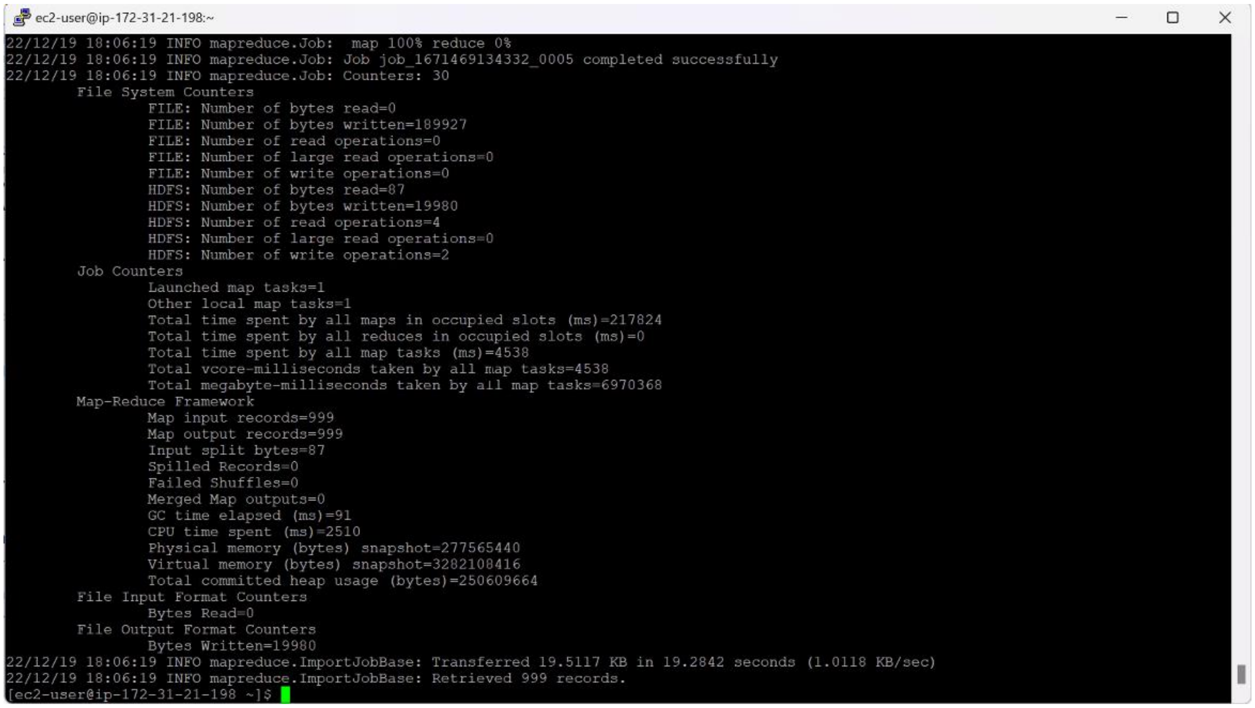
Data Ingestion from the RDS to HDFS using Sqoop

1. Running Sqoop command to import "member_score" table from RDS to HDFS.

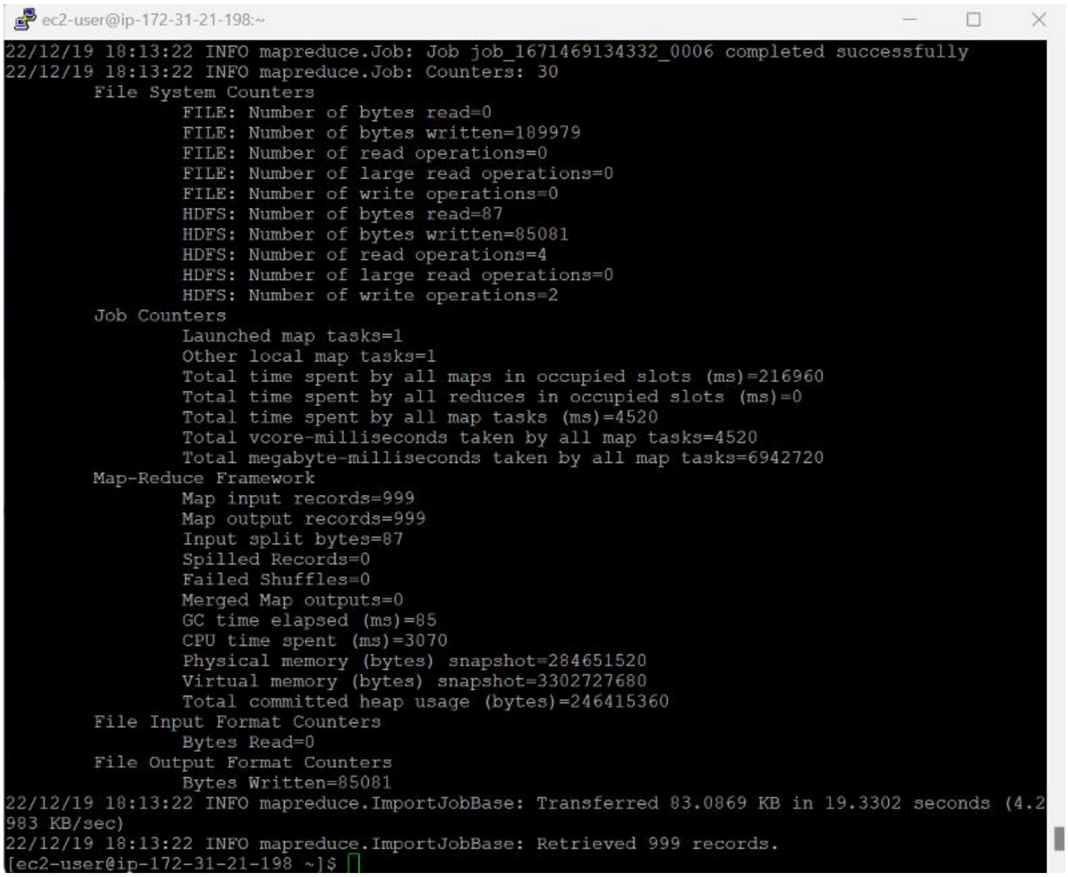
```
sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
--username upgraduser \
--password upgraduser \
--table member_score \
--null-string 'NA' \
--null-non-string '\\N' \
--delete-target-dir \
--target-dir '/user/hadoop/upgrad_ccfraud/card_transactions/member_score' \
-m 1
```



```
22/12/19 18:06:19 INFO mapreduce.Job:  map 100% reduce 0%
22/12/19 18:06:19 INFO mapreduce.Job: Job job_1671469134332_0005 completed successfully
22/12/19 18:06:19 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189927
                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)=217824
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4538
                Total vcore-milliseconds taken by all map tasks=4538
                Total megabyte-milliseconds taken by all map tasks=6970368
        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)=91
                CPU time spent (ms)=2510
                Physical memory (bytes) snapshot=277565440
                Virtual memory (bytes) snapshot=3282108416
                Total committed heap usage (bytes)=250609664
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
22/12/19 18:06:19 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 19.2842 seconds (1.0118 KB/sec)
22/12/19 18:06:19 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-21-198 ~]$
```

2. Running Sqoop command to import “**card_member**” table from RDS to HDFS.

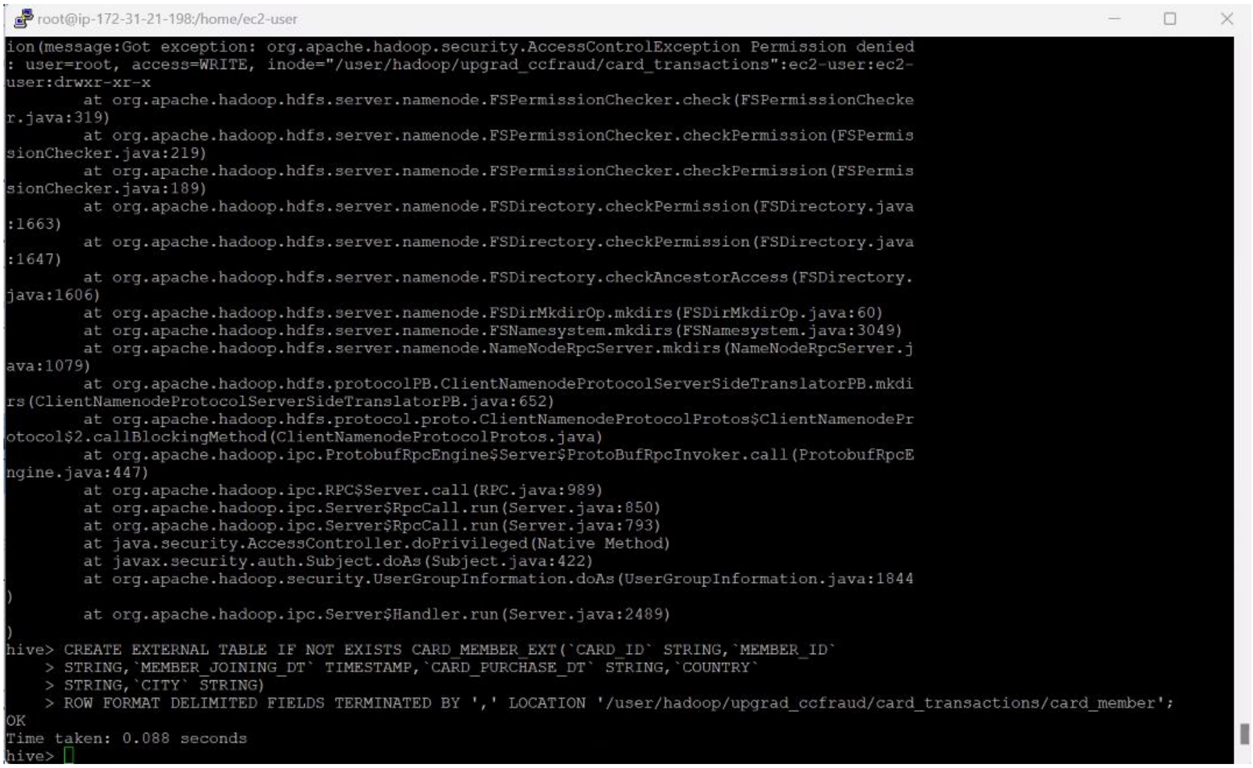
```
sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
--username upgraduser \
--password upgraduser \
--table card_member \
--null-string 'NA' \
--null-non-string '\\N' \
--delete-target-dir \
--target-dir '/user/hadoop/upgrad_ccfraud/card_transactions/card_member' \
-m 1
```



```
22/12/19 18:13:22 INFO mapreduce.Job: Job job_1671469134332_0006 completed successfully
22/12/19 18:13:22 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189979
                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)=216960
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4520
                Total vcore-milliseconds taken by all map tasks=4520
                Total megabyte-milliseconds taken by all map tasks=6942720
        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)=3070
                Physical memory (bytes) snapshot=284651520
                Virtual memory (bytes) snapshot=3302727680
                Total committed heap usage (bytes)=246415360
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85081
22/12/19 18:13:22 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 19.3302 seconds (4.2
983 KB/sec)
22/12/19 18:13:22 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-21-198 ~]$ 
```

3. Starting hive session and creating an external table "**card_member_ext**" to hold data from card_member table in RDS.

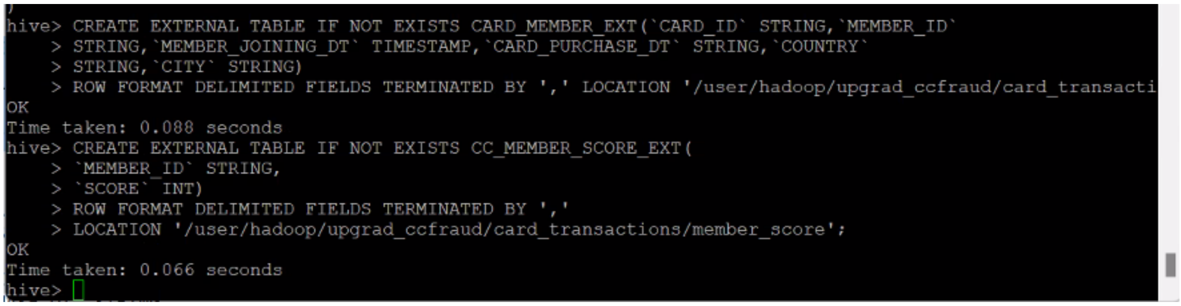
CREATE EXTERNAL TABLE IF NOT EXISTS CARD_MEMBER_EXT(`CARD_ID` STRING,`MEMBER_ID` STRING,`MEMBER_JOINING_DT` TIMESTAMP,`CARD_PURCHASE_DT` STRING,`COUNTRY` STRING,`CITY` STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hadoop/upgrad_ccfraud/card_transactions/card_member';



```
ion(message:Got exception: org.apache.hadoop.security.AccessControlException Permission denied
: user=root, access=WRITE, inode="/user/hadoop/upgrad_ccfraud/card_transactions":ec2-user:ec2-
user:drwxr-xr-x
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecke
r.java:319)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermis
sionChecker.java:219)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermis
sionChecker.java:189)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java
:1663)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java
:1647)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkAncestorAccess(FSDirectory.
java:1606)
        at org.apache.hadoop.hdfs.server.namenode.FSDirMkdirOp.mkdirs(FSDirMkdirOp.java:60)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.mkdirs(FSNamesystem.java:3049)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.mkdirs(NameNodeRpcServer.j
ava:1079)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.mkdi
rs(ClientNamenodeProtocolServerSideTranslatorPB.java:652)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodePr
otocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcE
ngine.java:447)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:989)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:850)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:793)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844
)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2489)
)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CARD_MEMBER_EXT(`CARD_ID` STRING,`MEMBER_ID`
    > STRING,`MEMBER_JOINING_DT` TIMESTAMP,`CARD_PURCHASE_DT` STRING,`COUNTRY`
    > STRING,`CITY` STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hadoop/upgrad_ccfraud/card_transactions/card_member';
OK
Time taken: 0.088 seconds
hive>
```

4. Create external table "**CC_MEMBER_SCORE_EXT** " to hold data from member_score table in RDS.

CREATE EXTERNAL TABLE IF NOT EXISTS CC_MEMBER_SCORE_EXT(
`MEMBER_ID` STRING,
`SCORE` INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION '/user/hadoop/upgrad_ccfraud/card_transactions/member_score';

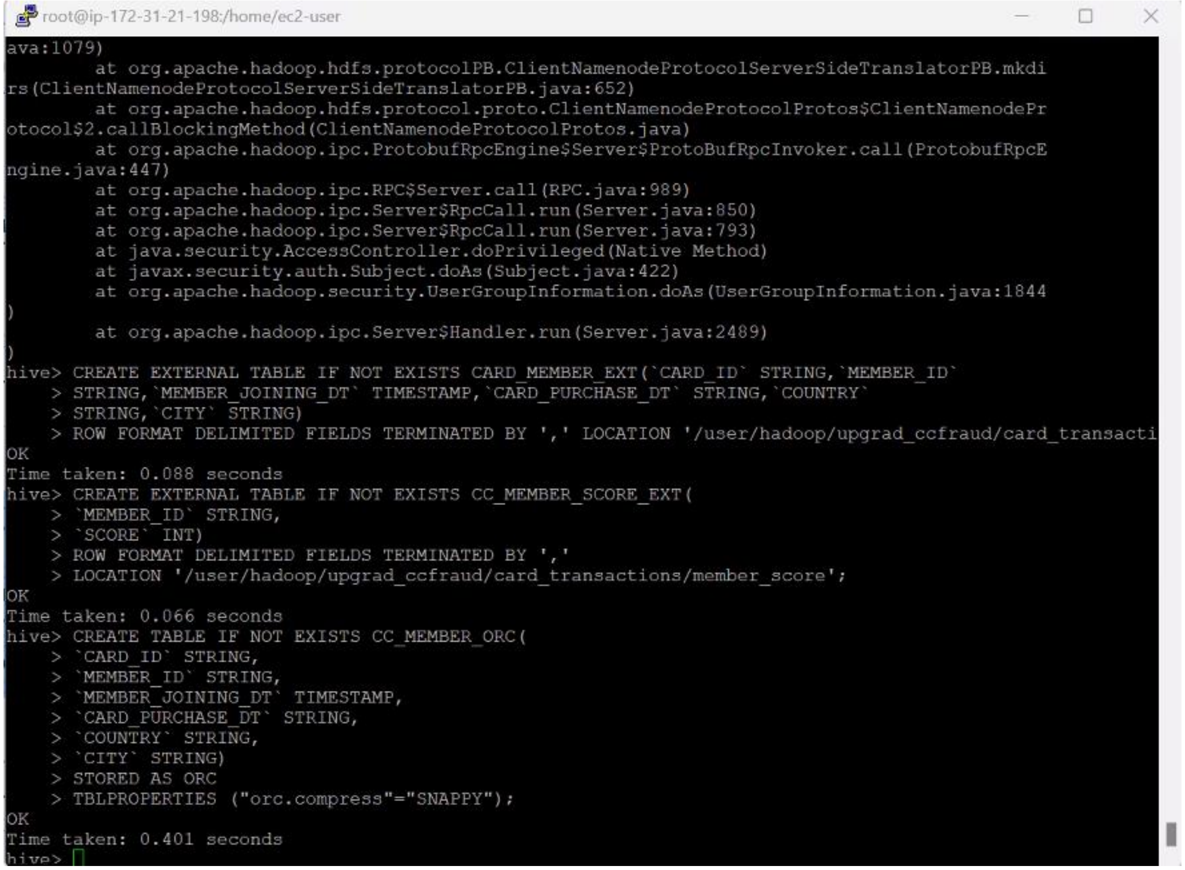


```
)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CARD_MEMBER_EXT(`CARD_ID` STRING,`MEMBER_ID`
    > STRING,`MEMBER_JOINING_DT` TIMESTAMP,`CARD_PURCHASE_DT` STRING,`COUNTRY`
    > STRING,`CITY` STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hadoop/upgrad_ccfraud/card_transacti
OK
Time taken: 0.088 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CC_MEMBER_SCORE_EXT(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/user/hadoop/upgrad_ccfraud/card_transactions/member_score';
OK
Time taken: 0.066 seconds
hive>
```

5. Create " **CC_MEMBER_ORC**" table with ORC for better performance.

CREATE TABLE IF NOT EXISTS CC_MEMBER_ORC(
`CARD_ID` STRING,
`MEMBER_ID` STRING,
`MEMBER_JOINING_DT` TIMESTAMP,
`CARD_PURCHASE_DT` STRING,
`COUNTRY` STRING,
`CITY` STRING)

STORED AS ORC
TBLPROPERTIES ("orc.compress"="SNAPPY");

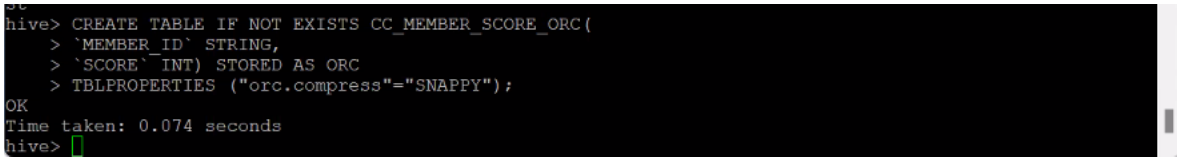


```
ava:1079)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.mkdi
rs(ClientNamenodeProtocolServerSideTranslatorPB.java:652)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodePr
otocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcE
ngine.java:447)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:989)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:850)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:793)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844
)

        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2489)
)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CARD_MEMBER_EXT(`CARD_ID` STRING,`MEMBER_ID`
    > STRING,`MEMBER_JOINING_DT` TIMESTAMP,`CARD_PURCHASE_DT` STRING,`COUNTRY`
    > STRING,`CITY` STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hadoop/upgrad_ccfraud/card_transacti
OK
Time taken: 0.088 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CC_MEMBER_SCORE_EXT(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/user/hadoop/upgrad_ccfraud/card_transactions/member_score';
OK
Time taken: 0.066 seconds
hive> CREATE TABLE IF NOT EXISTS CC_MEMBER_ORC(
    > `CARD_ID` STRING,
    > `MEMBER_ID` STRING,
    > `MEMBER_JOINING_DT` TIMESTAMP,
    > `CARD_PURCHASE_DT` STRING,
    > `COUNTRY` STRING,
    > `CITY` STRING)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.401 seconds
hive>
```

6. Create " **CC_MEMBER_SCORE_ORC** " table with ORC for better performance.

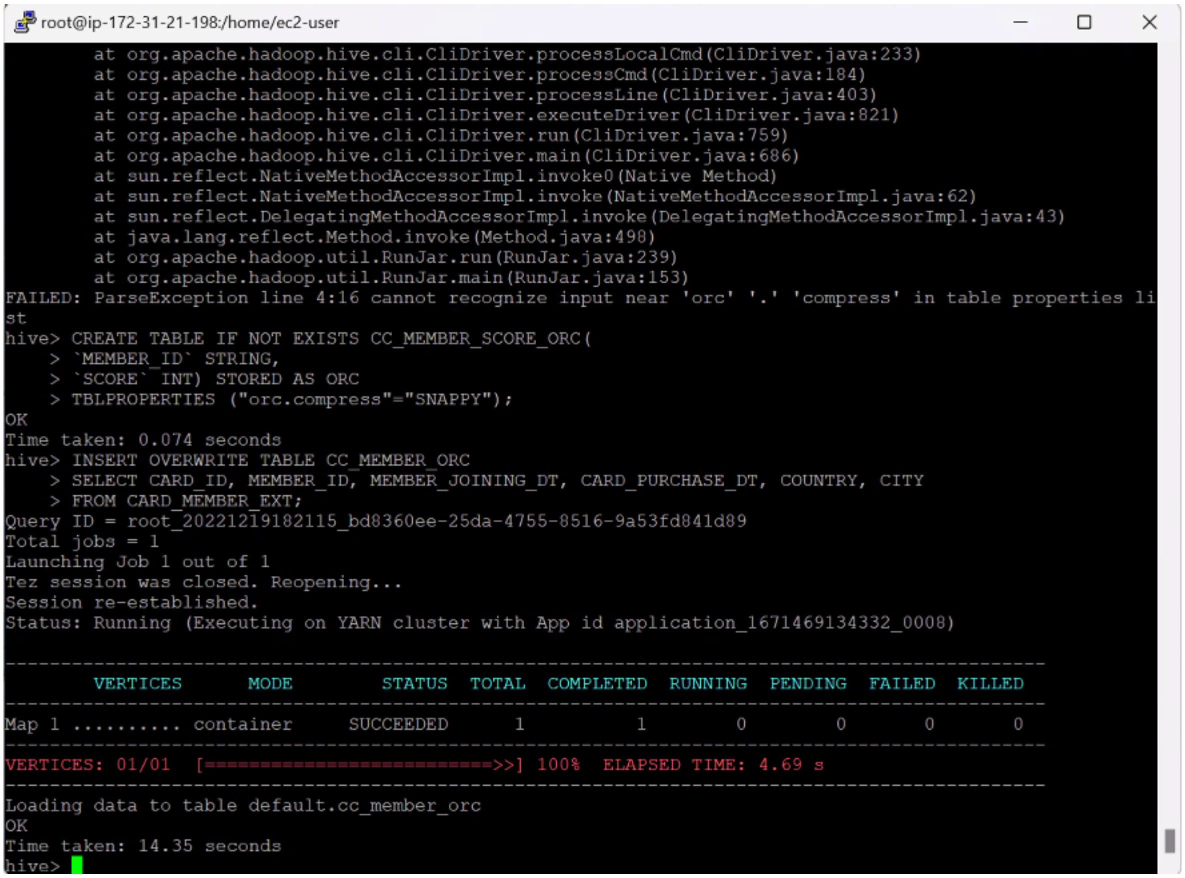
   CREATE TABLE IF NOT EXISTS CC_MEMBER_SCORE_ORC(
   `MEMBER_ID` STRING,
   `SCORE` INT) STORED AS ORC
   TBLPROPERTIES ("orc.compress"="SNAPPY");



```
hive> CREATE TABLE IF NOT EXISTS CC_MEMBER_SCORE_ORC(
    > `MEMBER_ID` STRING,
    > `SCORE` INT) STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.074 seconds
hive>
```

7. Load data into " **CC_MEMBER_ORC** " table from "**card_member_ext**" table.

   INSERT OVERWRITE TABLE CC_MEMBER_ORC
   SELECT CARD_ID, MEMBER_ID, MEMBER_JOINING_DT, CARD_PURCHASE_DT, COUNTRY, CITY
   FROM CARD_MEMBER_EXT;

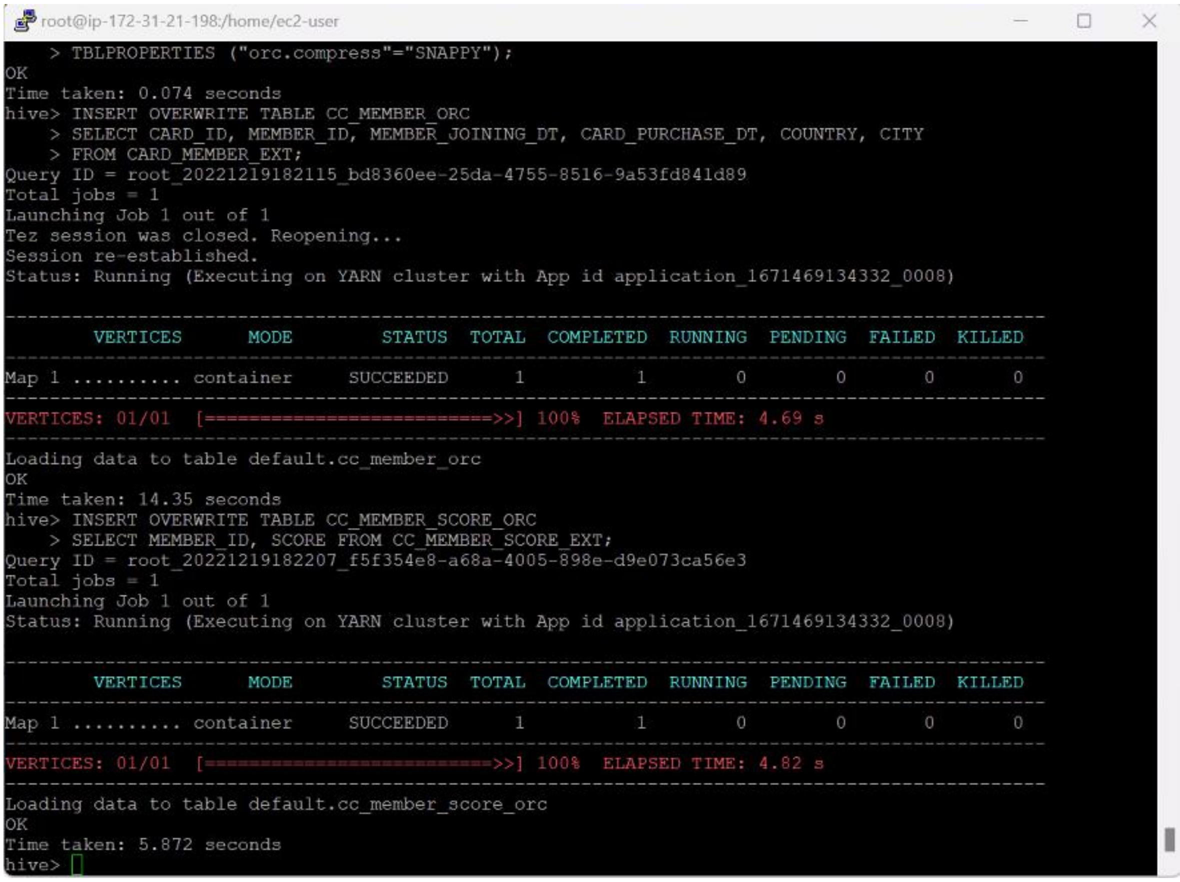


```
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 4:16 cannot recognize input near 'orc' '.' 'compress' in table properties li
st
hive> CREATE TABLE IF NOT EXISTS CC_MEMBER_SCORE_ORC(
    > `MEMBER_ID` STRING,
    > `SCORE` INT) STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.074 seconds
hive> INSERT OVERWRITE TABLE CC_MEMBER_ORC
    > SELECT CARD_ID, MEMBER_ID, MEMBER_JOINING_DT, CARD_PURCHASE_DT, COUNTRY, CITY
    > FROM CARD_MEMBER_EXT;
Query ID = root_20221219182115_bd8360ee-25da-4755-8516-9a53fd841d89
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_1671469134332_0008)

----------------------------------------------------------------------------------------------
        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: 4.69 s
----------------------------------------------------------------------------------------------
Loading data to table default.cc_member_orc
OK
Time taken: 14.35 seconds
hive> 
```

8. Load data into “**CC_MEMBER_SCORE_ORC**” table from “**CC_MEMBER_SCORE_EXT**” table.

INSERT OVERWRITE TABLE CC_MEMBER_SCORE_ORC
SELECT MEMBER_ID, SCORE FROM CC_MEMBER_SCORE_EXT;



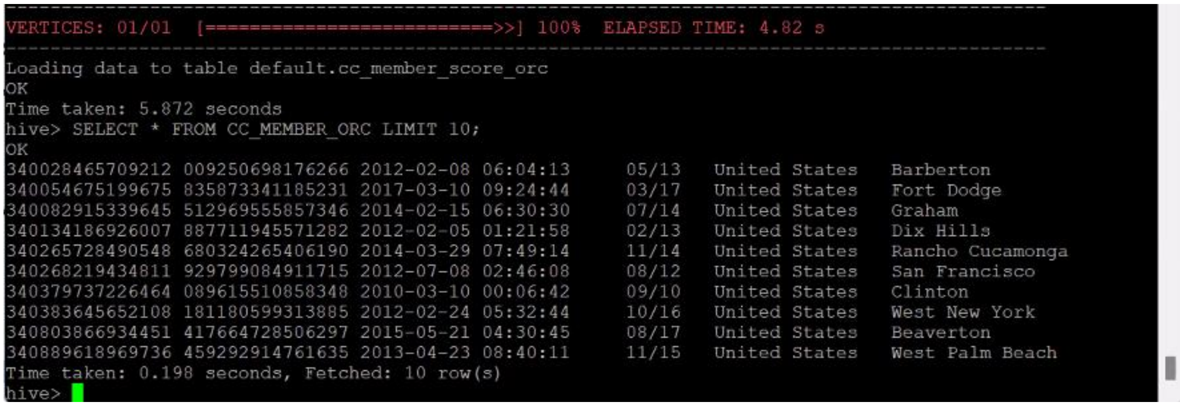
```
root@ip-172-31-21-198:/home/ec2-user
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.074 seconds
hive> INSERT OVERWRITE TABLE CC_MEMBER_ORC
    > SELECT CARD_ID, MEMBER_ID, MEMBER_JOINING_DT, CARD_PURCHASE_DT, COUNTRY, CITY
    > FROM CARD_MEMBER_EXT;
Query ID = root_20221219182115_bd8360ee-25da-4755-8516-9a53fd841d89
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_1671469134332_0008)

----------------------------------------------------------------------------------------------
        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: 4.69 s
----------------------------------------------------------------------------------------------
Loading data to table default.cc_member_orc
OK
Time taken: 14.35 seconds
hive> INSERT OVERWRITE TABLE CC_MEMBER_SCORE_ORC
    > SELECT MEMBER_ID, SCORE FROM CC_MEMBER_SCORE_EXT;
Query ID = root_20221219182207_f5f354e8-a68a-4005-898e-d9e073ca56e3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0008)

----------------------------------------------------------------------------------------------
        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: 4.82 s
----------------------------------------------------------------------------------------------
Loading data to table default.cc_member_score_orc
OK
Time taken: 5.872 seconds
hive> 
```

9. Verify data in “**CC_MEMBER_ORC** " table.

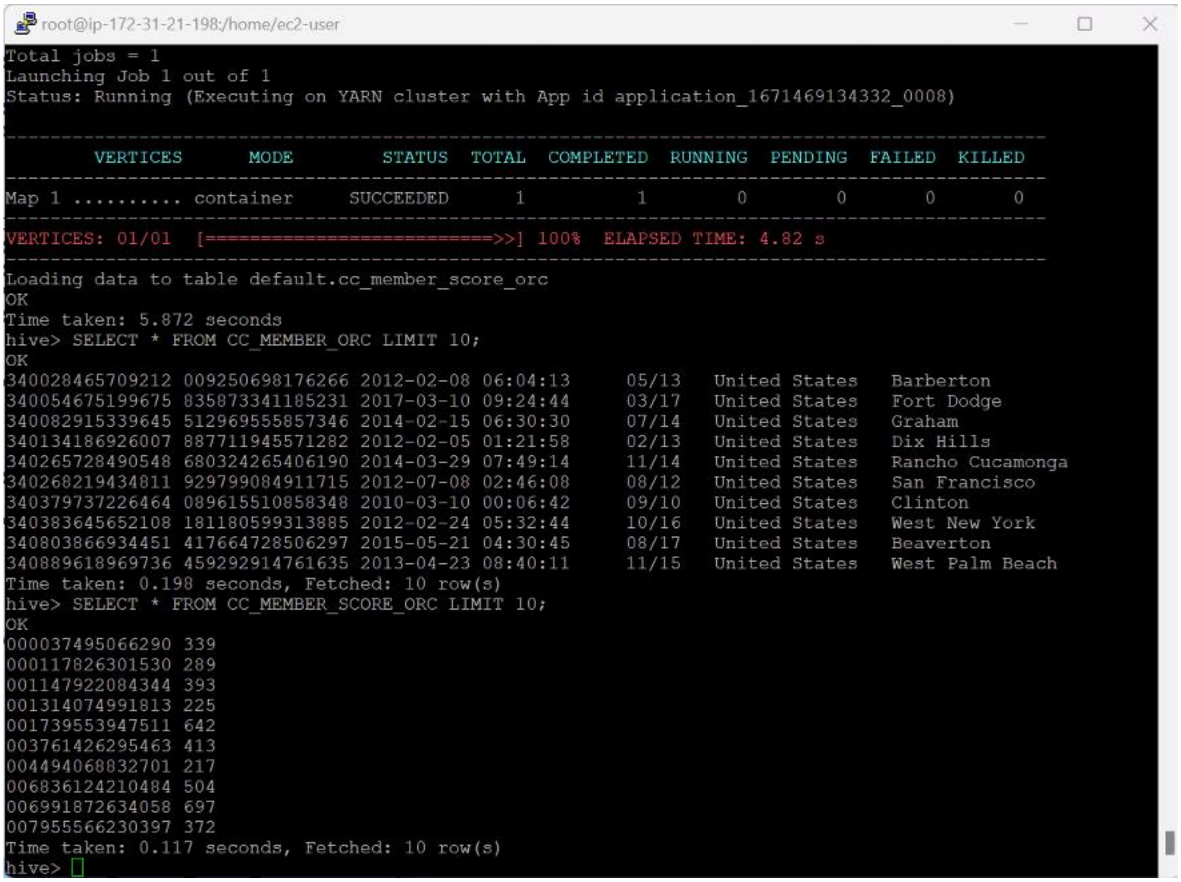
SELECT * FROM CC_MEMBER_ORC LIMIT 10;



```
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.82 s
----------------------------------------------------------------------------------------------
Loading data to table default.cc_member_score_orc
OK
Time taken: 5.872 seconds
hive> SELECT * FROM CC_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: 0.198 seconds, Fetched: 10 row(s)
hive> 
```

10. Verify data in "**CC_MEMBER_SCORE_ORC**" table.

SELECT * FROM CC_MEMBER_SCORE_ORC LIMIT 10;



```
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0008)

----------------------------------------------------------------------------------------------
        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: 4.82 s
----------------------------------------------------------------------------------------------
Loading data to table default.cc_member_score_orc
OK
Time taken: 5.872 seconds
hive> SELECT * FROM CC_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: 0.198 seconds, Fetched: 10 row(s)
hive> SELECT * FROM CC_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: 0.117 seconds, Fetched: 10 row(s)
hive> 
```