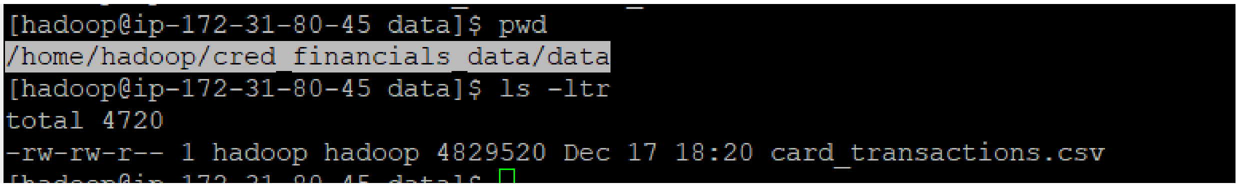
Loading Historical Transactions Data into NoSQL Database

Commands to load the past transactions data into NoSQL database

Here we are going to create HBase table for Card Transactions with Hive support (Hive-Hbase integration).

1. Create an empty Hbase table with the following command:
   create 'card_transactions', 'transaction_data'
2. Place the **card_transactions.csv** file into local file system in EMR cluster:
   /home/hadoop/cred_financials_data/data



```
[hadoop@ip-172-31-80-45 data]$ pwd
/home/hadoop/cred_financials_data/data
[hadoop@ip-172-31-80-45 data]$ ls -ltr
total 4720
-rw-rw-r-- 1 hadoop hadoop 4829520 Dec 17 18:20 card_transactions.csv
[hadoop@ip-172-31-80-45 data]$
```

3. Create Hive source table and load the CSV file data into it.
   CREATE TABLE IF NOT EXISTS card_transactions_src(
   card_id BIGINT,
   member_id BIGINT,
   amount INT,
   postcode INT,
   pos_id BIGINT,
   transaction_dt STRING,
   status STRING
)
   ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
   LOCATION '/user/hadoop/cred_financials_data/card_transactions_src'
   TBLPROPERTIES ('skip.header.line.count'='1');

   LOAD DATA LOCAL INPATH '/home/hadoop/cred_financials_data/data/card_transactions.csv' INTO TABLE card_transactions_src;

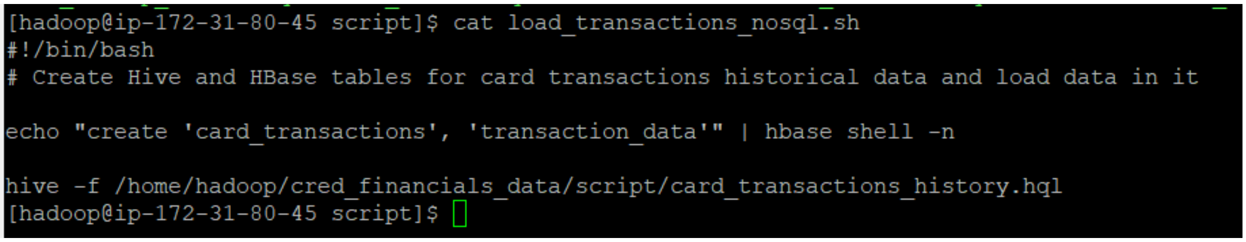
4. Create Hive staging table to perform type conversion on transaction_dt attribute (initially provided with different date format)
   CREATE TABLE IF NOT EXISTS card_transactions_stg(
   card_id BIGINT,
   member_id BIGINT,
   amount INT,

```
postcode INT,
pos_id BIGINT,
transaction_dt TIMESTAMP,
status STRING
)
location '/user/hadoop/cred_financials_data/card_transactions_stg';

INSERT INTO TABLE card_transactions_stg
SELECT card_id,
member_id,
amount,
postcode,
pos_id,
CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(transaction_dt, 'dd-MM-yyyy
HH:mm:ss'),'yyyy-MM-dd HH:mm:ss')AS TIMESTAMP) AS transaction_dt,
status
FROM card_transactions_src;
```

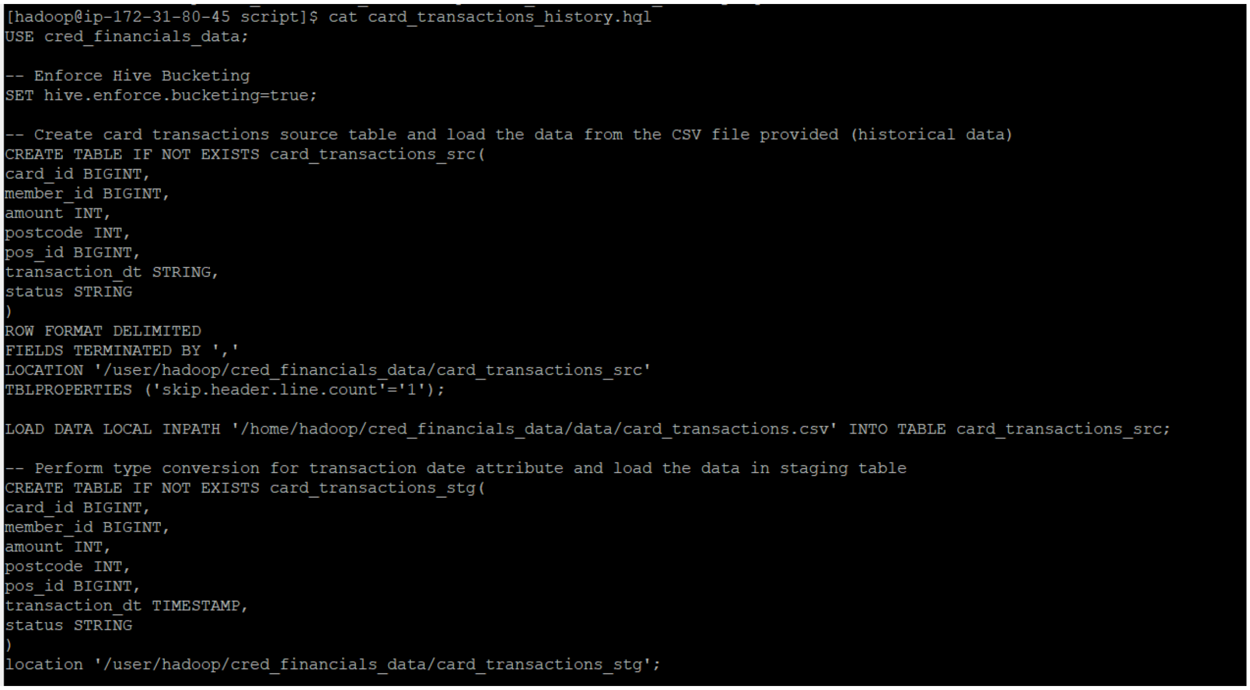
5. Finally create Hive-HBase table for card transactions with storage as 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with 16 buckets on card_id and member_id attributes for joins optimization and insert the data from staging table.

```
CREATE EXTERNAL TABLE IF NOT EXISTS card_transactions(
transaction_key STRING,
card_id BIGINT,
member_id BIGINT,
amount INT,
postcode INT,
pos_id BIGINT,
transaction_dt TIMESTAMP,
status STRING
)
CLUSTERED BY (card_id, member_id)INTO 16 BUCKETS
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping'=':key,
transaction_data:card_id, transaction_data:member_id,
transaction_data:amount, transaction_data:postcode,
transaction_data:pos_id, transaction_data:transaction_dt,
transaction_data:status')
TBLPROPERTIES ('hbase.table.name' = 'card_transactions');

INSERT INTO TABLE card_transactions
SELECT CONCAT_WS('|', CAST(card_id AS STRING), CAST(member_id AS
STRING), CAST(amount AS STRING), CAST(postcode AS STRING),
CAST(pos_id AS STRING), CAST(transaction_dt AS STRING), status) AS
transaction_key,
card_id,
member_id,
amount,
postcode,
pos_id,
transaction_dt,
status
FROM card_transactions_stg;
```

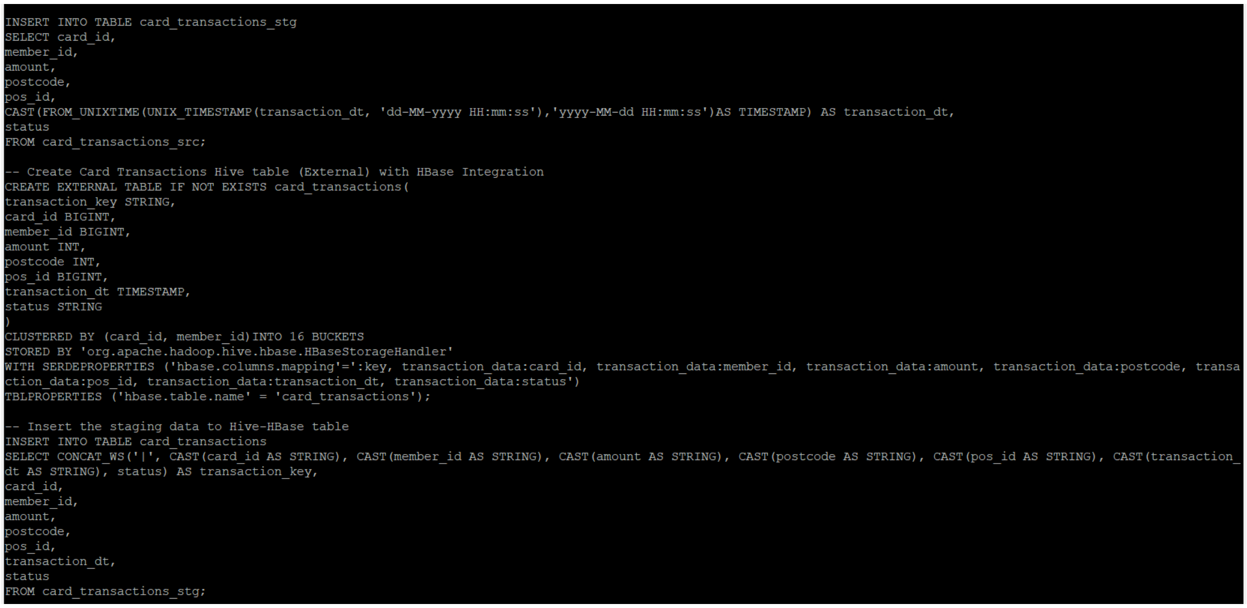
All the above steps are wrapped into a single shell script (creates empty HBase table and calls Hive script – **card_transactions_history.hql** to execute the above create and insert statements) with name **load_transactions_nosql.sh** placed in the path: **/home/hadoop/cred_financials_data/script/load_transactions_nosql.sh**



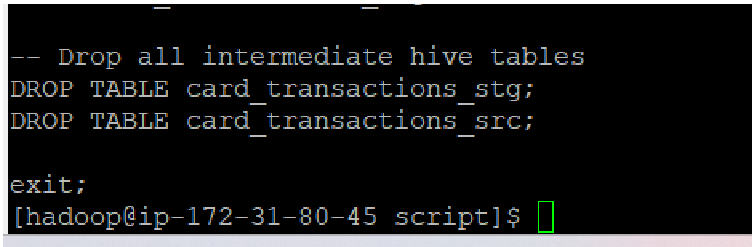
```
[hadoop@ip-172-31-80-45 script]$ cat load_transactions_nosql.sh
#!/bin/bash
# Create Hive and HBase tables for card transactions historical data and load data in it

echo "create 'card_transactions', 'transaction_data'" | hbase shell -n

hive -f /home/hadoop/cred_financials_data/script/card_transactions_history.hql
[hadoop@ip-172-31-80-45 script]$ 
```



```
[hadoop@ip-172-31-80-45 script]$ cat card_transactions_history.hql
USE cred_financials_data;

-- Enforce Hive Bucketing
SET hive.enforce.bucketing=true;

-- Create card transactions source table and load the data from the CSV file provided (historical data)
CREATE TABLE IF NOT EXISTS card_transactions_src(
card_id BIGINT,
member_id BIGINT,
amount INT,
postcode INT,
pos_id BIGINT,
transaction_dt STRING,
status STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION '/user/hadoop/cred_financials_data/card_transactions_src'
TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA LOCAL INPATH '/home/hadoop/cred_financials_data/data/card_transactions.csv' INTO TABLE card_transactions_src;

-- Perform type conversion for transaction date attribute and load the data in staging table
CREATE TABLE IF NOT EXISTS card_transactions_stg(
card_id BIGINT,
member_id BIGINT,
amount INT,
postcode INT,
pos_id BIGINT,
transaction_dt TIMESTAMP,
status STRING
)
location '/user/hadoop/cred_financials_data/card_transactions_stg';
```



```
INSERT INTO TABLE card_transactions_stg
SELECT card_id,
member_id,
amount,
postcode,
pos_id,
CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(transaction_dt, 'dd-MM-yyyy HH:mm:ss'),'yyyy-MM-dd HH:mm:ss')AS TIMESTAMP) AS transaction_dt,
status
FROM card_transactions_src;

-- Create Card Transactions Hive table (External) with HBase Integration
CREATE EXTERNAL TABLE IF NOT EXISTS card_transactions(
transaction_key STRING,
card_id BIGINT,
member_id BIGINT,
amount INT,
postcode INT,
pos_id BIGINT,
transaction_dt TIMESTAMP,
status STRING
)
CLUSTERED BY (card_id, member_id)INTO 16 BUCKETS
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping'=':key, transaction_data:card_id, transaction_data:member_id, transaction_data:amount, transaction_data:postcode, transa
ction_data:pos_id, transaction_data:transaction_dt, transaction_data:status')
TBLPROPERTIES ('hbase.table.name' = 'card_transactions');

-- Insert the staging data to Hive-HBase table
INSERT INTO TABLE card_transactions
SELECT CONCAT_WS('|', CAST(card_id AS STRING), CAST(member_id AS STRING), CAST(amount AS STRING), CAST(postcode AS STRING), CAST(pos_id AS STRING), CAST(transaction_
dt AS STRING), status) AS transaction_key,
card_id,
member_id,
amount,
postcode,
pos_id,
transaction_dt,
status
FROM card_transactions_stg;
```



```
-- Drop all intermediate hive tables
DROP TABLE card_transactions_stg;
DROP TABLE card_transactions_src;

exit;
[hadoop@ip-172-31-80-45 script]$
```

Execute the script by running the below command:

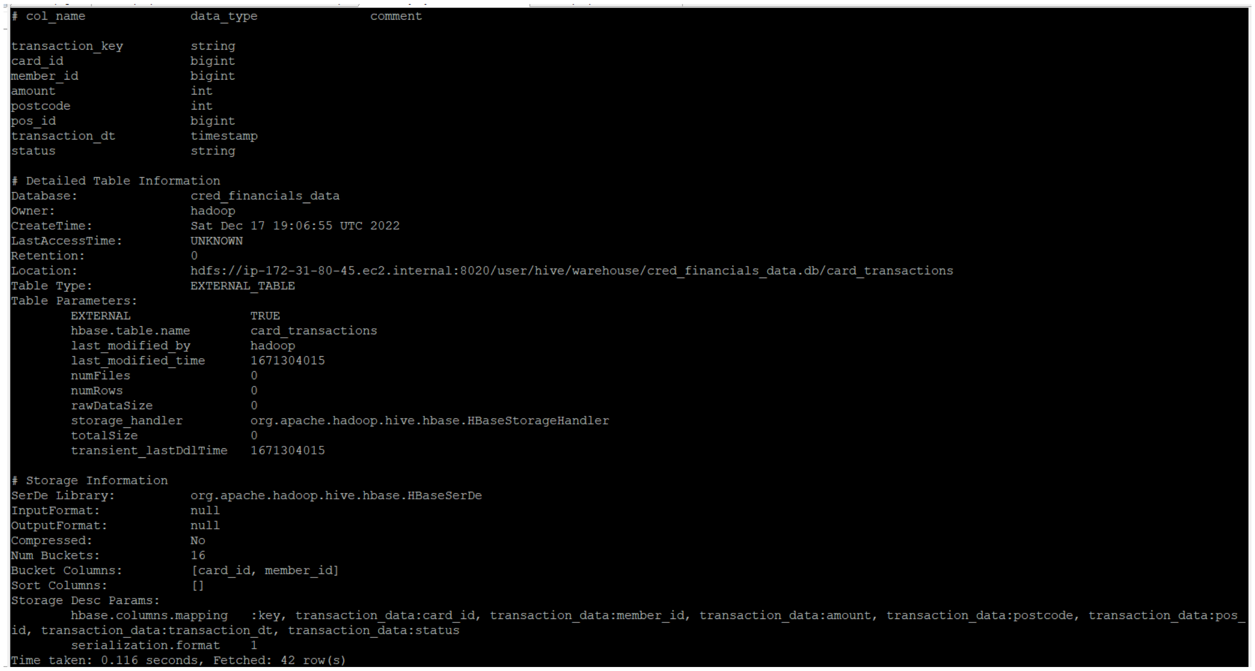
/home/hadoop/cred_financials_data/script/load_transactions_nosql.sh

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: -

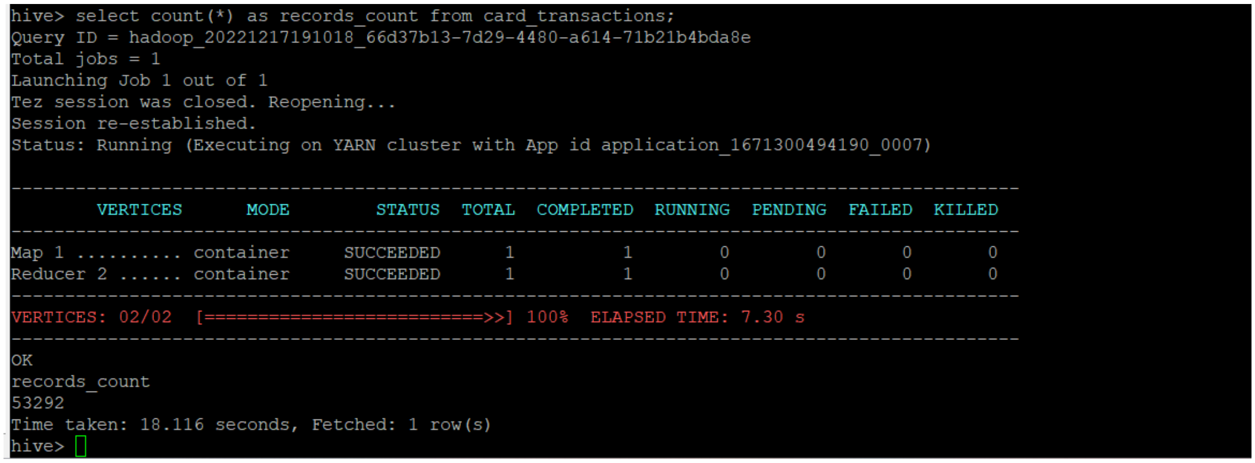
use cred_financials_data;

describe formatted card_transactions;



```
# col_name              data_type               comment

transaction_key         string
card_id                 bigint
member_id               bigint
amount                  int
postcode                int
pos_id                  bigint
transaction_dt          timestamp
status                  string

# Detailed Table Information
Database:               cred_financials_data
Owner:                  hadoop
CreateTime:             Sat Dec 17 19:06:55 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://ip-172-31-80-45.ec2.internal:8020/user/hive/warehouse/cred_financials_data.db/card_transactions
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        hbase.table.name        card_transactions
        last_modified_by        hadoop
        last_modified_time      1671304015
        numFiles                0
        numRows                 0
        rawDataSize             0
        storage_handler         org.apache.hadoop.hive.hbase.HBaseStorageHandler
        totalSize               0
        transient_lastDdlTime   1671304015

# Storage Information
SerDe Library:          org.apache.hadoop.hive.hbase.HBaseSerDe
InputFormat:            null
OutputFormat:           null
Compressed:             No
Num Buckets:            16
Bucket Columns:         [card_id, member_id]
Sort Columns:           []
Storage Desc Params:
        hbase.columns.mapping   :key, transaction_data:card_id, transaction_data:member_id, transaction_data:amount, transaction_data:postcode, transaction_data:pos_
id, transaction_data:transaction_dt, transaction_data:status
        serialization.format    1
Time taken: 0.116 seconds, Fetched: 42 row(s)
```

select count(*) as records_count from card_transactions;

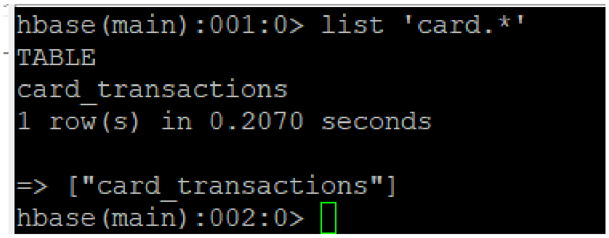


```
hive> select count(*) as records_count from card_transactions;
Query ID = hadoop_20221217191018_66d37b13-7d29-4480-a614-71b21b4bda8e
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1671300494190_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.30 s
----------------------------------------------------------------------------------------------
OK
records_count
53292
Time taken: 18.116 seconds, Fetched: 1 row(s)
hive>
```

Returned count as 53292 which matches with source card_transactions.csv file.

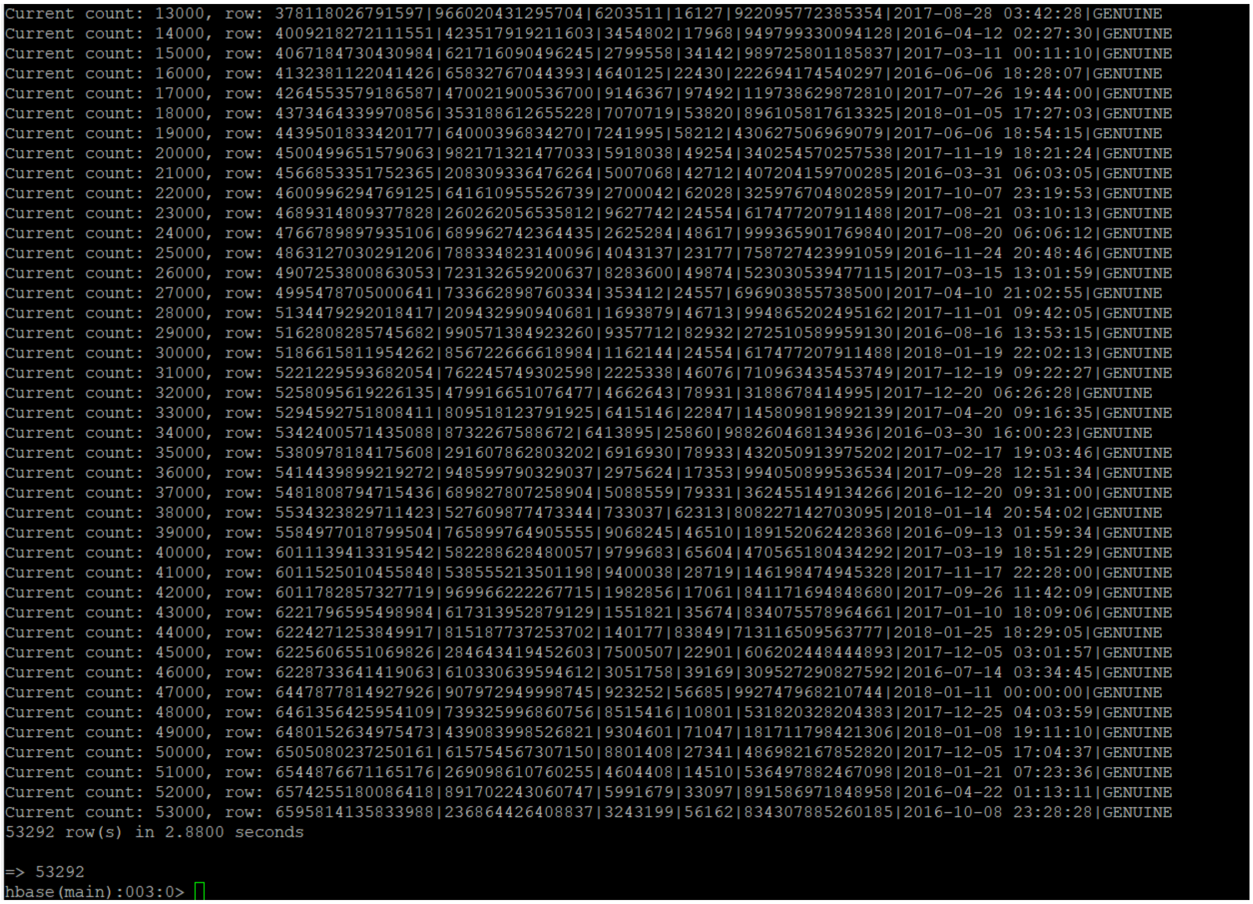
<u>HBase: -</u>

list ‘card.*’



```
hbase(main):001:0> list 'card.*'
TABLE
card_transactions
1 row(s) in 0.2070 seconds

=> ["card_transactions"]
hbase(main):002:0> 
```

count ‘card_transactions’



```
Current count: 13000, row: 378118026791597|966020431295704|6203511|16127|922095772385354|2017-08-28 03:42:28|GENUINE
Current count: 14000, row: 4009218272111551|423517919211603|3454802|17968|949799330094128|2016-04-12 02:27:30|GENUINE
Current count: 15000, row: 4067184730430984|621716090496245|2799558|34142|989725801185837|2017-03-11 00:11:10|GENUINE
Current count: 16000, row: 4132381122041426|65832767044393|4640125|22430|222694174540297|2016-06-06 18:28:07|GENUINE
Current count: 17000, row: 4264553579186587|470021900536700|9146367|97492|119738629872810|2017-07-26 19:44:00|GENUINE
Current count: 18000, row: 4373464339970856|353188612655228|7070719|53820|896105817613325|2018-01-05 17:27:03|GENUINE
Current count: 19000, row: 4439501833420177|64000396834270|7241995|58212|430627506969079|2017-06-06 18:54:15|GENUINE
Current count: 20000, row: 4500499651579063|982171321477033|5918038|49254|340254570257538|2017-11-19 18:21:24|GENUINE
Current count: 21000, row: 4566853351752365|208309336476264|5007068|42712|407204159700285|2016-03-31 06:03:05|GENUINE
Current count: 22000, row: 4600996294769125|641610955526739|2700042|62028|325976704802859|2017-10-07 23:19:53|GENUINE
Current count: 23000, row: 4689314809377828|260262056535812|9627742|24554|617477207911488|2017-08-21 03:10:13|GENUINE
Current count: 24000, row: 4766789897935106|689962742364435|2625284|48617|999365901769840|2017-08-20 06:06:12|GENUINE
Current count: 25000, row: 4863127030291206|788334823140096|4043137|23177|758727423991059|2016-11-24 20:48:46|GENUINE
Current count: 26000, row: 4907253800863053|723132659200637|8283600|49874|523030539477115|2017-03-15 13:01:59|GENUINE
Current count: 27000, row: 4995478705000641|733662898760334|353412|24557|696903855738500|2017-04-10 21:02:55|GENUINE
Current count: 28000, row: 5134479292018417|209432990940681|1693879|46713|994865202495162|2017-11-01 09:42:05|GENUINE
Current count: 29000, row: 5162808285745682|990571384923260|9357712|82932|272510589959130|2016-08-16 13:53:15|GENUINE
Current count: 30000, row: 5186615811954262|856722666618984|1162144|24554|617477207911488|2018-01-19 22:02:13|GENUINE
Current count: 31000, row: 5221229593682054|762245749302598|2225338|46076|710963435453749|2017-12-19 09:22:27|GENUINE
Current count: 32000, row: 5258095619226135|479916651076477|4662643|78931|3188678414995|2017-12-20 06:26:28|GENUINE
Current count: 33000, row: 5294592751808411|809518123791925|6415146|22847|145809819892139|2017-04-20 09:16:35|GENUINE
Current count: 34000, row: 5342400571435088|8732267588672|6413895|25860|988260468134936|2016-03-30 16:00:23|GENUINE
Current count: 35000, row: 5380978184175608|291607862803202|6916930|78933|432050913975202|2017-02-17 19:03:46|GENUINE
Current count: 36000, row: 5414439899219272|948599790329037|2975624|17353|994050899536534|2017-09-28 12:51:34|GENUINE
Current count: 37000, row: 5481808794715436|689827807258904|5088559|79331|362455149134266|2016-12-20 09:31:00|GENUINE
Current count: 38000, row: 5534323829711423|527609877473344|733037|62313|808227142703095|2018-01-14 20:54:02|GENUINE
Current count: 39000, row: 5584977018799504|765899764905555|9068245|46510|189152062428368|2016-09-13 01:59:34|GENUINE
Current count: 40000, row: 6011139413319542|582288628480057|9799683|65604|470565180434292|2017-03-19 18:51:29|GENUINE
Current count: 41000, row: 6011525010455848|538555213501198|9400038|28719|146198474945328|2017-11-17 22:28:00|GENUINE
Current count: 42000, row: 6011782857327719|969966222267715|1982856|17061|841171694848680|2017-09-26 11:42:09|GENUINE
Current count: 43000, row: 6221796595498984|617313952879129|1551821|35674|834075578964661|2017-01-10 18:09:06|GENUINE
Current count: 44000, row: 6224271253849917|815187737253702|140177|83849|713116509563777|2018-01-25 18:29:05|GENUINE
Current count: 45000, row: 6225606551069826|284643419452603|7500507|22901|606202448444893|2017-12-05 03:01:57|GENUINE
Current count: 46000, row: 6228733641419063|610330639594612|3051758|39169|309527290827592|2016-07-14 03:34:45|GENUINE
Current count: 47000, row: 6447877814927926|907972949998745|923252|56685|992747968210744|2018-01-11 00:00:00|GENUINE
Current count: 48000, row: 6461356425954109|739325996860756|8515416|10801|531820328204383|2017-12-25 04:03:59|GENUINE
Current count: 49000, row: 6480152634975473|439083998526821|9304601|71047|181711798421306|2018-01-08 19:11:10|GENUINE
Current count: 50000, row: 6505080237250161|615754567307150|8801408|27341|486982167852820|2017-12-05 17:04:37|GENUINE
Current count: 51000, row: 6544876671165176|269098610760255|4604408|14510|536497882467098|2018-01-21 07:23:36|GENUINE
Current count: 52000, row: 6574255180086418|891702243060747|5991679|33097|891586971848958|2016-04-22 01:13:11|GENUINE
Current count: 53000, row: 6595814135833988|236864426408837|3243199|56162|834307885260185|2016-10-08 23:28:28|GENUINE
53292 row(s) in 2.8800 seconds

=> 53292
hbase(main):003:0> 
```

Returned count as 53292 which matches with source card_transactions.csv file.