Loading Historical Transactions Data into NoSQL Database

Commands to load the past transactions data into NoSQL database

1. Creating External table in 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 '/user/CCFD_Project/card_transactions.csv'
TBLPROPERTIES ("skip.header.line.count"="1");
```

2. Loading the data from to external hive table

```
LOAD DATA LOCAL INPATH 'CARD_TRANSACTIONS.CSV' INTO TABLE
CARD_TRANSACTIONS_EXT;
```

3. Converting external table to ORC provides a more efficient and optimized data storage and retrieval solution for 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");
```

4. Inserting data into ORC table

```
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;
```

5. Creating hive-hbase integrated table which will be visible in HBase as well

```
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");
```

6. Inserting data into Hive-Hbase integrated table

```
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;
```

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

1. Verifying the data for CARD_TRANSACTIONS_EXT

```
SELECT COUNT(*) FROM CARD_TRANSACTIONS_EXT;
```

2. Verifying the data for CARD_TRANSACTIONS_ORC

```
SELECT * FROM CARD_TRANSACTIONS_ORC LIMIT 5;
```

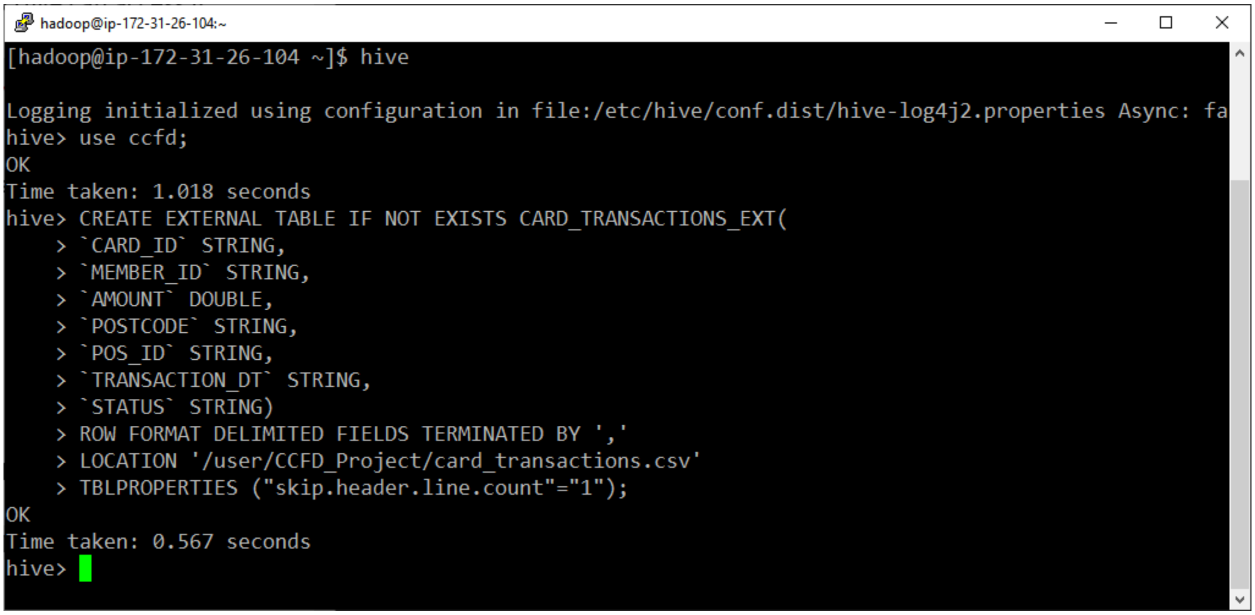
3. Verifying the data for CARD_TRANSACTIONS_HBASE

```
SELECT * FROM CARD_TRANSACTIONS_HBASE LIMIT 10;
```

4. Verifying if the timestamp is working properly.

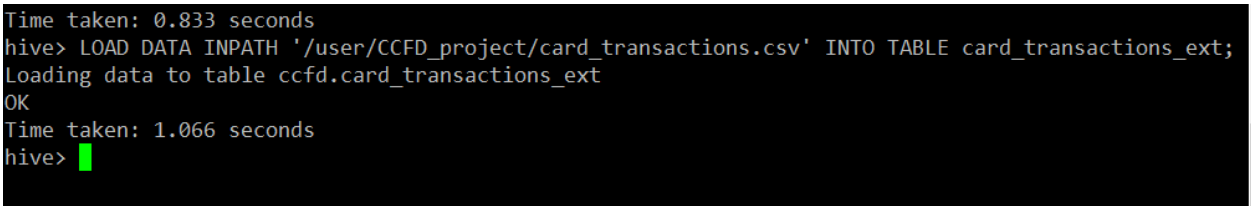
```
SELECT YEAR(TRANSACTION_DT),TRANSACTION_DT FROM
CARD_TRANSACTIONS_ORC LIMIT 10;
```

Screenshot of the table created



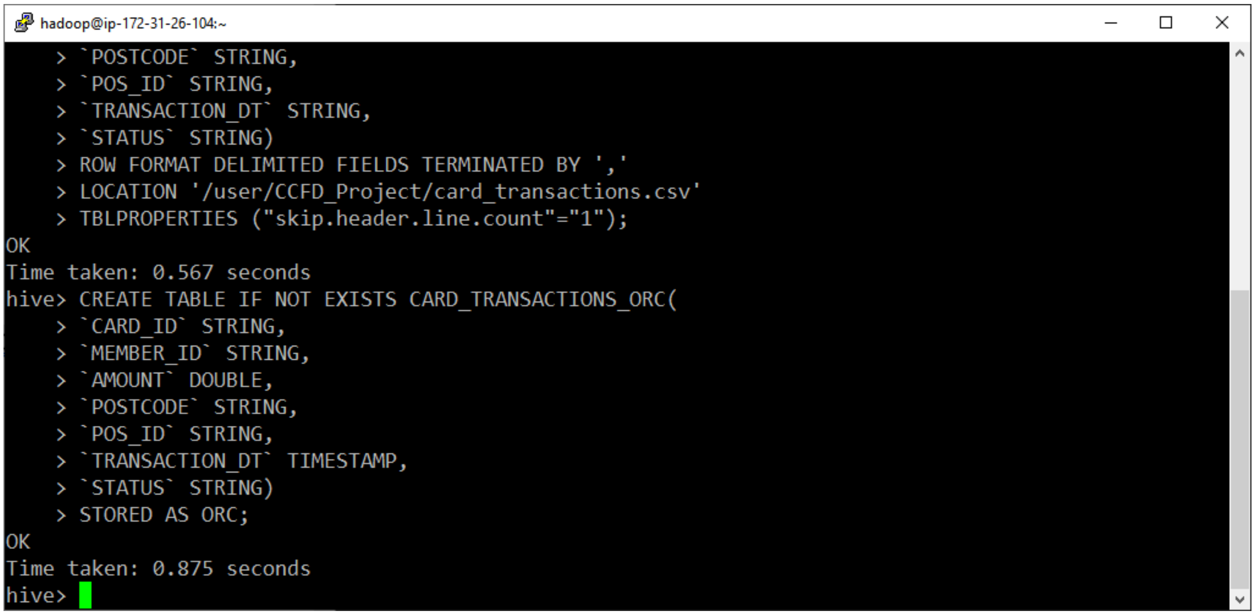
```
[hadoop@ip-172-31-26-104 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: fa
hive> use ccfd;
OK
Time taken: 1.018 seconds
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 '/user/CCFD_Project/card_transactions.csv'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.567 seconds
hive> 
```

Console Screenshot 1:Creating historical data External table.



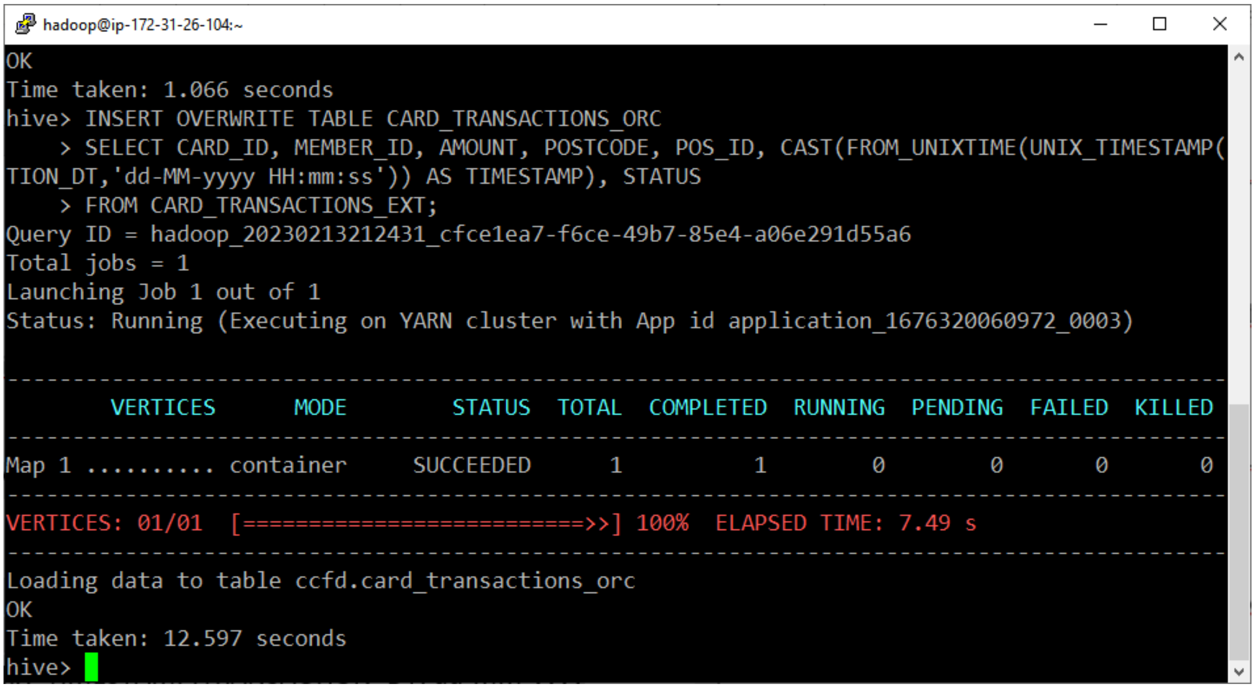
```
Time taken: 0.833 seconds
hive> LOAD DATA INPATH '/user/CCFD_project/card_transactions.csv' INTO TABLE card_transactions_ext;
Loading data to table ccfd.card_transactions_ext
OK
Time taken: 1.066 seconds
hive> 
```

Console Screenshot 2: Console Screenshot 7: loading data into external table



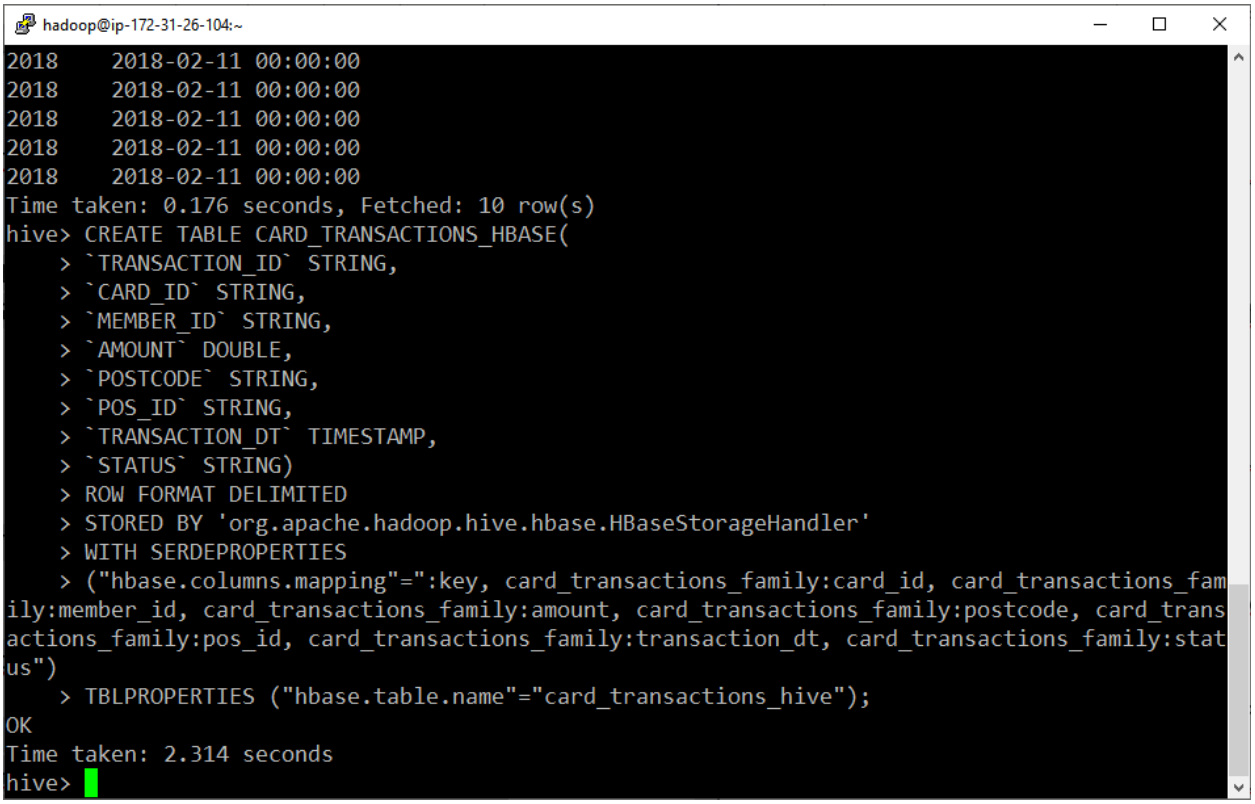
```
    > `POSTCODE` STRING,
    > `POS_ID` STRING,
    > `TRANSACTION_DT` STRING,
    > `STATUS` STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/user/CCFD_Project/card_transactions.csv'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.567 seconds
hive> 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;
OK
Time taken: 0.875 seconds
hive>
```

Console Screenshot 3: Creating ORC table from external table



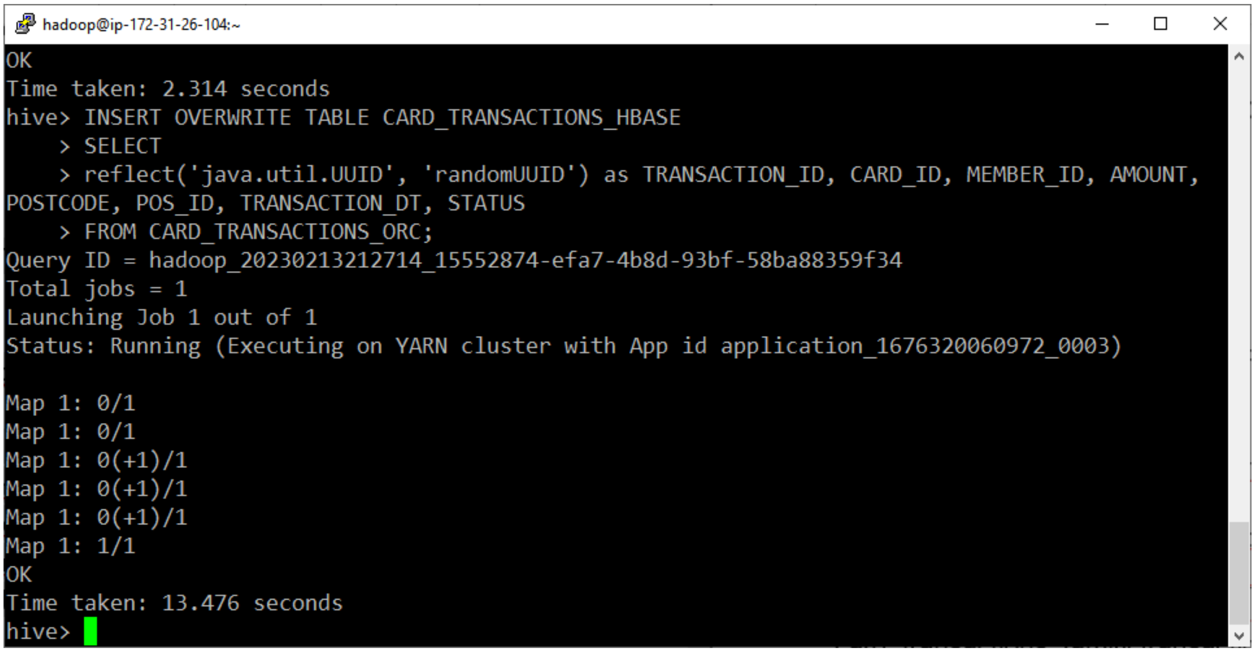
```
OK
Time taken: 1.066 seconds
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC
    > SELECT CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS_ID, CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(
TION_DT,'dd-MM-yyyy HH:mm:ss')) AS TIMESTAMP), STATUS
    > FROM CARD_TRANSACTIONS_EXT;
Query ID = hadoop_20230213212431_cfce1ea7-f6ce-49b7-85e4-a06e291d55a6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1676320060972_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.49 s
----------------------------------------------------------------------------------------------
Loading data to table ccfd.card_transactions_orc
OK
Time taken: 12.597 seconds
hive>
```

Console Screenshot 4: Inserting data into orc table.



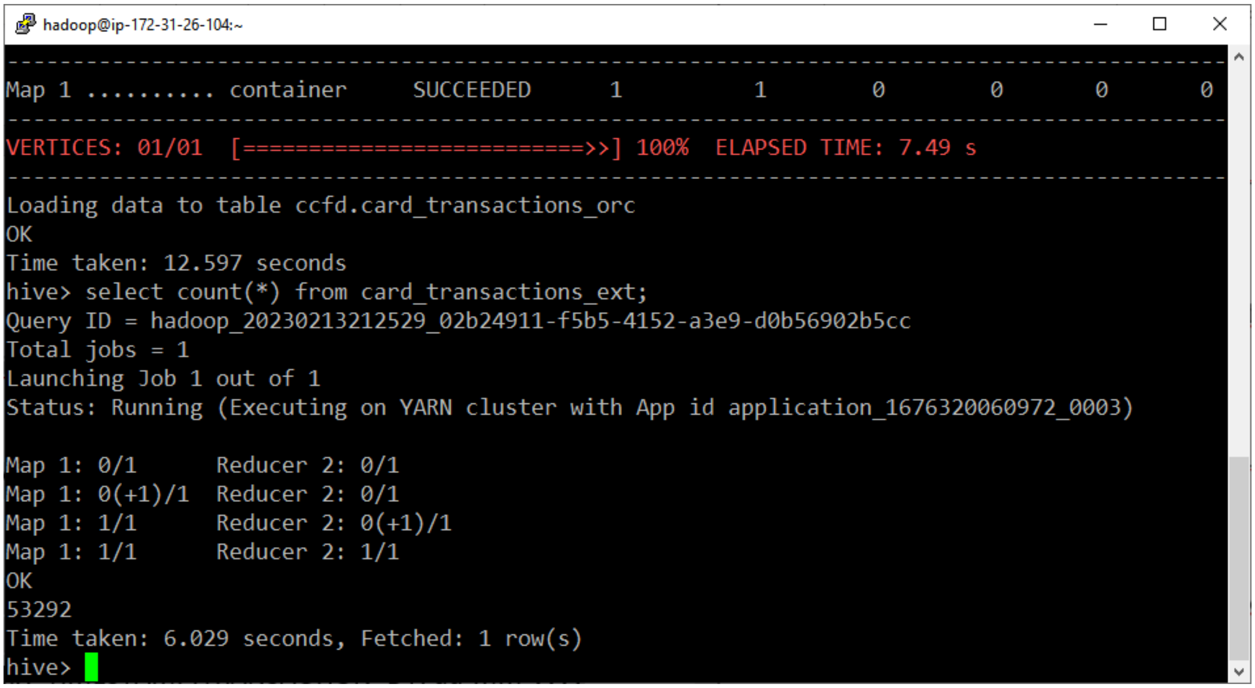
```
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
Time taken: 0.176 seconds, Fetched: 10 row(s)
hive> CREATE TABLE CARD_TRANSACTIONS_HBASE(
    > `TRANSACTION_ID` STRING,
    > `CARD_ID` STRING,
    > `MEMBER_ID` 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 SERDEPROPERTIES
    > ("hbase.columns.mapping"=":key, card_transactions_family:card_id, card_transactions_fam
ily:member_id, card_transactions_family:amount, card_transactions_family:postcode, card_trans
actions_family:pos_id, card_transactions_family:transaction_dt, card_transactions_family:stat
us")
    > TBLPROPERTIES ("hbase.table.name"="card_transactions_hive");
OK
Time taken: 2.314 seconds
hive>
```

Console Screenshot 5: Creating card_transactions_hbase hive-hbase integrated table



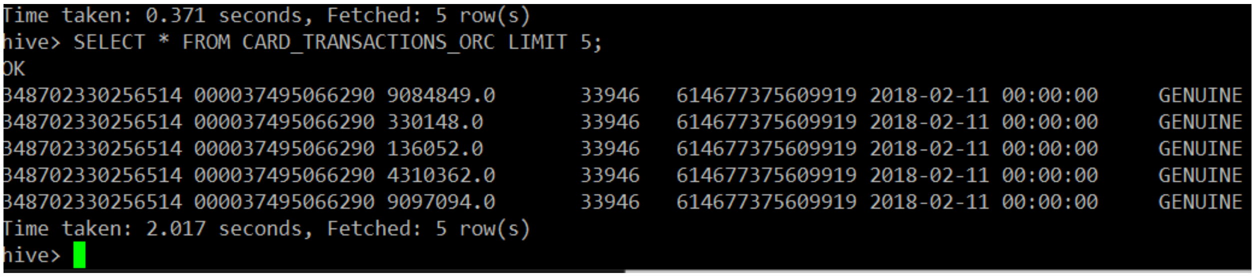
```
OK
Time taken: 2.314 seconds
hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_HBASE
    > SELECT
    > reflect('java.util.UUID', 'randomUUID') as TRANSACTION_ID, CARD_ID, MEMBER_ID, AMOUNT,
POSTCODE, POS_ID, TRANSACTION_DT, STATUS
    > FROM CARD_TRANSACTIONS_ORC;
Query ID = hadoop_20230213212714_15552874-efa7-4b8d-93bf-58ba88359f34
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1676320060972_0003)

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: 13.476 seconds
hive>
```

Console Screenshot 6: Inserting data into from orc table.



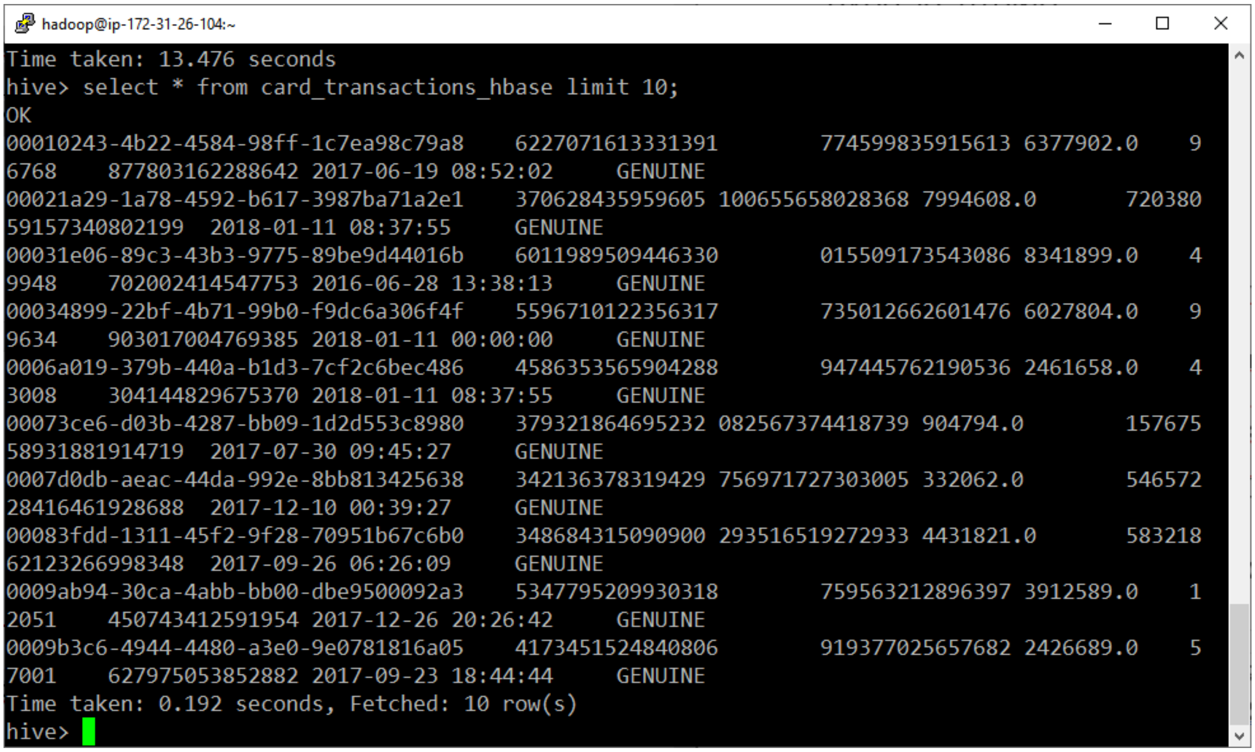
```
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.49 s
--------------------------------------------------------------------------------
Loading data to table ccfd.card_transactions_orc
OK
Time taken: 12.597 seconds
hive> select count(*) from card_transactions_ext;
Query ID = hadoop_20230213212529_02b24911-f5b5-4152-a3e9-d0b56902b5cc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1676320060972_0003)

Map 1: 0/1      Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/1
Map 1: 1/1      Reducer 2: 1/1
OK
53292
Time taken: 6.029 seconds, Fetched: 1 row(s)
hive>
```

Console Screenshot 7: Validating external table records.



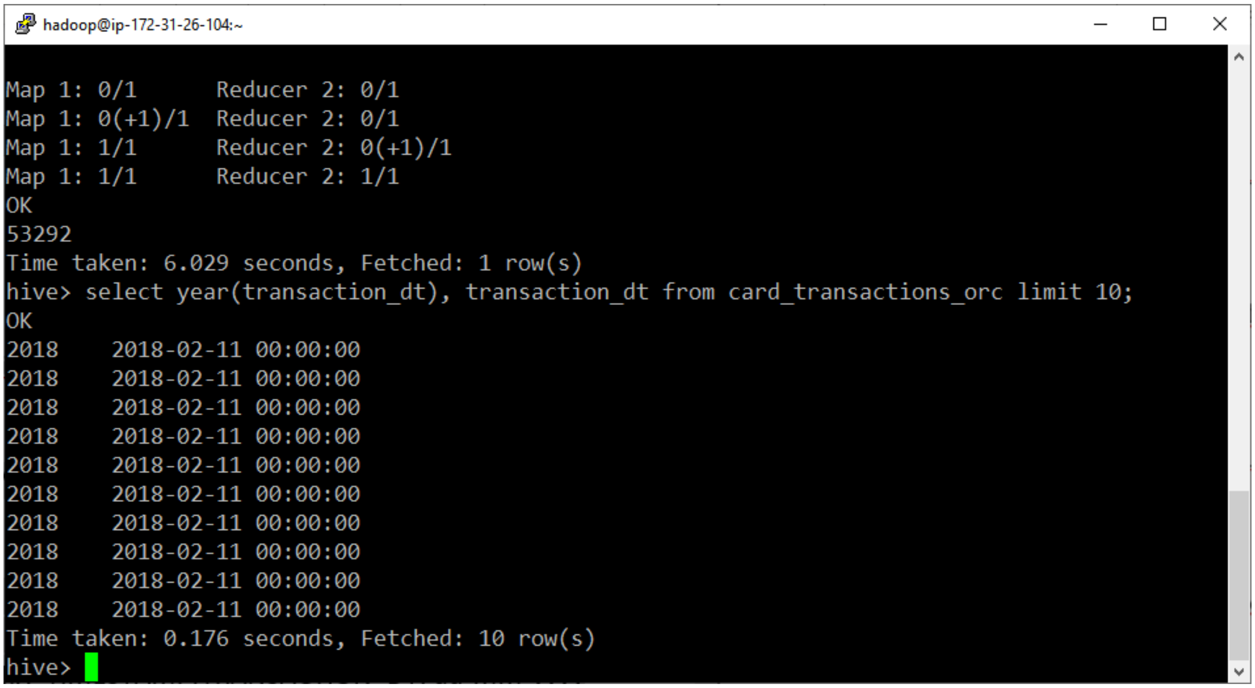
```
Time taken: 0.371 seconds, Fetched: 5 row(s)
hive> SELECT * FROM CARD_TRANSACTIONS_ORC LIMIT 5;
OK
348702330256514 000037495066290 9084849.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 330148.0        33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 136052.0        33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 4310362.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 9097094.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
Time taken: 2.017 seconds, Fetched: 5 row(s)
hive>
```

Console Screenshot 8: Validating external table records



```
Time taken: 13.476 seconds
hive> select * from card_transactions_hbase limit 10;
OK
00010243-4b22-4584-98ff-1c7ea98c79a8    6227071613331391        774599835915613 6377902.0    9
6768    877803162288642 2017-06-19 08:52:02     GENUINE
00021a29-1a78-4592-b617-3987ba71a2e1    370628435959605 100655658028368 7994608.0       720380
59157340802199  2018-01-11 08:37:55     GENUINE
00031e06-89c3-43b3-9775-89be9d44016b    6011989509446330        015509173543086 8341899.0    4
9948    702002414547753 2016-06-28 13:38:13     GENUINE
00034899-22bf-4b71-99b0-f9dc6a306f4f    5596710122356317        735012662601476 6027804.0    9
9634    903017004769385 2018-01-11 00:00:00     GENUINE
0006a019-379b-440a-b1d3-7cf2c6bec486    4586353565904288        947445762190536 2461658.0    4
3008    304144829675370 2018-01-11 08:37:55     GENUINE
00073ce6-d03b-4287-bb09-1d2d553c8980    379321864695232 082567374418739 904794.0        157675
58931881914719  2017-07-30 09:45:27     GENUINE
0007d0db-aeac-44da-992e-8bb813425638    342136378319429 756971727303005 332062.0        546572
28416461928688  2017-12-10 00:39:27     GENUINE
00083fdd-1311-45f2-9f28-70951b67c6b0    348684315090900 293516519272933 4431821.0       583218
62123266998348  2017-09-26 06:26:09     GENUINE
0009ab94-30ca-4abb-bb00-dbe9500092a3    5347795209930318        759563212896397 3912589.0    1
2051    450743412591954 2017-12-26 20:26:42     GENUINE
0009b3c6-4944-4480-a3e0-9e0781816a05    4173451524840806        919377025657682 2426689.0    5
7001    627975053852882 2017-09-23 18:44:44     GENUINE
Time taken: 0.192 seconds, Fetched: 10 row(s)
hive>
```

Console Screenshot 9: Validating HBase table data.



```
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/1
Map 1: 1/1      Reducer 2: 1/1
OK
53292
Time taken: 6.029 seconds, Fetched: 1 row(s)
hive> select year(transaction_dt), transaction_dt from card_transactions_orc limit 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: 0.176 seconds, Fetched: 10 row(s)
hive>
```

Console Screenshot 10 : Validating timestamp column in orc table.