First part of To Do activity

create database acadgilddb

use acadgilddb;

```
hive> create database acadgilddb;
OK
Time taken: 0.592 seconds
hive> use acadgilddb;
OK
Time taken: 0.171 seconds
hive>
```

```
CREATE TABLE CUSTOMER(
        custid INT,
        fname STRING,
        lname STRING,
        age INT,
        profession STRING)
row format delimited fields terminated by ',';
```

```
hive> CREATE TABLE CUSTOMER(
    > custid INT,
    > fname STRING,
    > lname STRING,
    > age INT,
    > profession STRING)
    > row format delimited fields terminated by ',';
OK
Time taken: 3.782 seconds
hive>
```

LOAD DATA LOCAL INPATH '<your location of attached custs.txt file>/custs.txt' into table CUSTOMER;

101,Amitabh,Bacchan,65,Actor
102,Sharukh,Khan,45,Doctor
103,Akshay,Kumar,38,Dentist
104,Anubahv,kumar,58,Business
105,Pawan,Trivedi,34,service
106,Aamir,Null,42,scientest
107,Salman,Khan,43,Surgen
108,Ranbir,Kapoor,26,Industrialist

```
[acadgild@localhost hive]$ cat > customer.txt
101,Amitabh,Bacchan,65,Actor
102,Sharukh,Khan,45,Doctor
103,Akshay,Kumar,38,Dentist
104,Anubahv,kumar,58,Business
105,Pawan,Trivedi,34,service
106,Aamir,Null,42,scientest
107,Salman,Khan,43,Surgen
108,Ranbir,Kapoor,26,Industrialist
^C
[acadgild@localhost hive]$
```

```
LOAD DATA LOCAL INPATH '/home/acadgild/hive/customer.txt'
into table CUSTOMER;
```

```
hive> LOAD DATA LOCAL INPATH '/home/acadgild/hive/customer.txt'
    > into table CUSTOMER;
Loading data to table acadgilddb.customer
OK
Time taken: 4.924 seconds
hive>
```

```
hive> select * from CUSTOMER;
OK
101     Amitabh Bacchan 65      Actor
102     Sharukh Khan    45      Doctor
103     Akshay  Kumar   38      Dentist
104     Anubahv kumar   58      Business
105     Pawan   Trivedi 34      service
106     Aamir   Null    42      scientest
107     Salman  Khan    43      Surgen
108     Ranbir  Kapoor  26      Industrialist
Time taken: 1.806 seconds, Fetched: 8 row(s)
hive>
```

```
CREATE TABLE TRANSACTIONS (
                txnno INT,
                txndate STRING,
                custno INT,
                amount DOUBLE,
                category STRING,
                product STRING,
                city STRING,
                state STRING,
                spendby STRING)
    row format delimited fields terminated by ',';
```

```
hive> CREATE TABLE TRANSACTIONS (
    > txnno INT,
    > txndate STRING,
    > custno INT,
    > amount DOUBLE,
    > category STRING,
    > product STRING,
    > city STRING,
    > state STRING,
    > spendby STRING)
    > row format delimited fields terminated by ',';
OK
Time taken: 0.756 seconds
hive>
```

```
cat > transactions.txt
97834,05/02/2018,101,965,Entertainment,Movie,Pune,Maharashtra,Daughter
98396,12/01/2018,102,239,Food,Grocery,Patna,Bihar,Self
34908,06/01/2018,101,875,Travel,Air,Bangalore,Karnataka,Spouse
70958,17/02/2018,104,439,Food,Restaurant,Delhi,Delhi,Wife
09874,21/01/2018,105,509,Entertainment,Park,Kolkata,West Bengal
94585,19/01/2018,106,629,Rent,House,Hyderabad,Telangana,Self
45509,20/01/2018,107,953,Travel,Rail,Chennai,Tamil Nadu,Brother
07864,01/02/2018,108,569,Rent,Parking,Goa,Goa,Wife
```

```
[acadgild@localhost hive]$ cat > transactions.txt
97834,05/02/2018,101,965,Entertainment,Movie,Pune,Maharashtra,Daughter
98396,12/01/2018,102,239,Food,Grocery,Patna,Bihar,Self
34908,06/01/2018,101,875,Travel,Air,Bangalore,Karnataka,Spouse
70958,17/02/2018,104,439,Food,Restaurant,Delhi,Delhi,Wife
09874,21/01/2018,105,509,Entertainment,Park,Kolkata,West Bengal
94585,19/01/2018,106,629,Rent,House,Hyderabad,Telangana,Self
45509,20/01/2018,107,953,Travel,Rail,Chennai,Tamil Nadu,Brother
07864,01/02/2018,108,569,Rent,Parking,Goa,Goa,Wife
^C
[acadgild@localhost hive]$
```

```
LOAD DATA LOCAL INPATH '/home/acadgild/hive/transactions.txt'
into table TRANSACTIONS;
```

```
select * from TRANSACTIONS;
```

```
    > LOAD DATA LOCAL INPATH '/home/acadgild/hive/transactions.txt'
    > into table TRANSACTIONS;
Loading data to table acadgilddb.transactions
OK
Time taken: 3.265 seconds
hive>
    > select * from TRANSACTIONS;
OK
97834   05/02/2018      101     965.0   Entertainment   Movie   Pune    Maharashtra     Daughter
98396   12/01/2018      102     239.0   Food    Grocery Patna   Bihar   Self
34908   06/01/2018      101     875.0   Travel  Air     Bangalore       Karnataka       Spouse
70958   17/02/2018      104     439.0   Food    Restaurant      Delhi   Delhi   Wife
9874    21/01/2018      105     509.0   Entertainment   Park    Kolkata West Bengal     NULL
94585   19/01/2018      106     629.0   Rent    House   Hyderabad       Telangana       Self
45509   20/01/2018      107     953.0   Travel  Rail    Chennai Tamil Nadu      Brother
7864    01/02/2018      108     569.0   Rent    Parking Goa     Goa     Wife
Time taken: 1.653 seconds, Fetched: 8 row(s)
hive>
    > show tables;
OK
customer
transactions
Time taken: 0.347 seconds, Fetched: 2 row(s)
hive>
```

```
hive> select count(*) from TRANSACTIONS;
```

```
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_20180218042059_ce22b56a-bab5-4b73-b65d-af88dc834b42
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1518881596510_0005, Tracking URL = http://localhost:8088/proxy/application_1518881596510_0005/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill job_1518881596510_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-02-18 04:22:01,169 Stage-1 map = 0%,  reduce = 0%
2018-02-18 04:22:30,711 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.4 sec
2018-02-18 04:23:12,113 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 15.16 sec
MapReduce Total cumulative CPU time: 16 seconds 50 msec
Ended Job = job_1518881596510_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 16.05 sec   HDFS Read: 9342 HDFS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 50 msec
OK
8
```

select a.fname, b.amount from CUSTOMER a join TRANSACTIONS b on a.custid =b.custno;

```
    > select a.fname, b.amount from CUSTOMER a join TRANSACTIONS b on a.custid=b.custno;
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_20180218042455_dfb6ec9e-1a27-4cc7-b903-de598de445c5
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/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]
2018-02-18 04:25:41     Starting to launch local task to process map join;      maximum memory = 518979584
2018-02-18 04:25:51     Dump the side-table for tag: 0 with group count: 8 into file: file:/tmp/acadgild/1ce62467-b331-4fd2-ade4-d5e8df644001/hive_2018-02-18_04-24-55_8
51_8045563776567241236-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile00--.hashtable
2018-02-18 04:25:51     Uploaded 1 File to: file:/tmp/acadgild/1ce62467-b331-4fd2-ade4-d5e8df644001/hive_2018-02-18_04-24-55_851_8045563776567241236-1/-local-10004/Hash
Table-Stage-3/MapJoin-mapfile00--.hashtable (469 bytes)
2018-02-18 04:25:51     End of local task; Time Taken: 9.864 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
```

```
SELECT CATEGORY, SUM(amount) FROM TRANSACTIONS group by category;

SELECT CATEGORY, SUM(amount) FROM TRANSACTIONS GROUP BY category
SORT BY category DESC;
```

```
    > select a.fname, b.amount from CUSTOMER a join TRANSACTIONS b on a.custid=b.custno;
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_20180218042455_dfb6ec9e-1a27-4cc7-b903-de598de445c5
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/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]
2018-02-18 04:25:41     Starting to launch local task to process map join;      maximum memory = 518979584
2018-02-18 04:25:51     Dump the side-table for tag: 0 with group count: 8 into file: file:/tmp/acadgild/1ce62467-b331-4fd2-ade4-d5e8df644001/hive_2018-02-18_04-24-55_8
51_8045563776567241236-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile00--.hashtable
2018-02-18 04:25:51     Uploaded 1 File to: file:/tmp/acadgild/1ce62467-b331-4fd2-ade4-d5e8df644001/hive_2018-02-18_04-24-55_851_8045563776567241236-1/-local-10004/Hash
Table-Stage-3/MapJoin-mapfile00--.hashtable (469 bytes)
2018-02-18 04:25:51     End of local task; Time Taken: 9.864 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1518881596510_0006, Tracking URL = http://localhost:8088/proxy/application_1518881596510_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill job_1518881596510_0006
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2018-02-18 04:26:56,367 Stage-3 map = 0%,  reduce = 0%
2018-02-18 04:27:42,467 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 9.93 sec
MapReduce Total cumulative CPU time: 9 seconds 930 msec
Ended Job = job_1518881596510_0006
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 9.93 sec   HDFS Read: 7454 HDFS Write: 289 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 930 msec
OK
Amitabh 965.0
Sharukh 239.0
Amitabh 875.0
Anubahv 439.0
Pawan   509.0
Aamir   629.0
Salman  953.0
Ranbir  569.0
Time taken: 169.035 seconds, Fetched: 8 row(s)
hive>
```

1. **Find out the number of transaction done by each customer (These should be take up in module 8 itself)**

```
hive> describe TRANSACTIONS;
OK
txnno                   int
txndate                 string
custno                  int
amount                  double
category                string
product                 string
city                    string
state                   string
spendby                 string
Time taken: 0.537 seconds, Fetched: 9 row(s)
hive> describe CUSTOMER
    >
    > ;
OK
custid                  int
fname                   string
lname                   string
age                     int
profession              string
Time taken: 0.586 seconds, Fetched: 5 row(s)
hive>
```

```
hive> select custno,count(*) as totaltrx from TRANSACTIONS
    > group by custno;
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_20180218043029_e401e52f-2f0b-4ea4-a93c-7e092d6ce37d
Total jobs = 1
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_1518881596510_0007, Tracking URL = http://localhost:8088/proxy/application_1518881596510_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill job_1518881596510_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-02-18 04:31:33,425 Stage-1 map = 0%,  reduce = 0%
2018-02-18 04:32:09,464 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.22 sec
2018-02-18 04:32:52,219 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 16.44 sec
2018-02-18 04:32:55,941 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 18.7 sec
MapReduce Total cumulative CPU time: 18 seconds 700 msec
Ended Job = job_1518881596510_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 18.7 sec   HDFS Read: 9719 HDFS Write: 213 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 700 msec
OK
101     2
102     1
104     1
105     1
106     1
107     1
108     1
```

2. Create a new table called TRANSACTIONS_COUNT. This table should have 3 fields - custid, fname and count. (Again to be done in module 8)

```
hive> create table TRANSACTIONS_COUNT(
    > fname string,
    > count int);
OK
Time taken: 0.524 seconds
hive>
```

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. (This has to be done in module 9).

```
hive> insert into TRANSACTIONS_COUNT
    > select a.fname, count(1) as count from CUSTOMER a join TRANSACTIONS b on a.custid=b.custno
    > group by a.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_20180218043419_f55e2a8e-5f18-4780-8ebd-71760c517333
Total jobs = 1
```

4. Now lets make the TRANSACTIONS_COUNT table Hbase complaint. In the sence, use Ser Des And Storate handler features of hive to change the TRANSACTIONS_COUNT table to be able to create a TRANSACTIONS table in Hbase. (This has to be done in module 10)

```
describe TRANSACTIONS_COUNT;

CREATE TABLE TRANSACTIONS_HBase(username STRING, count INT)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = ':key,stats:count');

--in hbase
describe 'acadgilddb.transactions_hbase'

insert into TRANSACTIONS_HBase
select * from TRANSACTIONS_COUNT;
```

```
hive> use acadgilddb;
```

```
OK
Time taken: 30.682 seconds
hive>
    > describe TRANSACTIONS_COUNT;
```

```
hive>
    > describe TRANSACTIONS_COUNT;
OK
fname                   string
count                   int
Time taken: 2.232 seconds, Fetched: 2 row(s)
hive>
```

```
hive> CREATE TABLE TRANSACTIONS_HBase(username STRING, count INT)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ('hbase.columns.mapping' = ':key,stats:count');
```

```
hive> show tables;
OK
customer
t
transactions
transactions_count
transactions_hbase
Time taken: 0.656 seconds, Fetched: 5 row(s)
```

```
hbase(main):001:0> describe 'acadgilddb.transactions_hbase'
Table acadgilddb.transactions_hbase is ENABLED
acadgilddb.transactions_hbase
COLUMN FAMILIES DESCRIPTION
{NAME => 'stats', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_E
NCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536
', REPLICATION_SCOPE => '0'}
1 row(s) in 3.0080 seconds

hbase(main):002:0>
```

5. Now insert the data in TRANSACTIONS_COUNT table using the query in step 3 again, this should populate the Hbase TRANSACTIONS table automatically (This has to be done in module 10)

```
hive> insert into TRANSACTIONS_HBase
    > select * from TRANSACTIONS_COUNT;
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_20180224112253_3ae0b1a9-5431-4952-ab1c-f685996a1dfe
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_1519437016588_0004, Tracking URL = http://localhost:8088/prox
y/application_1519437016588_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job  -kill
job_1519437016588_0004
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2018-02-24 11:24:05,027 Stage-3 map = 0%,  reduce = 0%
```

```
2018-02-24 11:24:36,559 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 9.05 se
c
MapReduce Total cumulative CPU time: 9 seconds 50 msec
Ended Job = job_1519437016588_0004
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 9.05 sec   HDFS Read: 10809 HDFS Write:
0 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 50 msec
OK
Time taken: 104.903 seconds
hive>
```

6. Now from the Hbase level, write the Hbase java API code to access and scan the TRANSACTIONS table data from java level.

```
hbase(main):002:0> scan 'acadgilddb.transactions_hbase'
ROW                           COLUMN+CELL
 Aamir                        column=stats:count, timestamp=1519451674423, value=2
 Amitabh                      column=stats:count, timestamp=1519451674423, value=2
 Anubahv                      column=stats:count, timestamp=1519451674423, value=2
 Pawan                        column=stats:count, timestamp=1519451674423, value=2
 Ranbir                       column=stats:count, timestamp=1519451674423, value=2
 Salman                       column=stats:count, timestamp=1519451674423, value=2
 Sharukh                      column=stats:count, timestamp=1519451674423, value=2
7 row(s) in 0.9030 seconds

hbase(main):003:0>
```