6ca7af16-54d1-422a-9cc0-f2e4e8e3dfc6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1526507568593_0001, Tracking URL = http://localhost:8088/proxy/application_15
26507568593_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill job_152650756859
3_0001
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2018-05-17 05:33:31,789 Stage-3 map = 0%,  reduce = 0%
2018-05-17 05:33:52,695 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 4.6 sec
MapReduce Total cumulative CPU time: 4 seconds 600 msec
Ended Job = job_1526507568593_0001
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1    Cumulative CPU: 4.6 sec   HDFS Read: 11032 HDFS Write: 0 SUCCESS
```

➔ describe 'acadgilddb.transactions_hbase1'

➔ scan 'acadgilddb.transactions_hbase1'   //to check if any rows present in it.

```
acadgild@localhost:~/install/hbase/hbase-1.2.6/bin                                    —   □   ×

hbase(main):001:0> describe 'acadgilddb.TRANSACTIONS_hbase1'

ERROR: Unknown table acadgilddb.TRANSACTIONS_hbase1!

Here is some help for this command:
Describe the named table. For example:
  hbase> describe 't1'
  hbase> describe 'ns1:t1'

Alternatively, you can use the abbreviated 'desc' for the same thing.
  hbase> desc 't1'
  hbase> desc 'ns1:t1'


hbase(main):002:0> describe 'acadgilddb.transactions_hbase1'
Table acadgilddb.transactions_hbase1 is ENABLED
acadgilddb.transactions_hbase1
COLUMN FAMILIES DESCRIPTION
{NAME => 'ID', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP
_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMP
RESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '6553
6', REPLICATION_SCOPE => '0'}
{NAME => 'stats', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', K
EEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', C
OMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '6
5536', REPLICATION_SCOPE => '0'}
2 row(s) in 0.2600 seconds

hbase(main):003:0> scan 'acadgilddb.transactions_hbase1'
ROW                      COLUMN+CELL
0 row(s) in 0.0750 seconds

hbase(main):004:0> scan 'acadgilddb.transactions_hbase1'
ROW                      COLUMN+CELL
```

5. populate the data into transactions_hbase1   from the hive table

command used:

➔ insert into transactions_hbase1 select * from transactions_count_1;

```
acadgild@localhost:~                                                                  —   □   ×
Time taken: 0.099 seconds, Fetched: 6 row(s)
hive> insert into TRANSACTIONS_HBase1
    > select * from TRANSACTIONS_COUNT_1;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Con
sider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180517053222_6ca7af16-54d1-422a-9cc0-f2e4e8e3dfc6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1526507568593_0001, Tracking URL = http://localhost:8088/proxy/application_15
26507568593_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill job_152650756859
3_0001
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2018-05-17 05:33:31,789 Stage-3 map = 0%,  reduce = 0%
2018-05-17 05:33:52,695 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 4.6 sec
MapReduce Total cumulative CPU time: 4 seconds 600 msec
Ended Job = job_1526507568593_0001
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 4.6 sec   HDFS Read: 11032 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 600 msec
OK
Time taken: 91.569 seconds
hive>
    >
    >
    >
    >
    >
    >
    >
    >
    >
    >
    >
```

6. to check the data into hbase shell

command used:

➔ scan 'acadgilddb.transactions_hbase1'

```
hbase(main):003:0> scan 'acadgilddb.transactions_hbase1'
ROW                    COLUMN+CELL
0 row(s) in 0.0750 seconds

hbase(main):004:0> scan 'acadgilddb.transactions_hbase1'
ROW                    COLUMN+CELL
 Aamir                 column=ID:id, timestamp=1526515431222, value=106
 Aamir                 column=stats:count, timestamp=1526515431222, value=1
 Amitabh               column=ID:id, timestamp=1526515431222, value=101
 Amitabh               column=stats:count, timestamp=1526515431222, value=2
 Anubahv               column=ID:id, timestamp=1526515431222, value=104
 Anubahv               column=stats:count, timestamp=1526515431222, value=1
 Pawan                 column=ID:id, timestamp=1526515431222, value=105
 Pawan                 column=stats:count, timestamp=1526515431222, value=1
 Ranbir                column=ID:id, timestamp=1526515431222, value=108
 Ranbir                column=stats:count, timestamp=1526515431222, value=1
 Salman                column=ID:id, timestamp=1526515431222, value=107
 Salman                column=stats:count, timestamp=1526515431222, value=1
 Sharukh               column=ID:id, timestamp=1526515431222, value=102
 Sharukh               column=stats:count, timestamp=1526515431222, value=1
7 row(s) in 0.3030 seconds

hbase(main):005:0>
hbase(main):006:0*
hbase(main):007:0*
hbase(main):008:0*
hbase(main):009:0*
hbase(main):010:0*
hbase(main):011:0*
hbase(main):012:0*
hbase(main):013:0*
hbase(main):014:0*
hbase(main):015:0*
hbase(main):016:0*
```