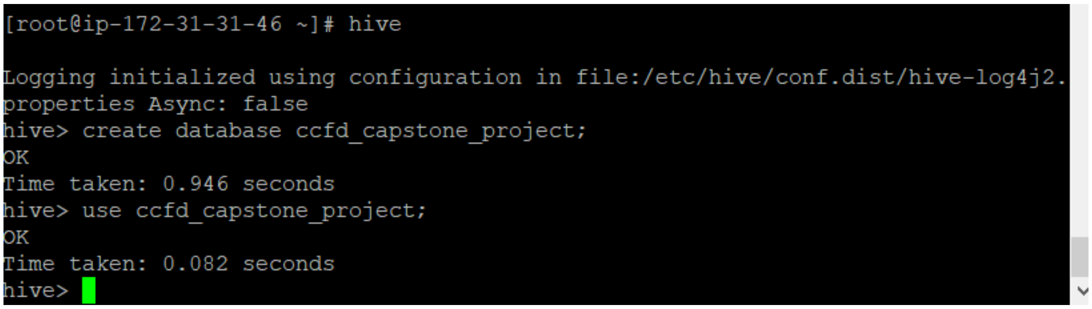
Scripts Execution

- ## Screenshots of the execution of the scripts written

a) Load the transactions history data (card_transactions.csv) in a NoSQL database.

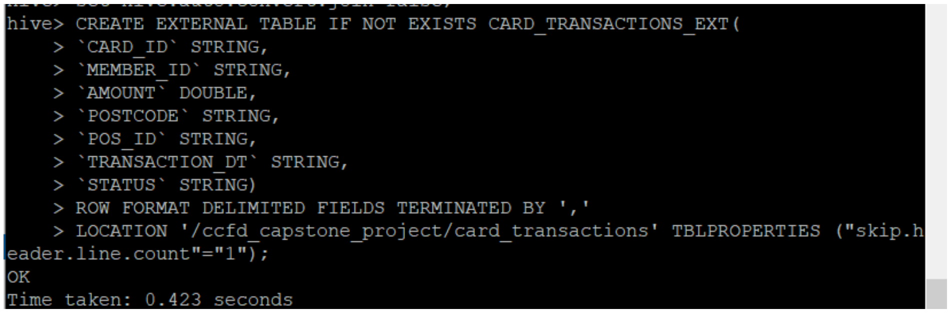
---------- Hive Operations: Starts Here ----------

1. Start hive and create new database named ccfd_capstone_project -> switch to



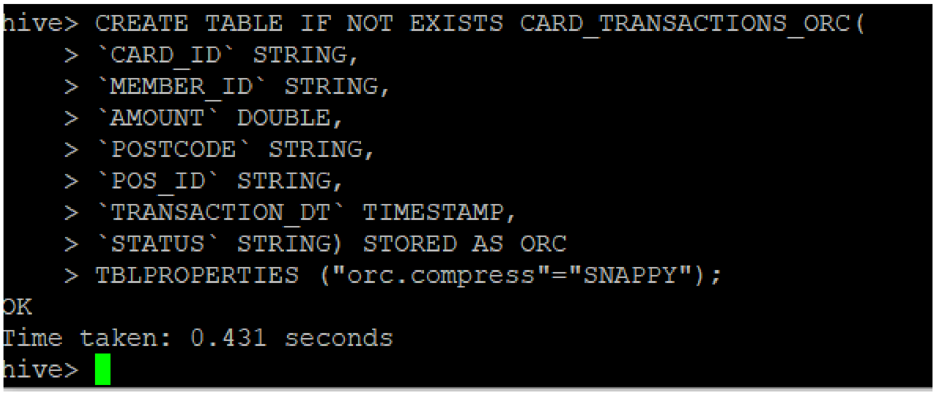
```
[root@ip-172-31-31-46 ~]# hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.
properties Async: false
hive> create database ccfd_capstone_project;
OK
Time taken: 0.946 seconds
hive> use ccfd_capstone_project;
OK
Time taken: 0.082 seconds
hive>
```

2. Create an external table “**card_transactions_ext**”



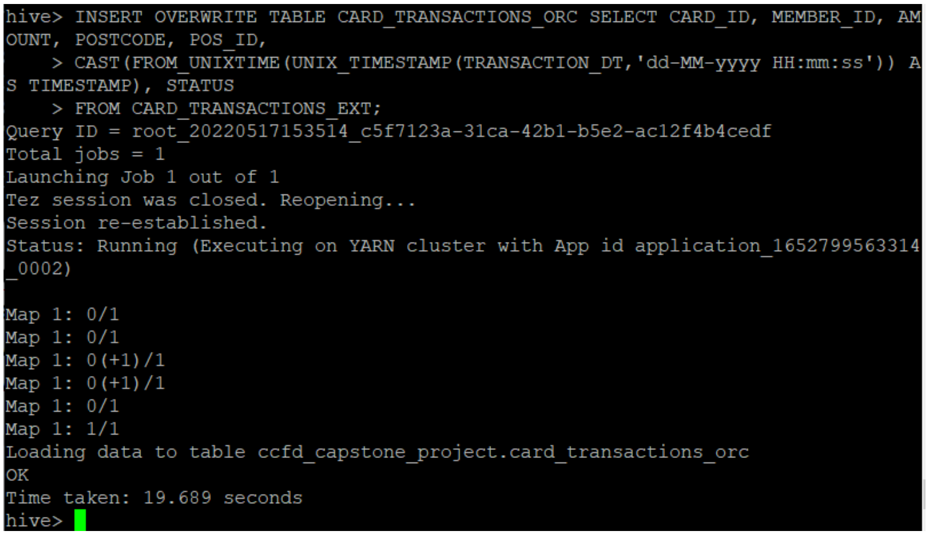
```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CARD_TRANSACTIONS_EXT(
    > `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` STRING,
    > `STATUS` STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/ccfd_capstone_project/card_transactions' TBLPROPERTIES ("skip.h
eader.line.count"="1");
OK
Time taken: 0.423 seconds
```

3. Create table "**card_transactions_orc**" in ORC format for better performance.



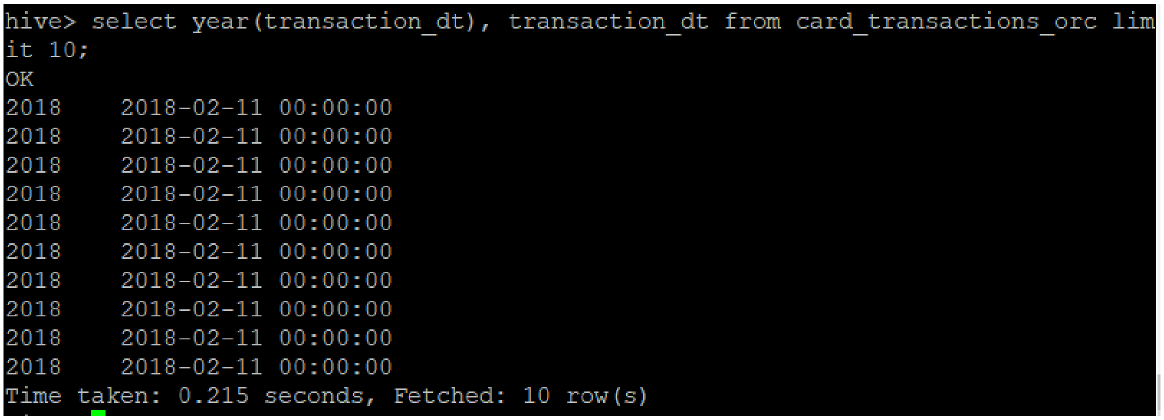
```
hive> CREATE TABLE IF NOT EXISTS CARD_TRANSACTIONS_ORC(
    > `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` TIMESTAMP,
    > `STATUS` STRING) STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.431 seconds
hive>
```

4. Load data in “**card_transactions_orc**” table and type cast **transaction_dt** column in timestamp format



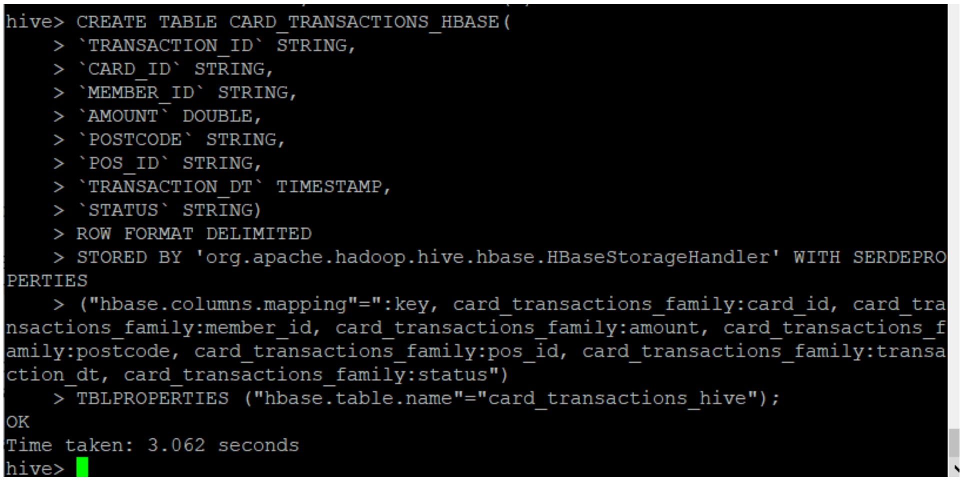
```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC SELECT CARD_ID, MEMBER_ID, AM
OUNT, POSTCODE, POS_ID,
    > CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTION_DT,'dd-MM-yyyy HH:mm:ss')) A
S TIMESTAMP), STATUS
    > FROM CARD_TRANSACTIONS_EXT;
Query ID = root_20220517153514_c5f7123a-31ca-42b1-b5e2-ac12f4b4cedf
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1652799563314
_0002)

Map 1: 0/1
Map 1: 0/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 0/1
Map 1: 1/1
Loading data to table ccfd_capstone_project.card_transactions_orc
OK
Time taken: 19.689 seconds
hive>
```

5. Verify **transaction_dt** and **year** columns in "**card_transactions_orc**" table.



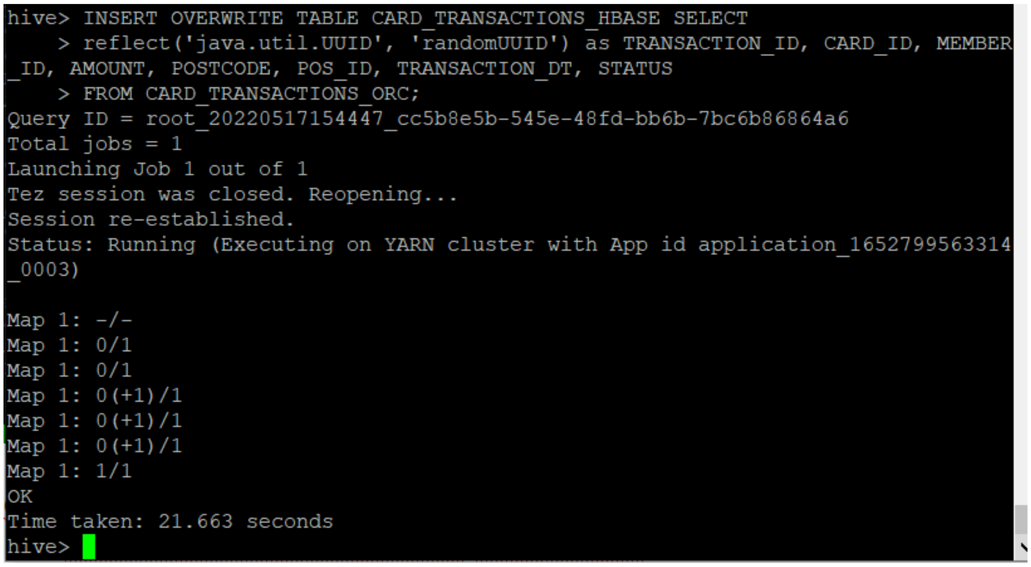
```
hive> select year(transaction_dt), transaction_dt from card_transactions_orc lim
it 10;
OK
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
Time taken: 0.215 seconds, Fetched: 10 row(s)
```

6. Create hive-hbase integrated table which will be visible in HBase as well. "**card_transactions_hbase**" table



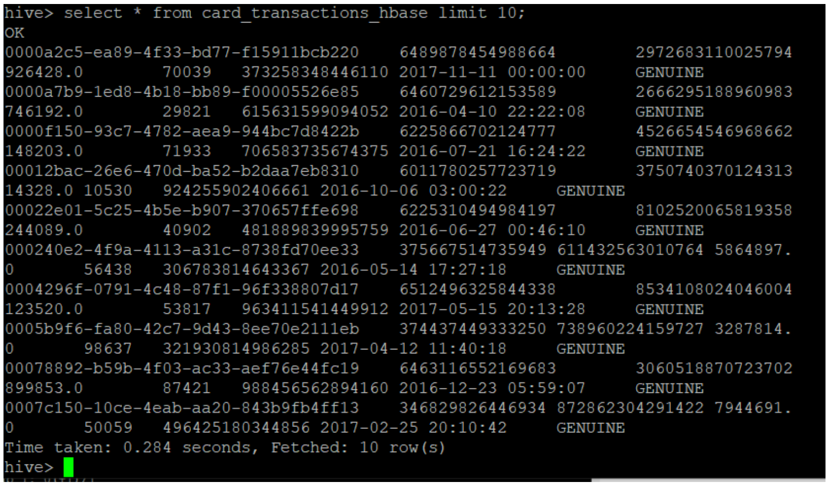
```
hive> CREATE TABLE CARD_TRANSACTIONS_HBASE(
    > `TRANSACTION_ID` STRING,
    > `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` TIMESTAMP,
    > `STATUS` STRING)
    > ROW FORMAT DELIMITED
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH SERDEPRO
PERTIES
    > ("hbase.columns.mapping"=":key, card_transactions_family:card_id, card_tra
nsactions_family:member_id, card_transactions_family:amount, card_transactions_f
amily:postcode, card_transactions_family:pos_id, card_transactions_family:transa
ction_dt, card_transactions_family:status")
    > TBLPROPERTIES ("hbase.table.name"="card_transactions_hive");
OK
Time taken: 3.062 seconds
hive>
```

7. Load data in "**card_transactions_hbase**" table which will be visible in HBase as well with table name as "**card_transactions_hive**".Using randomUUID to populate TRANSACTION_ID field (row key).



```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_HBASE SELECT
    > reflect('java.util.UUID', 'randomUUID') as TRANSACTION_ID, CARD_ID, MEMBER
_ID, AMOUNT, POSTCODE, POS_ID, TRANSACTION_DT, STATUS
    > FROM CARD_TRANSACTIONS_ORC;
Query ID = root_20220517154447_cc5b8e5b-545e-48fd-bb6b-7bc6b86864a6
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1652799563314
_0003)

Map 1: -/-
Map 1: 0/1
Map 1: 0/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 1/1
OK
Time taken: 21.663 seconds
hive>
```

8. Verify data in "**card_transactions_hbase**” table.

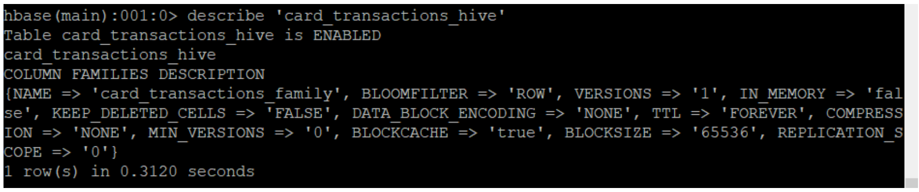


```
hive> select * from card_transactions_hbase limit 10;
OK
0000a2c5-ea89-4f33-bd77-f15911bcb220    6489878454988664        2972683110025794
926428.0        70039   373258348446110 2017-11-11 00:00:00     GENUINE
0000a7b9-1ed8-4b18-bb89-f00005526e85    6460729612153589        2666295188960983
746192.0        29821   615631599094052 2016-04-10 22:22:08     GENUINE
0000f150-93c7-4782-aea9-944bc7d8422b    6225866702124777        4526654546968662
148203.0        71933   706583735674375 2016-07-21 16:24:22     GENUINE
00012bac-26e6-470d-ba52-b2daa7eb8310    6011780257723719        3750740370124313
14328.0 10530   924255902406661 2016-10-06 03:00:22     GENUINE
00022e01-5c25-4b5e-b907-370657ffe698    6225310494984197        8102520065819358
244089.0        40902   481889839995759 2016-06-27 00:46:10     GENUINE
000240e2-4f9a-4113-a31c-8738fd70ee33    375667514735949 611432563010764 5864897.
0       56438   306783814643367 2016-05-14 17:27:18     GENUINE
0004296f-0791-4c48-87f1-96f338807d17    6512496325844338        8534108024046004
123520.0        53817   963411541449912 2017-05-15 20:13:28     GENUINE
0005b9f6-fa80-42c7-9d43-8ee70e2111eb    374437449333250 738960224159727 3287814.
0       98637   321930814986285 2017-04-12 11:40:18     GENUINE
00078892-b59b-4f03-ac33-aef76e44fc19    6463116552169683        3060518870723702
899853.0        87421   988456562894160 2016-12-23 05:59:07     GENUINE
0007c150-10ce-4eab-aa20-843b9fb4ff13    346829826446934 872862304291422 7944691.
0       50059   496425180344856 2017-02-25 20:10:42     GENUINE
Time taken: 0.284 seconds, Fetched: 10 row(s)
hive>
```

---------- Hive Operations: Ends Here ----------

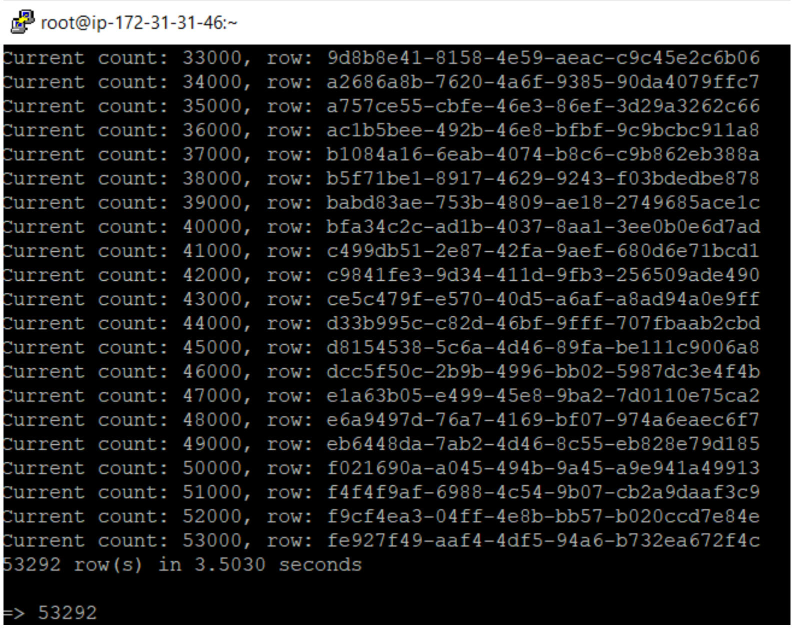
---------- HBase Operations: Starts Here ----------

1. Start HBase and verify details of "**card_transactions_hive**" table (hive-hbase integrated table).



```
hbase(main):001:0> describe 'card_transactions_hive'
Table card_transactions_hive is ENABLED
card_transactions_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'card_transactions_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'fal
se', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESS
ION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_S
COPE => '0'}
1 row(s) in 0.3120 seconds
```

2. Verify count of "**card_transactions_hive**" table



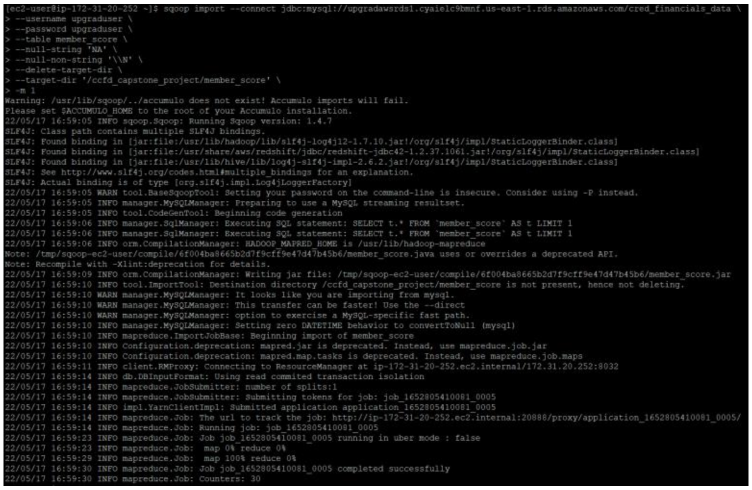
```
Current count: 33000, row: 9d8b8e41-8158-4e59-aeac-c9c45e2c6b06
Current count: 34000, row: a2686a8b-7620-4a6f-9385-90da4079ffc7
Current count: 35000, row: a757ce55-cbfe-46e3-86ef-3d29a3262c66
Current count: 36000, row: ac1b5bee-492b-46e8-bfbf-9c9bcbc911a8
Current count: 37000, row: b1084a16-6eab-4074-b8c6-c9b862eb388a
Current count: 38000, row: b5f71be1-8917-4629-9243-f03bdedbe878
Current count: 39000, row: babd83ae-753b-4809-ae18-2749685ace1c
Current count: 40000, row: bfa34c2c-ad1b-4037-8aa1-3ee0b0e6d7ad
Current count: 41000, row: c499db51-2e87-42fa-9aef-680d6e71bcd1
Current count: 42000, row: c9841fe3-9d34-411d-9fb3-256509ade490
Current count: 43000, row: ce5c479f-e570-40d5-a6af-a8ad94a0e9ff
Current count: 44000, row: d33b995c-c82d-46bf-9fff-707fbaab2cbd
Current count: 45000, row: d8154538-5c6a-4d46-89fa-be111c9006a8
Current count: 46000, row: dcc5f50c-2b9b-4996-bb02-5987dc3e4f4b
Current count: 47000, row: e1a63b05-e499-45e8-9ba2-7d0110e75ca2
Current count: 48000, row: e6a9497d-76a7-4169-bf07-974a6eaec6f7
Current count: 49000, row: eb6448da-7ab2-4d46-8c55-eb828e79d185
Current count: 50000, row: f021690a-a045-494b-9a45-a9e941a49913
Current count: 51000, row: f4f4f9af-6988-4c54-9b07-cb2a9daaf3c9
Current count: 52000, row: f9cf4ea3-04ff-4e8b-bb57-b020ccd7e84e
Current count: 53000, row: fe927f49-aaf4-4df5-94a6-b732ea672f4c
53292 row(s) in 3.5030 seconds

=> 53292
```

Total number for record is **53292** which is matching with given requirement.

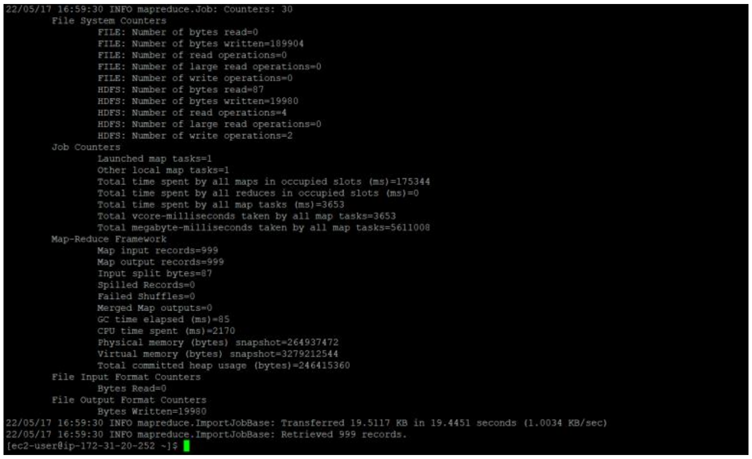
---------- HBase Operations: Ends Here ----------

b) Task 2: Ingest the relevant data from AWS RDS to Hadoop.

- Run Sqoop command to import "**member_score**" table from RDS to HDFS.

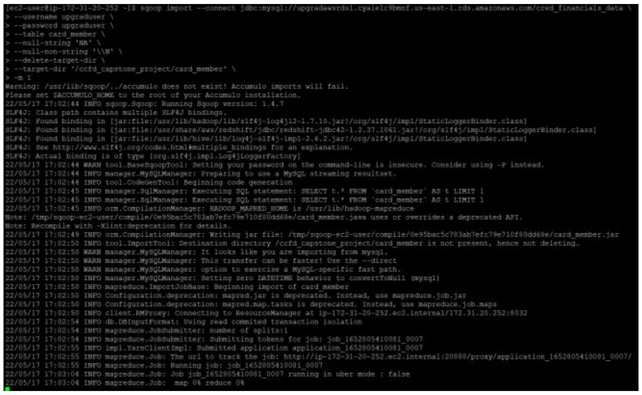


```
[ec2-user@ip-172-31-20-252 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table member_score \
> --null-string 'NA' \
> --null-non-string '\\N' \
> --delete-target-dir \
> --target-dir '/ccfd_capstone_project/member_score' \
> -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/05/17 16:59:05 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.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.slf4j.impl.Log4jLoggerFactory]
22/05/17 16:59:05 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/17 16:59:05 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/17 16:59:05 INFO tool.CodeGenTool: Beginning code generation
22/05/17 16:59:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/17 16:59:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/17 16:59:06 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/6f004ba8665b2d7f9cff9e47d47b45b6/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/17 16:59:09 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/6f004ba8665b2d7f9cff9e47d47b45b6/member_score.jar
22/05/17 16:59:10 INFO tool.ImportTool: Destination directory /ccfd_capstone_project/member_score is not present, hence not deleting.
22/05/17 16:59:10 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/17 16:59:10 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/17 16:59:10 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/17 16:59:10 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/17 16:59:10 INFO mapreduce.ImportJobBase: Beginning import of member_score
22/05/17 16:59:10 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/17 16:59:10 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/17 16:59:11 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-20-252.ec2.internal/172.31.20.252:8032
22/05/17 16:59:14 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/17 16:59:14 INFO mapreduce.JobSubmitter: number of splits:1
22/05/17 16:59:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652805410081_0005
22/05/17 16:59:14 INFO impl.YarnClientImpl: Submitted application application_1652805410081_0005
22/05/17 16:59:14 INFO mapreduce.Job: The url to track the job: http://ip-172-31-20-252.ec2.internal:20888/proxy/application_1652805410081_0005/
22/05/17 16:59:14 INFO mapreduce.Job: Running job: job_1652805410081_0005
22/05/17 16:59:23 INFO mapreduce.Job: Job job_1652805410081_0005 running in uber mode : false
22/05/17 16:59:23 INFO mapreduce.Job:  map 0% reduce 0%
22/05/17 16:59:29 INFO mapreduce.Job:  map 100% reduce 0%
22/05/17 16:59:30 INFO mapreduce.Job: Job job_1652805410081_0005 completed successfully
22/05/17 16:59:30 INFO mapreduce.Job: Counters: 30
```

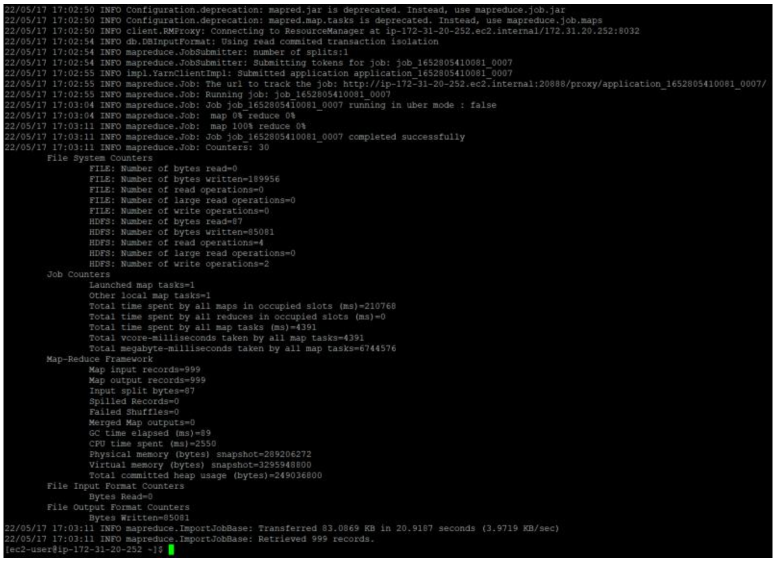


```
22/05/17 16:59:30 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189904
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=19980
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=175344
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3653
                Total vcore-milliseconds taken by all map tasks=3653
                Total megabyte-milliseconds taken by all map tasks=5611008
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=85
                CPU time spent (ms)=2170
                Physical memory (bytes) snapshot=264937472
                Virtual memory (bytes) snapshot=3279212544
                Total committed heap usage (bytes)=246415360
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
22/05/17 16:59:30 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 19.4451 seconds (1.0034 KB/sec)
22/05/17 16:59:30 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-20-252 ~]$
```

- Run Sqoop command to import “**card_member**” table from RDS to HDFS.

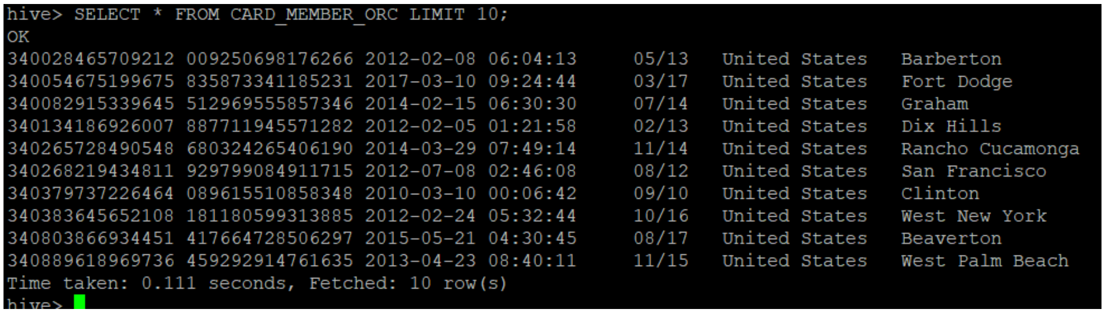


```
[ec2-user@ip-172-31-20-252 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table card_member \
> --null-string 'NA' \
> --null-non-string '\\N' \
> --delete-target-dir \
> --target-dir '/ccfd_capstone_project/card_member' \
> -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/05/17 17:02:44 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.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.slf4j.impl.Log4jLoggerFactory]
22/05/17 17:02:44 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/17 17:02:44 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/17 17:02:44 INFO tool.CodeGenTool: Beginning code generation
22/05/17 17:02:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/17 17:02:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/17 17:02:45 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/0e95bac5c783ab7efc79e710f80dd68e/card_member.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/17 17:02:49 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/0e95bac5c783ab7efc79e710f80dd68e/card_member.jar
22/05/17 17:02:50 INFO tool.ImportTool: Destination directory /ccfd_capstone_project/card_member is not present, hence not deleting.
22/05/17 17:02:50 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/17 17:02:50 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/17 17:02:50 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/17 17:02:50 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/17 17:02:50 INFO mapreduce.ImportJobBase: Beginning import of card_member
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/17 17:02:50 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-20-252.ec2.internal/172.31.20.252:8032
22/05/17 17:02:54 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: number of splits:1
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652805410081_0007
22/05/17 17:02:55 INFO impl.YarnClientImpl: Submitted application application_1652805410081_0007
22/05/17 17:02:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-20-252.ec2.internal:20888/proxy/application_1652805410081_0007/
22/05/17 17:02:55 INFO mapreduce.Job: Running job: job_1652805410081_0007
22/05/17 17:03:04 INFO mapreduce.Job: Job job_1652805410081_0007 running in uber mode : false
22/05/17 17:03:04 INFO mapreduce.Job:  map 0% reduce 0%
```



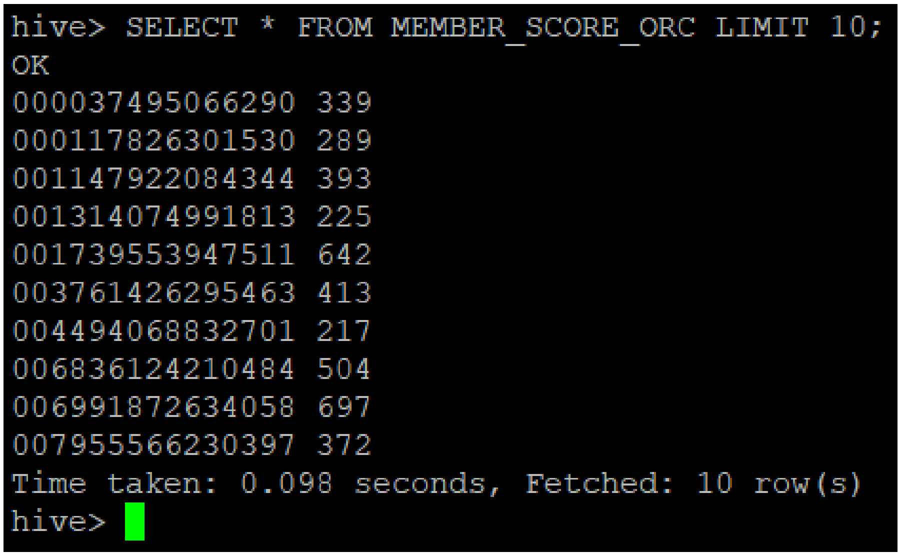
```
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/17 17:02:50 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-20-252.ec2.internal/172.31.20.252:8032
22/05/17 17:02:54 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: number of splits:1
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652805410081_0007
22/05/17 17:02:55 INFO impl.YarnClientImpl: Submitted application application_1652805410081_0007
22/05/17 17:02:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-20-252.ec2.internal:20888/proxy/application_1652805410081_0007/
22/05/17 17:02:55 INFO mapreduce.Job: Running job: job_1652805410081_0007
22/05/17 17:03:04 INFO mapreduce.Job: Job job_1652805410081_0007 running in uber mode : false
22/05/17 17:03:04 INFO mapreduce.Job:  map 0% reduce 0%
22/05/17 17:03:11 INFO mapreduce.Job:  map 100% reduce 0%
22/05/17 17:03:11 INFO mapreduce.Job: Job job_1652805410081_0007 completed successfully
22/05/17 17:03:11 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189956
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=85081
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=210768
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4391
                Total vcore-milliseconds taken by all map tasks=4391
                Total megabyte-milliseconds taken by all map tasks=6744576
        Map-Reduce Framework
                Map input records=999
                Map output records=999
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=89
                CPU time spent (ms)=2550
                Physical memory (bytes) snapshot=289206272
                Virtual memory (bytes) snapshot=3295948800
                Total committed heap usage (bytes)=249036800
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85081
22/05/17 17:03:11 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 20.9187 seconds (3.9719 KB/sec)
22/05/17 17:03:11 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-20-252 ~]$
```

- Verify data in “**card_member_orc**" table.



```
hive> SELECT * FROM CARD_MEMBER_ORC LIMIT 10;
OK
340028465709212 009250698176266 2012-02-08 06:04:13     05/13   United States   Barberton
340054675199675 835873341185231 2017-03-10 09:24:44     03/17   United States   Fort Dodge
340082915339645 512969555857346 2014-02-15 06:30:30     07/14   United States   Graham
340134186926007 887711945571282 2012-02-05 01:21:58     02/13   United States   Dix Hills
340265728490548 680324265406190 2014-03-29 07:49:14     11/14   United States   Rancho Cucamonga
340268219434811 929799084911715 2012-07-08 02:46:08     08/12   United States   San Francisco
340379737226464 089615510858348 2010-03-10 00:06:42     09/10   United States   Clinton
340383645652108 181180599313885 2012-02-24 05:32:44     10/16   United States   West New York
340803866934451 417664728506297 2015-05-21 04:30:45     08/17   United States   Beaverton
340889618969736 459292914761635 2013-04-23 08:40:11     11/15   United States   West Palm Beach
Time taken: 0.111 seconds, Fetched: 10 row(s)
hive>
```

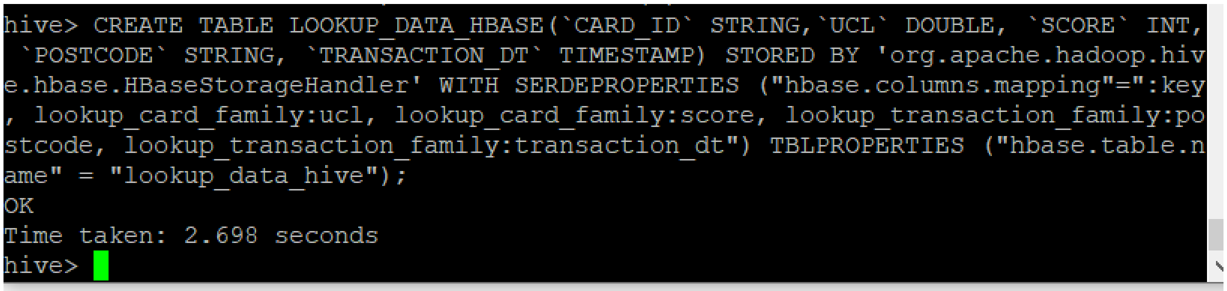
- Verify data in "member_score_orc" table.



```
hive> SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;
OK
000037495066290 339
000117826301530 289
001147922084344 393
001314074991813 225
001739553947511 642
003761426295463 413
004494068832701 217
006836124210484 504
006991872634058 697
007955566230397 372
Time taken: 0.098 seconds, Fetched: 10 row(s)
hive>
```

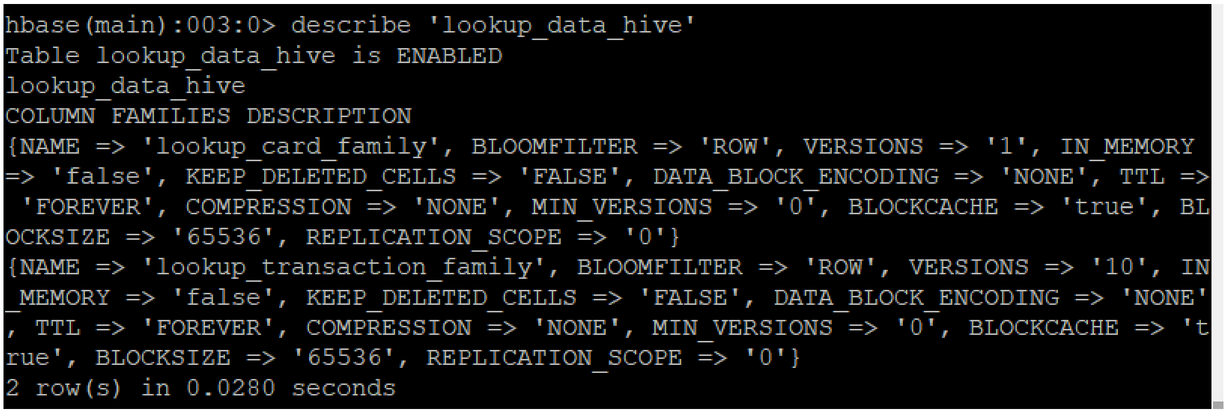
c) Task 3: Create a look-up table with columns specified earlier in the problem statement.

- Create “lookup_data_hbase” table (hive-hbase integrated table) which will be visible in HBase (lookup_data_hive).



```
hive> CREATE TABLE LOOKUP_DATA_HBASE(`CARD_ID` STRING,`UCL` DOUBLE, `SCORE` INT,
 `POSTCODE` STRING, `TRANSACTION_DT` TIMESTAMP) STORED BY 'org.apache.hadoop.hiv
e.hbase.HBaseStorageHandler' WITH SERDEPROPERTIES ("hbase.columns.mapping"=":key
, lookup_card_family:ucl, lookup_card_family:score, lookup_transaction_family:po
stcode, lookup_transaction_family:transaction_dt") TBLPROPERTIES ("hbase.table.n
ame" = "lookup_data_hive");
OK
Time taken: 2.698 seconds
hive> 
```

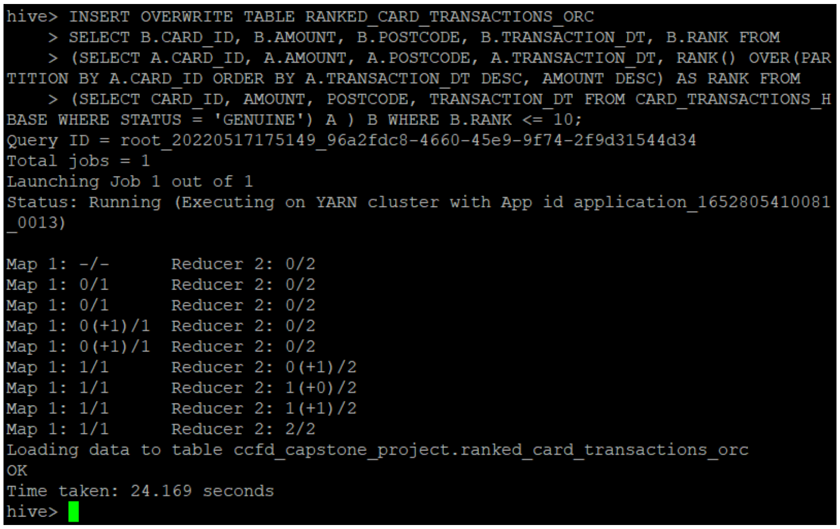
- Verify details of **lookup_data_hive** (hive-hbase integrated) table :



```
hbase(main):003:0> describe 'lookup_data_hive'
Table lookup_data_hive is ENABLED
lookup_data_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'lookup_card_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY
=> 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL =>
 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BL
OCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'lookup_transaction_family', BLOOMFILTER => 'ROW', VERSIONS => '10', IN
_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE'
, TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 't
rue', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
2 row(s) in 0.0280 seconds
```

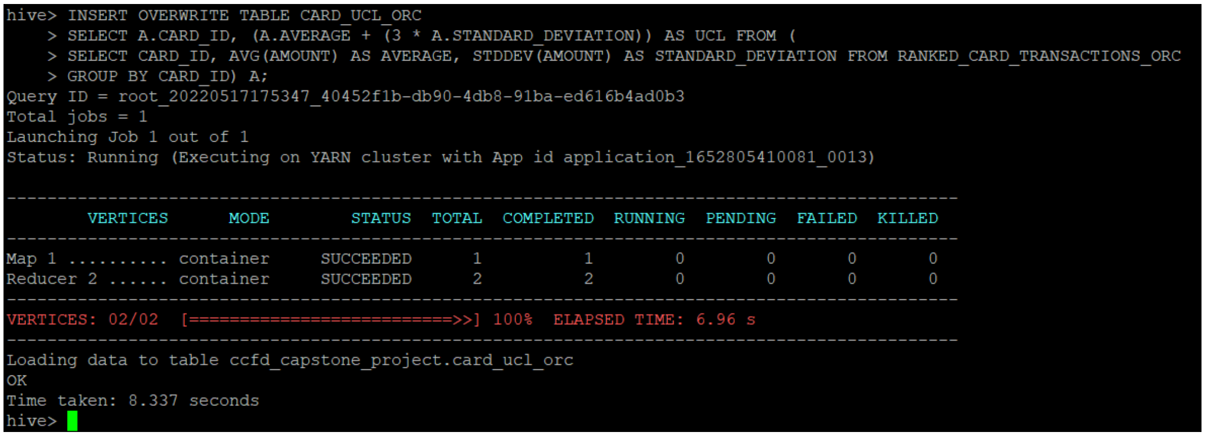
d) Task 4: After creating the table, you need to load the relevant data in the lookup table.

- Load data in “**ranked_card_transactions_orc**” table



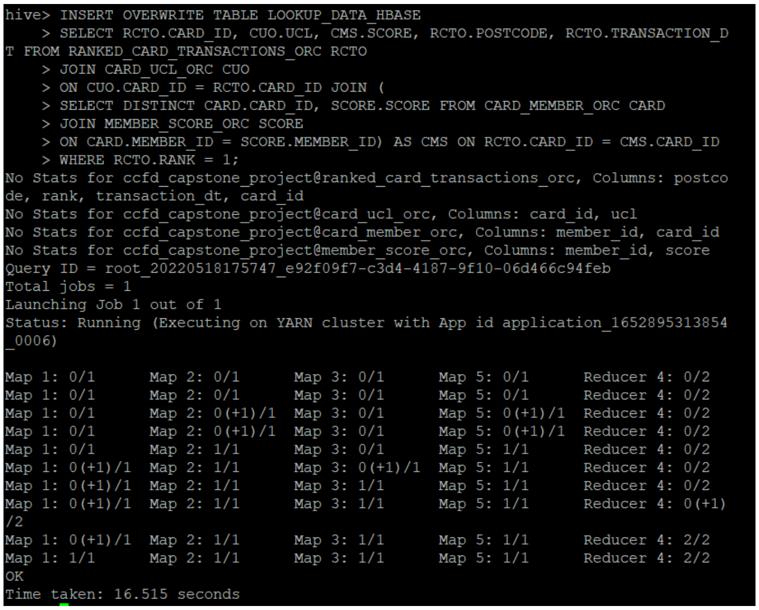
```
hive> INSERT OVERWRITE TABLE RANKED_CARD_TRANSACTIONS_ORC
    > SELECT B.CARD_ID, B.AMOUNT, B.POSTCODE, B.TRANSACTION_DT, B.RANK FROM
    > (SELECT A.CARD_ID, A.AMOUNT, A.POSTCODE, A.TRANSACTION_DT, RANK() OVER(PAR
TITION BY A.CARD_ID ORDER BY A.TRANSACTION_DT DESC, AMOUNT DESC) AS RANK FROM
    > (SELECT CARD_ID, AMOUNT, POSTCODE, TRANSACTION_DT FROM CARD_TRANSACTIONS_H
BASE WHERE STATUS = 'GENUINE') A ) B WHERE B.RANK <= 10;
Query ID = root_20220517175149_96a2fdc8-4660-45e9-9f74-2f9d31544d34
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652805410081
_0013)

Map 1: -/-      Reducer 2: 0/2
Map 1: 0/1      Reducer 2: 0/2
Map 1: 0/1      Reducer 2: 0/2
Map 1: 0(+1)/1  Reducer 2: 0/2
Map 1: 0(+1)/1  Reducer 2: 0/2
Map 1: 1/1      Reducer 2: 0(+1)/2
Map 1: 1/1      Reducer 2: 1(+0)/2
Map 1: 1/1      Reducer 2: 1(+1)/2
Map 1: 1/1      Reducer 2: 2/2
Loading data to table ccfd_capstone_project.ranked_card_transactions_orc
OK
Time taken: 24.169 seconds
hive> 
```

- Load data in "**card_ucl_orc**" table



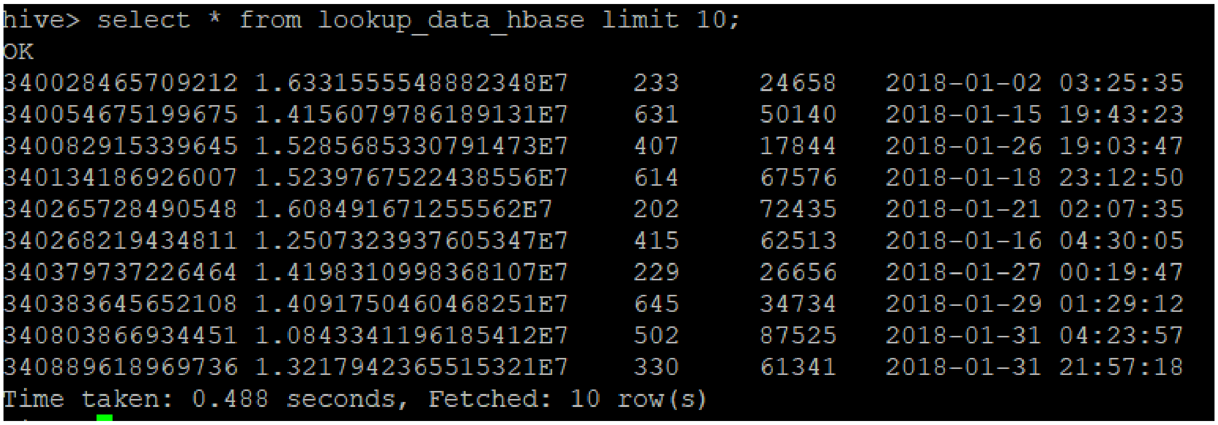
```
hive> INSERT OVERWRITE TABLE CARD_UCL_ORC
    > SELECT A.CARD_ID, (A.AVERAGE + (3 * A.STANDARD_DEVIATION)) AS UCL FROM (
    > SELECT CARD_ID, AVG(AMOUNT) AS AVERAGE, STDDEV(AMOUNT) AS STANDARD_DEVIATION FROM RANKED_CARD_TRANSACTIONS_ORC
    > GROUP BY CARD_ID) A;
Query ID = root_20220517175347_40452f1b-db90-4db8-91ba-ed616b4ad0b3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652805410081_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.96 s
----------------------------------------------------------------------------------------------
Loading data to table ccfd_capstone_project.card_ucl_orc
OK
Time taken: 8.337 seconds
hive>
```

- Load data in **lookup_data_hbase** table.



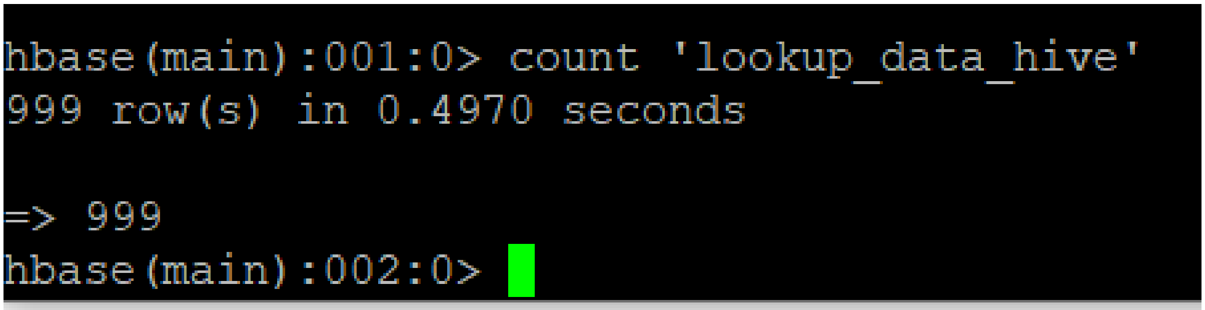
```
hive> INSERT OVERWRITE TABLE LOOKUP_DATA_HBASE
    > SELECT RCTO.CARD_ID, CUO.UCL, CMS.SCORE, RCTO.POSTCODE, RCTO.TRANSACTION_D
T FROM RANKED_CARD_TRANSACTIONS_ORC RCTO
    > JOIN CARD_UCL_ORC CUO
    > ON CUO.CARD_ID = RCTO.CARD_ID JOIN (
    > SELECT DISTINCT CARD.CARD_ID, SCORE.SCORE FROM CARD_MEMBER_ORC CARD
    > JOIN MEMBER_SCORE_ORC SCORE
    > ON CARD.MEMBER_ID = SCORE.MEMBER_ID) AS CMS ON RCTO.CARD_ID = CMS.CARD_ID
    > WHERE RCTO.RANK = 1;
No Stats for ccfd_capstone_project@ranked_card_transactions_orc, Columns: postco
de, rank, transaction_dt, card_id
No Stats for ccfd_capstone_project@card_ucl_orc, Columns: card_id, ucl
No Stats for ccfd_capstone_project@card_member_orc, Columns: member_id, card_id
No Stats for ccfd_capstone_project@member_score_orc, Columns: member_id, score
Query ID = root_20220518175747_e92f09f7-c3d4-4187-9f10-06d466c94feb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652895313854
_0006)

Map 1: 0/1      Map 2: 0/1      Map 3: 0/1      Map 5: 0/1      Reducer 4: 0/2
Map 1: 0/1      Map 2: 0/1      Map 3: 0/1      Map 5: 0/1      Reducer 4: 0/2
Map 1: 0/1      Map 2: 0(+1)/1  Map 3: 0/1      Map 5: 0(+1)/1  Reducer 4: 0/2
Map 1: 0/1      Map 2: 0(+1)/1  Map 3: 0/1      Map 5: 0(+1)/1  Reducer 4: 0/2
Map 1: 0/1      Map 2: 1/1      Map 3: 0/1      Map 5: 1/1      Reducer 4: 0/2
Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 0(+1)/1  Map 5: 1/1      Reducer 4: 0/2
Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 0/2
Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 0(+1)
/2
Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 2/2
Map 1: 1/1      Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 2/2
OK
Time taken: 16.515 seconds
```

- Verify some data in “**lookup_data_hbase**” table.



```
hive> select * from lookup_data_hbase limit 10;
OK
340028465709212 1.6331555548882348E7    233     24658   2018-01-02 03:25:35
340054675199675 1.4156079786189131E7    631     50140   2018-01-15 19:43:23
340082915339645 1.5285685330791473E7    407     17844   2018-01-26 19:03:47
340134186926007 1.5239767522438556E7    614     67576   2018-01-18 23:12:50
340265728490548 1.608491671255562E7     202     72435   2018-01-21 02:07:35
340268219434811 1.2507323937605347E7    415     62513   2018-01-16 04:30:05
340379737226464 1.4198310998368107E7    229     26656   2018-01-27 00:19:47
340383645652108 1.4091750460468251E7    645     34734   2018-01-29 01:29:12
340803866934451 1.0843341196185412E7    502     87525   2018-01-31 04:23:57
340889618969736 1.3217942365515321E7    330     61341   2018-01-31 21:57:18
Time taken: 0.488 seconds, Fetched: 10 row(s)
```

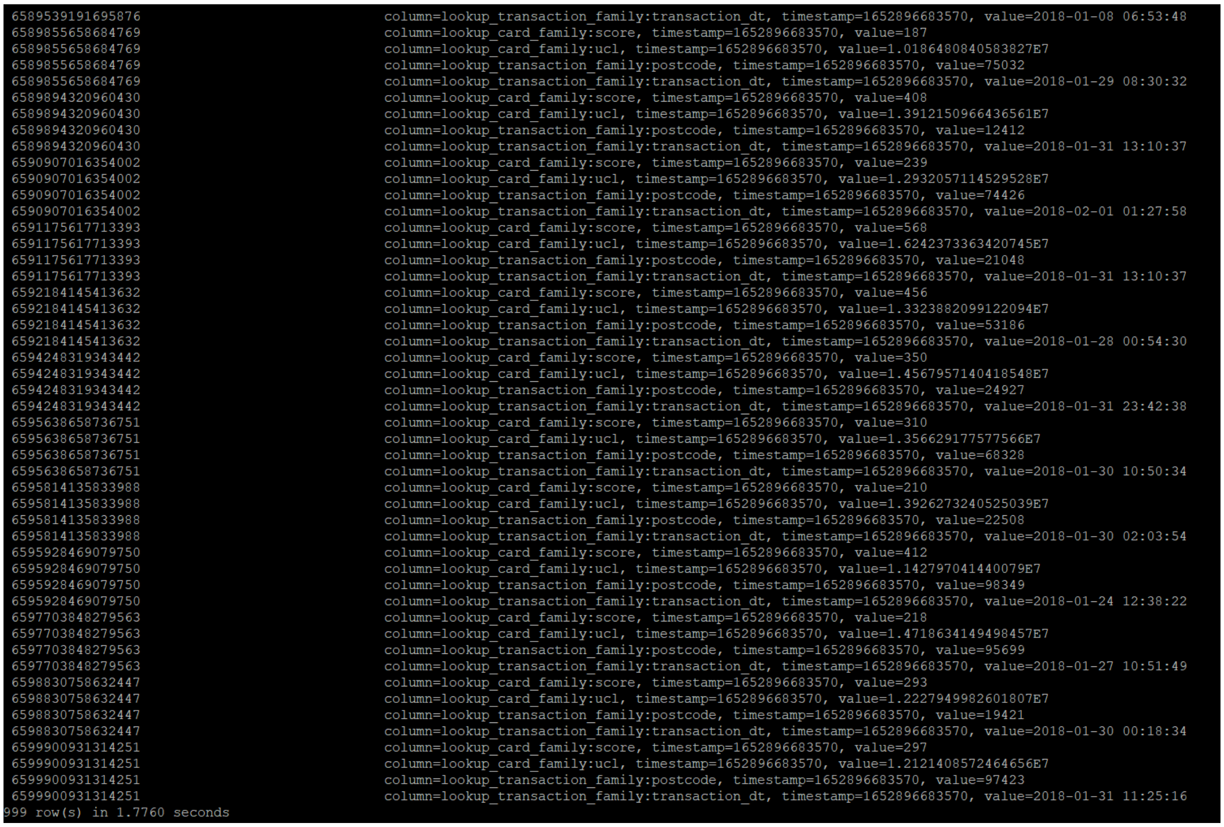
- Verify count in “**lookup_data_hive**” table.Record count is **999** which matches with the given requirement



```
hbase(main):001:0> count 'lookup_data_hive'
999 row(s) in 0.4970 seconds

=> 999
hbase(main):002:0>
```

Total number for record is 999 which is matching with given requirement.

- Verify data in “**lookup_data_hive**” table. Record count is **999** which matches with the given requirement



```
 6589539191695876                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-08 06:53:48
 6589855658684769                               column=lookup_card_family:score, timestamp=1652896683570, value=187
 6589855658684769                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.0186480840583827E7
 6589855658684769                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=75032
 6589855658684769                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-29 08:30:32
 6589894320960430                               column=lookup_card_family:score, timestamp=1652896683570, value=408
 6589894320960430                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.3912150966436561E7
 6589894320960430                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=12412
 6589894320960430                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-31 13:10:37
 6590907016354002                               column=lookup_card_family:score, timestamp=1652896683570, value=239
 6590907016354002                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.2932057114529528E7
 6590907016354002                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=74426
 6590907016354002                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-02-01 01:27:58
 6591175617713393                               column=lookup_card_family:score, timestamp=1652896683570, value=568
 6591175617713393                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.6242373363420745E7
 6591175617713393                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=21048
 6591175617713393                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-31 13:10:37
 6592184145413632                               column=lookup_card_family:score, timestamp=1652896683570, value=456
 6592184145413632                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.3323882099122094E7
 6592184145413632                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=53186
 6592184145413632                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-28 00:54:30
 6594248319343442                               column=lookup_card_family:score, timestamp=1652896683570, value=350
 6594248319343442                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.4567957140418548E7
 6594248319343442                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=24927
 6594248319343442                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-31 23:42:38
 6595638658736751                               column=lookup_card_family:score, timestamp=1652896683570, value=310
 6595638658736751                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.356629177577566E7
 6595638658736751                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=68328
 6595638658736751                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-30 10:50:34
 6595814135833988                               column=lookup_card_family:score, timestamp=1652896683570, value=210
 6595814135833988                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.3926273240525039E7
 6595814135833988                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=22508
 6595814135833988                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-30 02:03:54
 6595928469079750                               column=lookup_card_family:score, timestamp=1652896683570, value=412
 6595928469079750                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.142797041440079E7
 6595928469079750                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=98349
 6595928469079750                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-24 12:38:22
 6597703848279563                               column=lookup_card_family:score, timestamp=1652896683570, value=218
 6597703848279563                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.4718634149498457E7
 6597703848279563                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=95699
 6597703848279563                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-27 10:51:49
 6598830758632447                               column=lookup_card_family:score, timestamp=1652896683570, value=293
 6598830758632447                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.2227949982601807E7
 6598830758632447                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=19421
 6598830758632447                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-30 00:18:34
 6599900931314251                               column=lookup_card_family:score, timestamp=1652896683570, value=297
 6599900931314251                               column=lookup_card_family:ucl, timestamp=1652896683570, value=1.2121408572464656E7
 6599900931314251                               column=lookup_transaction_family:postcode, timestamp=1652896683570, value=97423
 6599900931314251                               column=lookup_transaction_family:transaction_dt, timestamp=1652896683570, value=2018-01-31 11:25:16
999 row(s) in 1.7760 seconds
```