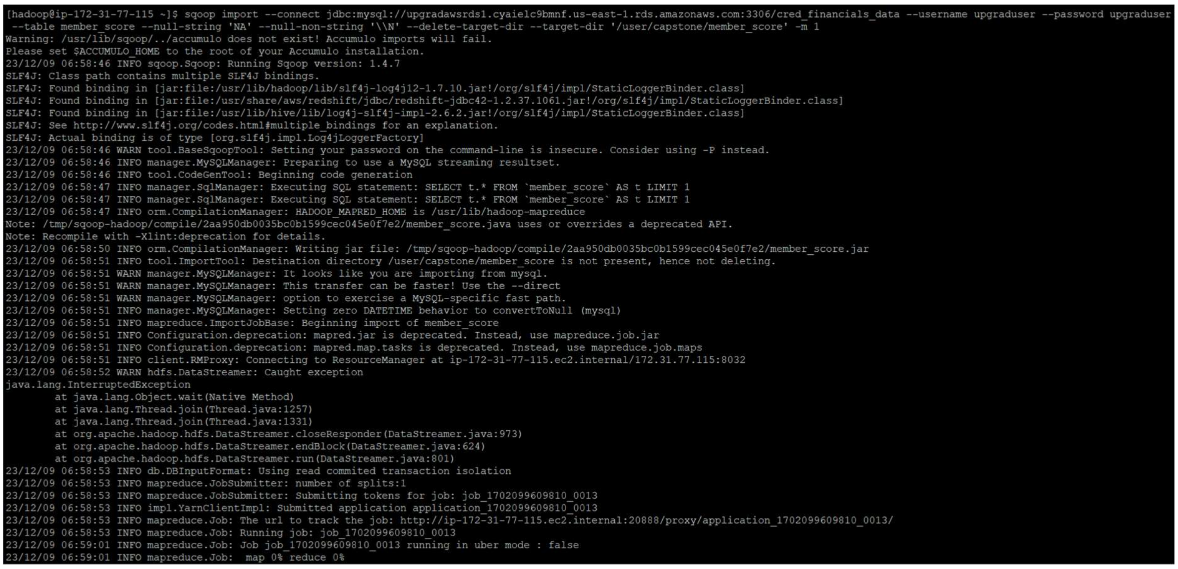
Data Ingestion from the RDS to HDFS using Sqoop

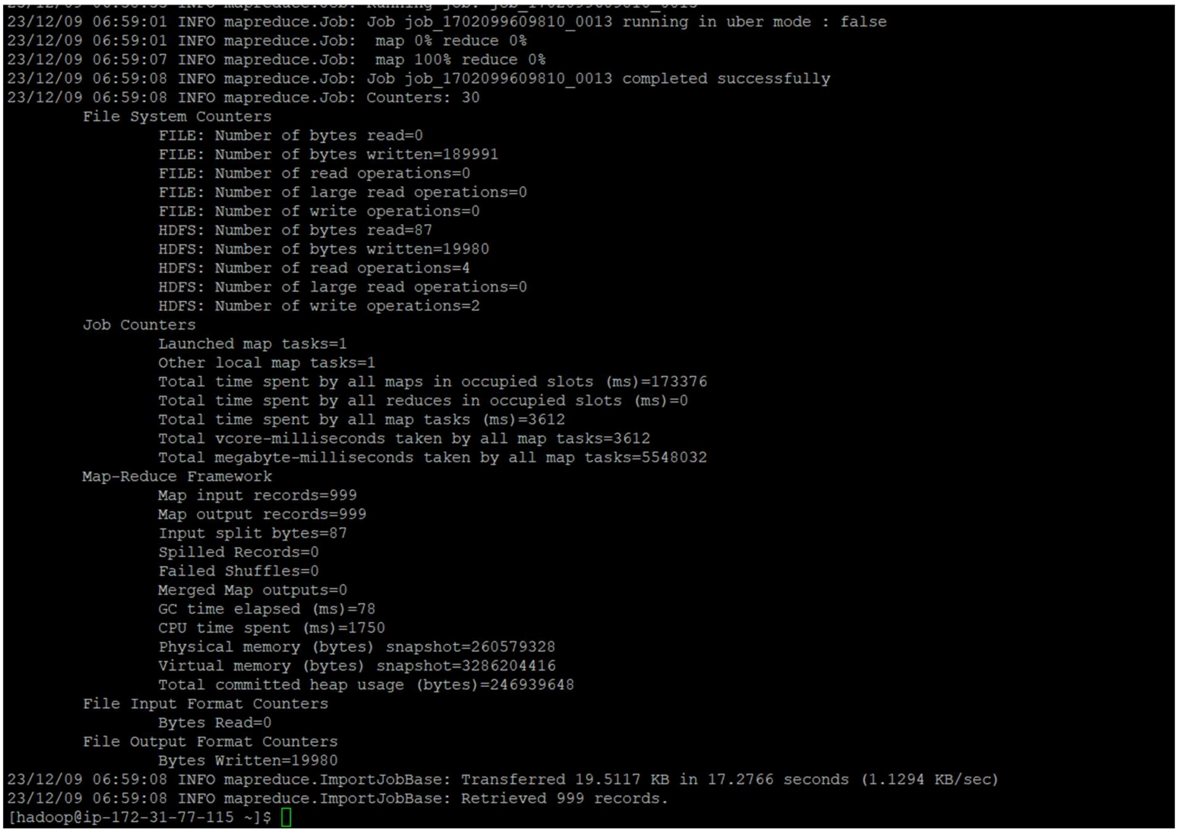
Sqoop command used for importing table from RDS to HDFS

- Command to create prerequisite HDFS directories
  - *hadoop fs -mkdir /user/capstone/card_member*
  - *hadoop fs -mkdir /user/capstone/member_score*
- Setting Hive parameters (observed performance improvements by this)
  - *Set Hive parameters:*
  - *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;*
- Sqoop Command to import member score data

  sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/cred_financials_data --username upgraduser --password upgraduser --table member_score --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/user/capstone/member_score' -m 1



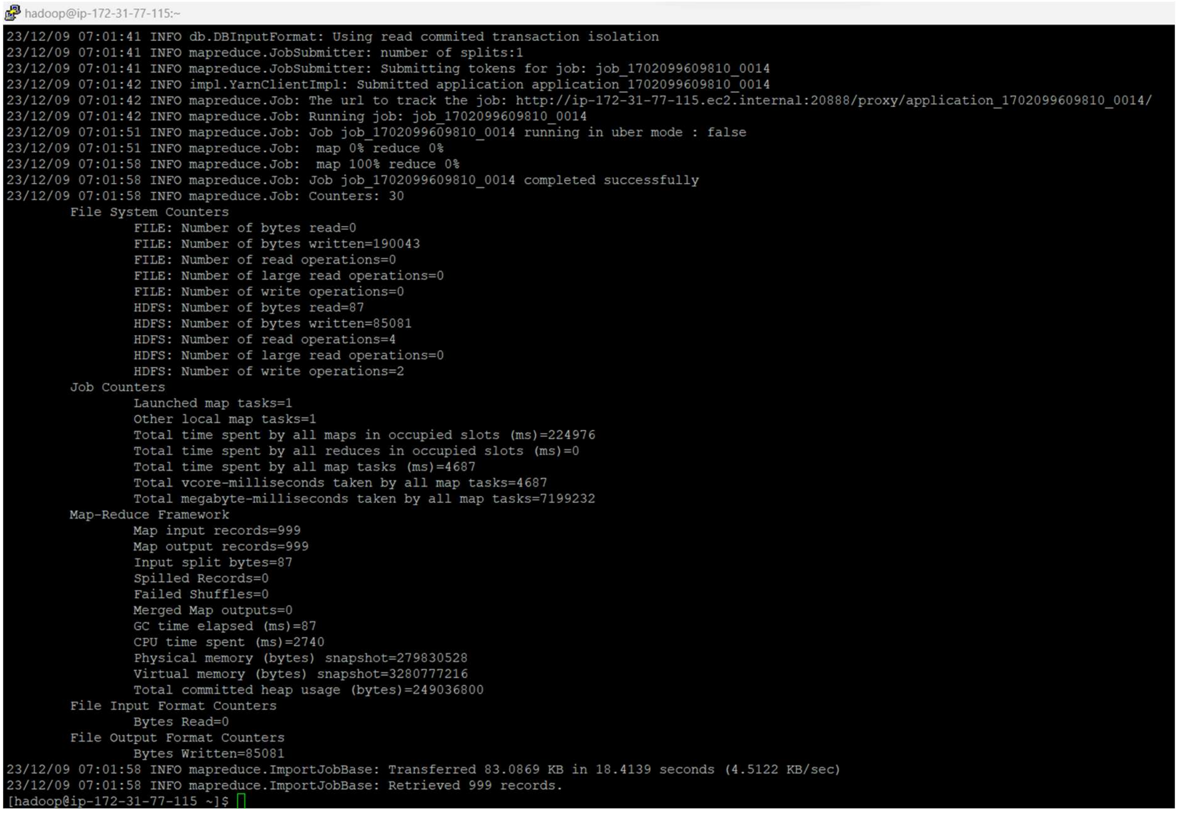
```
[hadoop@ip-172-31-77-115 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/cred_financials_data --username upgraduser --password upgraduser
 --table member_score --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/user/capstone/member_score' -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.
23/12/09 06:58:46 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]
23/12/09 06:58:46 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/12/09 06:58:46 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/12/09 06:58:46 INFO tool.CodeGenTool: Beginning code generation
23/12/09 06:58:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/12/09 06:58:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/12/09 06:58:47 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/2aa950db0035bc0b1599cec045e0f7e2/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/12/09 06:58:50 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/2aa950db0035bc0b1599cec045e0f7e2/member_score.jar
23/12/09 06:58:51 INFO tool.ImportTool: Destination directory /user/capstone/member_score is not present, hence not deleting.
23/12/09 06:58:51 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/12/09 06:58:51 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/12/09 06:58:51 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/12/09 06:58:51 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/12/09 06:58:51 INFO mapreduce.ImportJobBase: Beginning import of member_score
23/12/09 06:58:51 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/12/09 06:58:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/12/09 06:58:51 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-77-115.ec2.internal/172.31.77.115:8032
23/12/09 06:58:52 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1257)
        at java.lang.Thread.join(Thread.java:1331)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:973)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:624)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:801)
23/12/09 06:58:53 INFO db.DBInputFormat: Using read commited transaction isolation
23/12/09 06:58:53 INFO mapreduce.JobSubmitter: number of splits:1
23/12/09 06:58:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1702099609810_0013
23/12/09 06:58:53 INFO impl.YarnClientImpl: Submitted application application_1702099609810_0013
23/12/09 06:58:53 INFO mapreduce.Job: The url to track the job: http://ip-172-31-77-115.ec2.internal:20888/proxy/application_1702099609810_0013/
23/12/09 06:58:53 INFO mapreduce.Job: Running job: job_1702099609810_0013
23/12/09 06:59:01 INFO mapreduce.Job: Job job_1702099609810_0013 running in uber mode : false
23/12/09 06:59:01 INFO mapreduce.Job:  map 0% reduce 0%
```



```
23/12/09 06:59:01 INFO mapreduce.Job: Job job_1702099609810_0013 running in uber mode : false
23/12/09 06:59:01 INFO mapreduce.Job:  map 0% reduce 0%
23/12/09 06:59:07 INFO mapreduce.Job:  map 100% reduce 0%
23/12/09 06:59:08 INFO mapreduce.Job: Job job_1702099609810_0013 completed successfully
23/12/09 06:59:08 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189991
                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)=173376
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3612
                Total vcore-milliseconds taken by all map tasks=3612
                Total megabyte-milliseconds taken by all map tasks=5548032
        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)=78
                CPU time spent (ms)=1750
                Physical memory (bytes) snapshot=260579328
                Virtual memory (bytes) snapshot=3286204416
                Total committed heap usage (bytes)=246939648
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
23/12/09 06:59:08 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 17.2766 seconds (1.1294 KB/sec)
23/12/09 06:59:08 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-77-115 ~]$
```

- Sqoop command to import card member data

  ```
  sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/cred_financials_data --username upgraduser --password upgraduser --table card_member --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/user/capstone/card_member'  -m 1
  ```



```
hadoop@ip-172-31-77-115:~
23/12/09 07:01:41 INFO db.DBInputFormat: Using read commited transaction isolation
23/12/09 07:01:41 INFO mapreduce.JobSubmitter: number of splits:1
23/12/09 07:01:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1702099609810_0014
23/12/09 07:01:42 INFO impl.YarnClientImpl: Submitted application application_1702099609810_0014
23/12/09 07:01:42 INFO mapreduce.Job: The url to track the job: http://ip-172-31-77-115.ec2.internal:20888/proxy/application_1702099609810_0014/
23/12/09 07:01:42 INFO mapreduce.Job: Running job: job_1702099609810_0014
23/12/09 07:01:51 INFO mapreduce.Job: Job job_1702099609810_0014 running in uber mode : false
23/12/09 07:01:51 INFO mapreduce.Job:  map 0% reduce 0%
23/12/09 07:01:58 INFO mapreduce.Job:  map 100% reduce 0%
23/12/09 07:01:58 INFO mapreduce.Job: Job job_1702099609810_0014 completed successfully
23/12/09 07:01:58 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=190043
                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)=224976
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4687
                Total vcore-milliseconds taken by all map tasks=4687
                Total megabyte-milliseconds taken by all map tasks=7199232
        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)=87
                CPU time spent (ms)=2740
                Physical memory (bytes) snapshot=279830528
                Virtual memory (bytes) snapshot=3280777216
                Total committed heap usage (bytes)=249036800
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85081
23/12/09 07:01:58 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 18.4139 seconds (4.5122 KB/sec)
23/12/09 07:01:58 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-77-115 ~]$
```

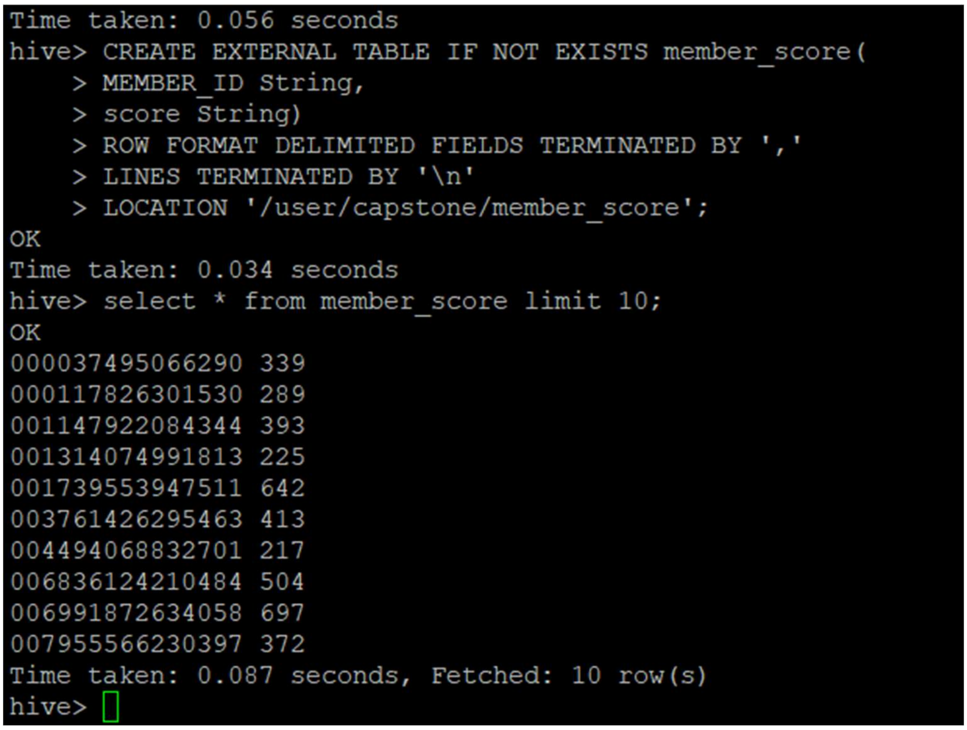
- Hadoop command used to see imported data
  - *hadoop fs -ls /user/capstone/member_score*
  - *hadoop fs -ls /user/capstone/card_member*

Screenshot of the imported data

Result -> Above command shows Retrieved 999 records.

- Create Hive tables for above file:

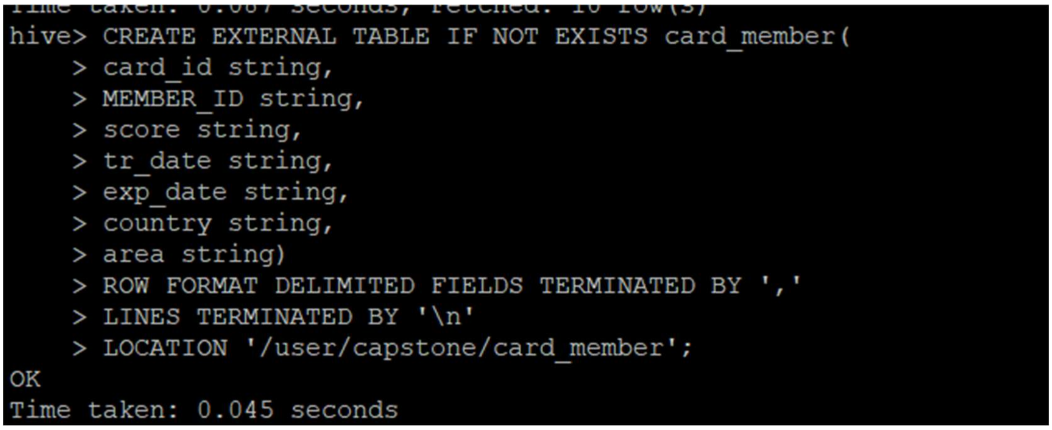
CREATE EXTERNAL TABLE IF NOT EXISTS member_score(
MEMBER_ID String,
score String)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
LOCATION '/user/capstone/member_score';



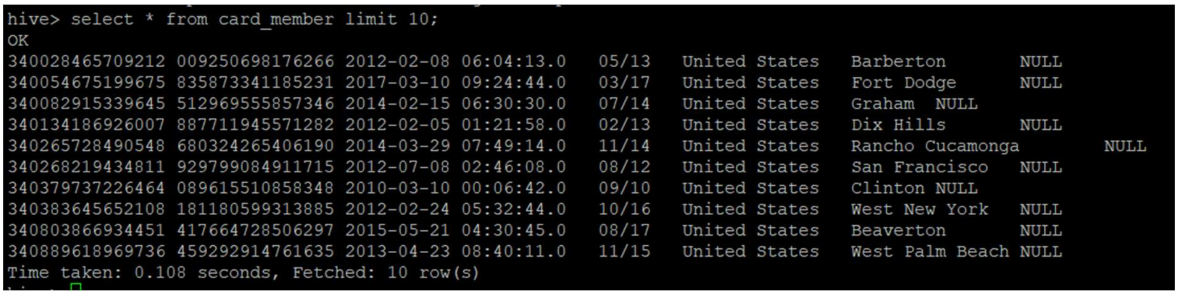
```
Time taken: 0.056 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS member_score(
    > MEMBER_ID String,
    > score String)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > LOCATION '/user/capstone/member_score';
OK
Time taken: 0.034 seconds
hive> select * from member_score 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: 0.087 seconds, Fetched: 10 row(s)
hive> 
```

- Create Hive table for card_member:

CREATE EXTERNAL TABLE IF NOT EXISTS card_member(
card_id string,
MEMBER_ID string,
score string,
tr_date string,
exp_date string,
country string,
area string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
LOCATION '/user/capstone/card_member';



```
Time taken: 0.087 seconds, Fetched: 10 row(s)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS card_member(
    > card_id string,
    > MEMBER_ID string,
    > score string,
    > tr_date string,
    > exp_date string,
    > country string,
    > area string)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > LOCATION '/user/capstone/card_member';
OK
Time taken: 0.045 seconds
```



```
hive> select * from card_member limit 10;
OK
340028465709212 009250698176266 2012-02-08 06:04:13.0   05/13   United States   Barberton       NULL
340054675199675 835873341185231 2017-03-10 09:24:44.0   03/17   United States   Fort Dodge      NULL
340082915339645 512969555857346 2014-02-15 06:30:30.0   07/14   United States   Graham  NULL
340134186926007 887711945571282 2012-02-05 01:21:58.0   02/13   United States   Dix Hills       NULL
340265728490548 680324265406190 2014-03-29 07:49:14.0   11/14   United States   Rancho Cucamonga        NULL
340268219434811 929799084911715 2012-07-08 02:46:08.0   08/12   United States   San Francisco   NULL
340379737226464 089615510858348 2010-03-10 00:06:42.0   09/10   United States   Clinton NULL
340383645652108 181180599313885 2012-02-24 05:32:44.0   10/16   United States   West New York   NULL
340803866934451 417664728506297 2015-05-21 04:30:45.0   08/17   United States   Beaverton       NULL
340889618969736 459292914761635 2013-04-23 08:40:11.0   11/15   United States   West Palm Beach NULL
Time taken: 0.108 seconds, Fetched: 10 row(s)
```