Loading Historical Transactions Data into NoSQL Database

1. Commands to load the past transactions data into NoSQL database

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

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

```
create database ccfd_capstone_project;
use ccfd_capstone_project;
```

- Set below parameters for the hive session

```
set hive.auto.convert.join=false;
set hive.stats.autogather=true;
set orc.compress=SNAPPY;
set hive.exec.compress.output=true;
set mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCodec; set
mapred.output.compression.type=BLOCK;
set mapreduce.map.java.opts=-Xmx5G; set mapreduce.reduce.java.opts=-Xmx5G;
set mapred.child.java.opts=-Xmx5G -XX:+UseConcMarkSweepGC -XX:-UseGCOverheadLimit;
```

- Create an external table "**card_transactions_ext**" which will point to HDFS location created in ccfd_capstone_project location.

```
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.header.line.count"="1");
```

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

```
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");
```

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

```
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 SERDEPROPERTIES
("hbase.columns.mapping"=":key, card_transactions_family:card_id,
card_transactions_family:member_id, card_transactions_family:amount,
card_transactions_family:postcode, card_transactions_family:pos_id,
card_transactions_family:transaction_dt, card_transactions_family:status")
TBLPROPERTIES ("hbase.table.name"="card_transactions_hive");
```

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

2. <Command to list the table in which the data is loaded and the command to get the count of the rows of the table>

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

- Load data in "**card_transactions_orc**" table and type cast **transaction_dt** column in timestamp format

```
INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC SELECT CARD_ID, MEMBER_ID,
AMOUNT, POSTCODE,
POS_ID,CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTION_DT,'dd-MM-yyyy HH:mm:ss'))
AS TIMESTAMP), STATUS FROM CARD_TRANSACTIONS_EXT;
```

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

```
select year(transaction_dt), transaction_dt from card_transactions_orc limit 10;
```

- 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).

```
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;
```

- Verify data in "**card_transactions_hbase**" table.

```
select * from card_transactions_hbase limit 10;
```

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

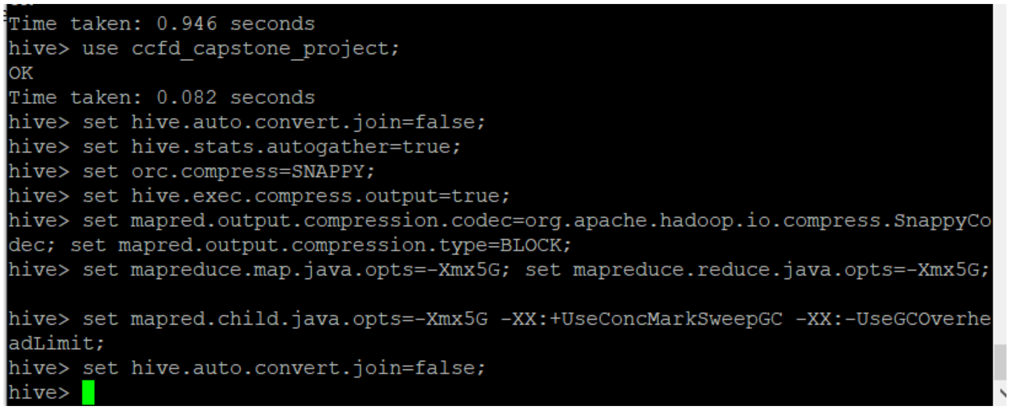
----------Hbase Operations: Starts Here----------

- Verify details of "**card_transactions_hive**" table (hive-hbase integrated table).
  describe 'card_transactions_hive'

- Check count of "**card_transactions_hive**" table.
  count 'card_transactions_hive''
  Expected output : 53292 rows
  ---------- Hbase Operations: Ends Here----------

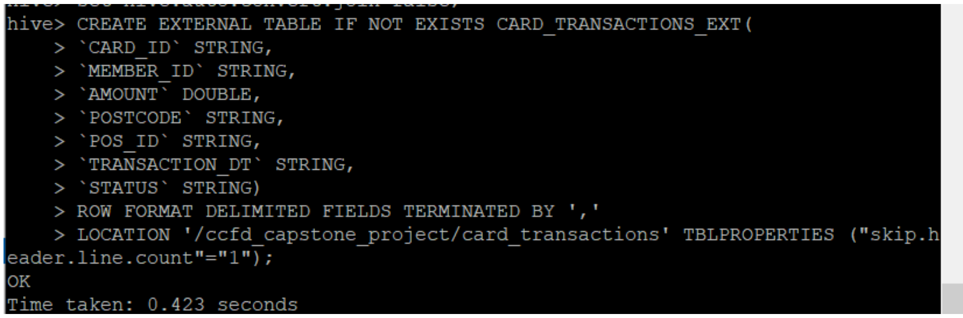
3. Screenshot of the table created

- create new database named **ccfd_capstone_project** and set parameters



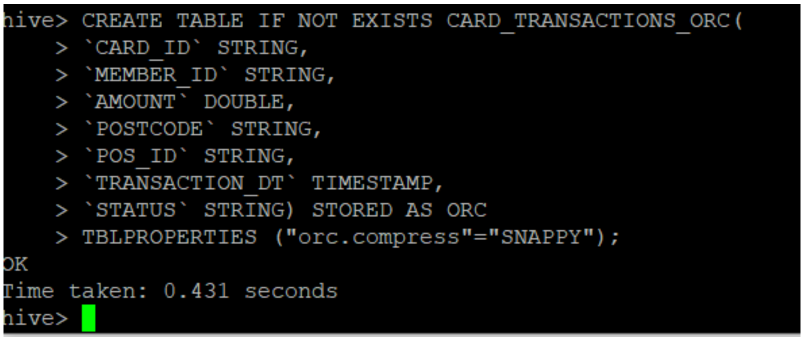
```
Time taken: 0.946 seconds
hive> use ccfd_capstone_project;
OK
Time taken: 0.082 seconds
hive> set hive.auto.convert.join=false;
hive> set hive.stats.autogather=true;
hive> set orc.compress=SNAPPY;
hive> set hive.exec.compress.output=true;
hive> set mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCo
dec; set mapred.output.compression.type=BLOCK;
hive> set mapreduce.map.java.opts=-Xmx5G; set mapreduce.reduce.java.opts=-Xmx5G;

hive> set mapred.child.java.opts=-Xmx5G -XX:+UseConcMarkSweepGC -XX:-UseGCOverhe
adLimit;
hive> set hive.auto.convert.join=false;
hive> 
```

- Create an external table "**card_transactions_ext**"



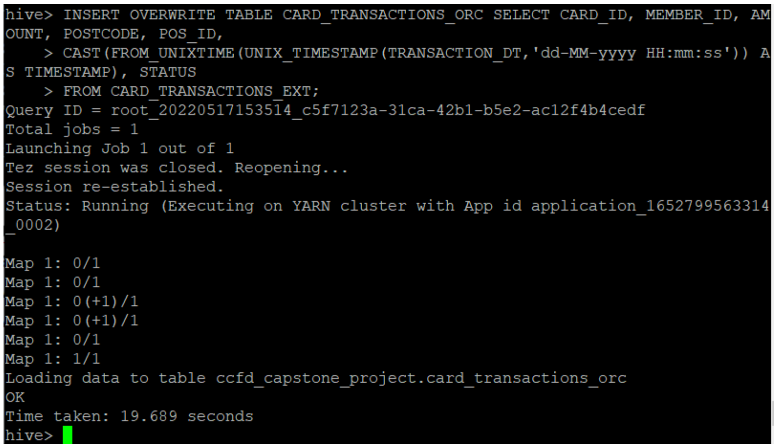
```
hive> set hive.auto.convert.join=false;
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CARD_TRANSACTIONS_EXT(
    > `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` STRING,
    > `STATUS` STRING)
    > ROW FORMAT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

- Create table "**card_transactions_orc**"



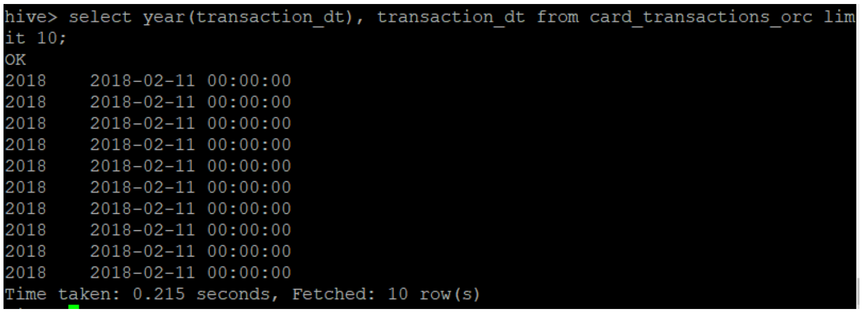
```
hive> CREATE TABLE IF NOT EXISTS CARD_TRANSACTIONS_ORC(
    > `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` TIMESTAMP,
    > `STATUS` STRING) STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK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

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



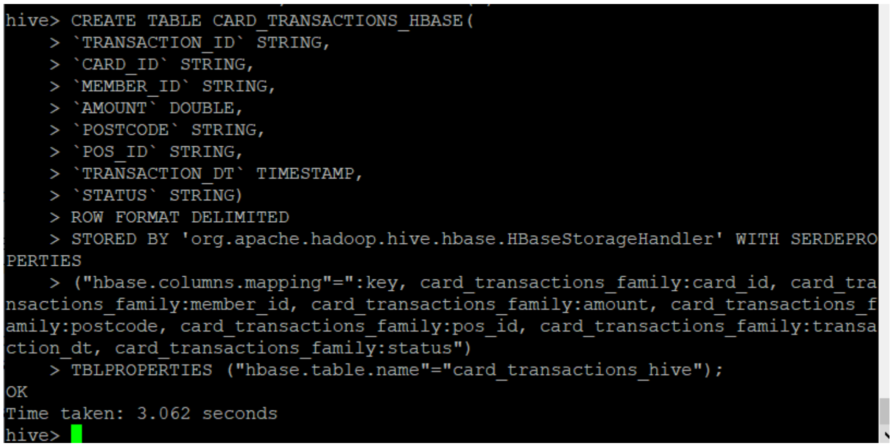
```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC SELECT CARD_ID, MEMBER_ID, AM
OUNT, POSTCODE, POS_ID,
    > CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTION_DT,'dd-MM-yyyy HH:mm:ss')) A
S TIMESTAMP), STATUS
    > FROM CARD_TRANSACTIONS_EXT;
Query ID = root_20220517153514_c5f7123a-31ca-42b1-b5e2-ac12f4b4cedf
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>
```

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



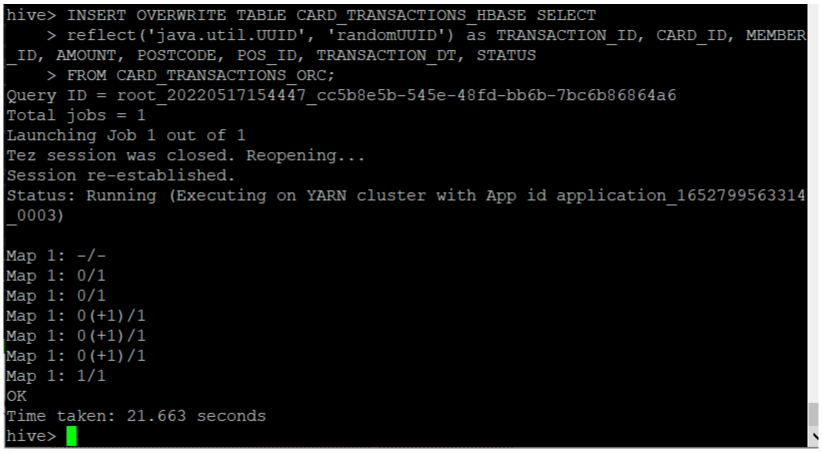
```
hive> select year(transaction_dt), transaction_dt from card_transactions_orc lim
it 10;
OK
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 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)
```

- Create hive-hbase integrated table. "**card_transactions_hbase**" table



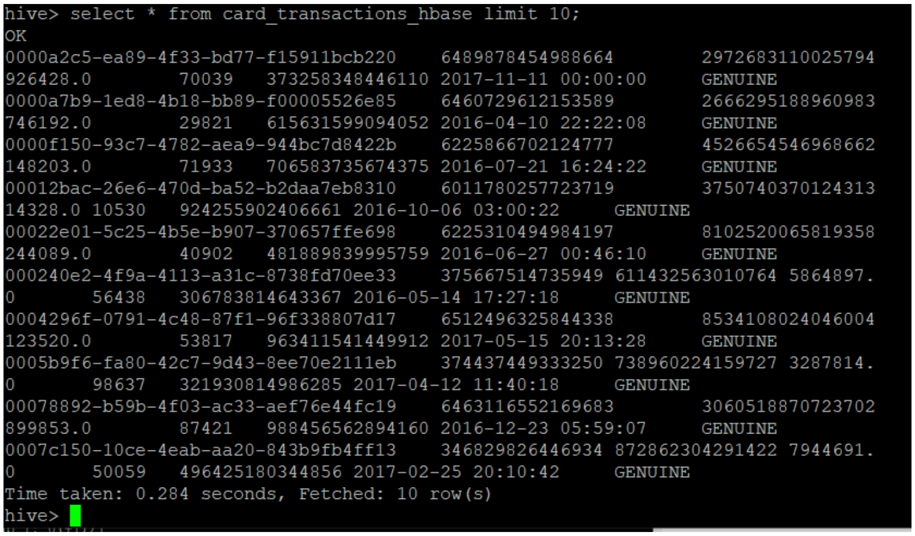
```
hive> CREATE TABLE CARD_TRANSACTIONS_HBASE(
    > `TRANSACTION_ID` STRING,
    > `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `AMOUNT` DOUBLE,
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` TIMESTAMP,
    > `STATUS` STRING)
    > ROW FORMAT DELIMITED
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' 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>
```

- Load data in "**card_transactions_hbase**" table



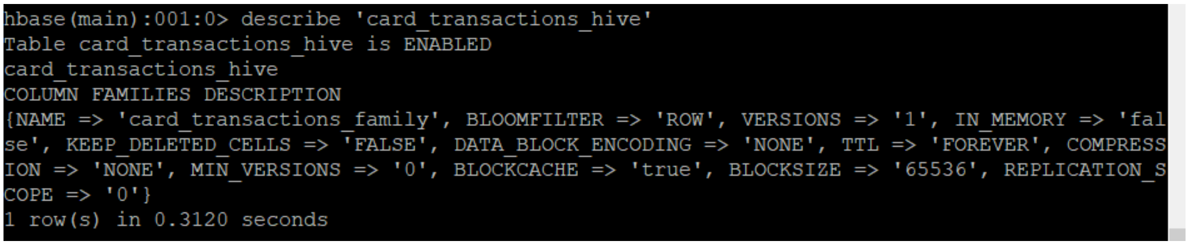
```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_HBASE SELECT
    > reflect('java.util.UUID', 'randomUUID') as TRANSACTION_ID, CARD_ID, MEMBER
_ID, AMOUNT, POSTCODE, POS_ID, TRANSACTION_DT, STATUS
    > FROM CARD_TRANSACTIONS_ORC;
Query ID = root_20220517154447_cc5b8e5b-545e-48fd-bb6b-7bc6b86864a6
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> 
```

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



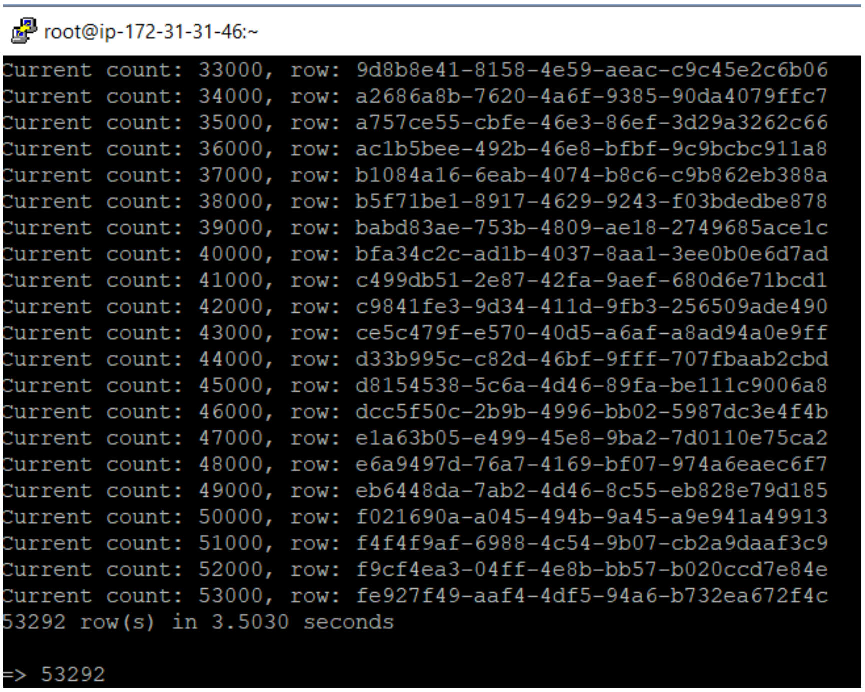
```
hive> select * from card_transactions_hbase limit 10;
OK
0000a2c5-ea89-4f33-bd77-f15911bcb220    6489878454988664        2972683110025794
926428.0        70039   373258348446110 2017-11-11 00:00:00     GENUINE
0000a7b9-1ed8-4b18-bb89-f00005526e85    6460729612153589        2666295188960983
746192.0        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> 
```

- Verify details of "**card_transactions_hive**" 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
```

- 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.