Data Ingestion from the RDS to HDFS using Sqoop

1. Sqoop command used for importing table from RDS to HDFS

- Run 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 '/ccfd_capstone_project/member_score' \
-m 1
```

- Run 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 '/ccfd_capstone_project/card_member' \
-m 1
```

2. <Command to see the list of imported data in HDFS>

- Create 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
'/ccfd_capstone_project/card_member';
```

- Create external table "**member_score_ext**" to hold data from member_score table in RDS.

  ```
  CREATE EXTERNAL TABLE IF NOT EXISTS MEMBER_SCORE_EXT(
  `MEMBER_ID` STRING,
  `SCORE` INT)
  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
  LOCATION '/ccfd_capstone_project/member_score';
  ```

- Create "**card_member_orc**" table. For better performance.

  ```
  CREATE TABLE IF NOT EXISTS CARD_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");
  ```

- Create "**member_score_orc**" table. For better performance.

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

- Load data into "**card_member_orc**" table from "**card_member_ext**" table.

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

- Load data into “**member_score_orc**” table from “**member_score_ext**” table.

  ```
  INSERT OVERWRITE TABLE MEMBER_SCORE_ORC
  SELECT MEMBER_ID, SCORE FROM MEMBER_SCORE_EXT;
  ```

- Verify data in “**card_member_orc**" table.

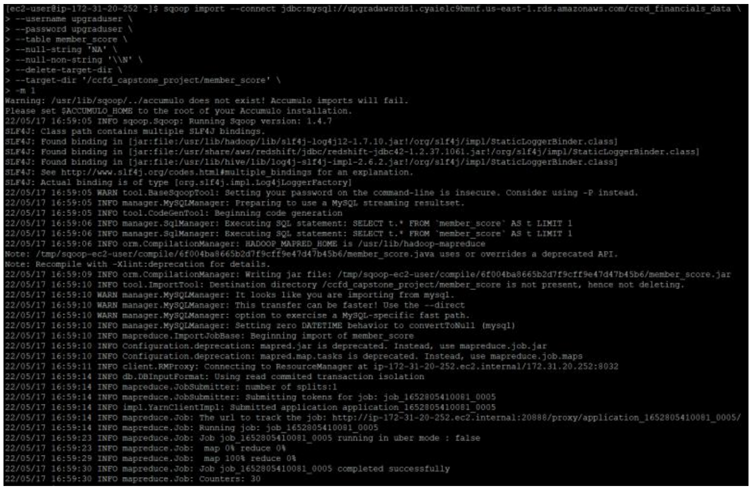
  ```
  SELECT * FROM CARD_MEMBER_ORC LIMIT 10;
  ```

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

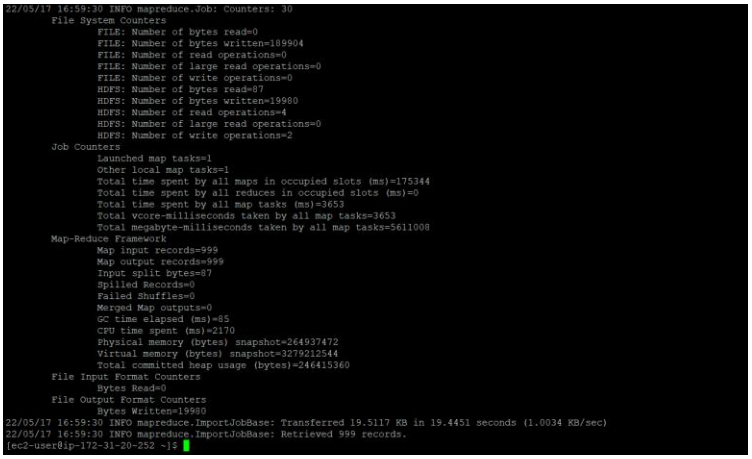
  ```
  SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;
  ```

3. Screenshot of the imported data

- Run Sqoop command to import "**member_score**" table from RDS to HDFS.

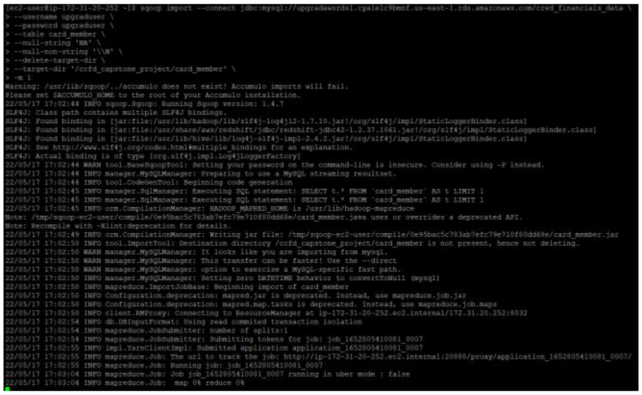


```
[ec2-user@ip-172-31-20-252 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table member_score \
> --null-string 'NA' \
> --null-non-string '\\N' \
> --delete-target-dir \
> --target-dir '/ccfd_capstone_project/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.
22/05/17 16:59:05 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]
22/05/17 16:59:05 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/17 16:59:05 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/17 16:59:05 INFO tool.CodeGenTool: Beginning code generation
22/05/17 16:59:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/17 16:59:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/17 16:59:06 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/6f004ba8665b2d7f9cff9e47d47b45b6/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/17 16:59:09 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/6f004ba8665b2d7f9cff9e47d47b45b6/member_score.jar
22/05/17 16:59:10 INFO tool.ImportTool: Destination directory /ccfd_capstone_project/member_score is not present, hence not deleting.
22/05/17 16:59:10 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/17 16:59:10 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/17 16:59:10 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/17 16:59:10 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/17 16:59:10 INFO mapreduce.ImportJobBase: Beginning import of member_score
22/05/17 16:59:10 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/17 16:59:10 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/17 16:59:11 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-20-252.ec2.internal/172.31.20.252:8032
22/05/17 16:59:14 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/17 16:59:14 INFO mapreduce.JobSubmitter: number of splits:1
22/05/17 16:59:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652805410081_0005
22/05/17 16:59:14 INFO impl.YarnClientImpl: Submitted application application_1652805410081_0005
22/05/17 16:59:14 INFO mapreduce.Job: The url to track the job: http://ip-172-31-20-252.ec2.internal:20888/proxy/application_1652805410081_0005/
22/05/17 16:59:14 INFO mapreduce.Job: Running job: job_1652805410081_0005
22/05/17 16:59:23 INFO mapreduce.Job: Job job_1652805410081_0005 running in uber mode : false
22/05/17 16:59:23 INFO mapreduce.Job:  map 0% reduce 0%
22/05/17 16:59:29 INFO mapreduce.Job:  map 100% reduce 0%
22/05/17 16:59:30 INFO mapreduce.Job: Job job_1652805410081_0005 completed successfully
22/05/17 16:59:30 INFO mapreduce.Job: Counters: 30
```

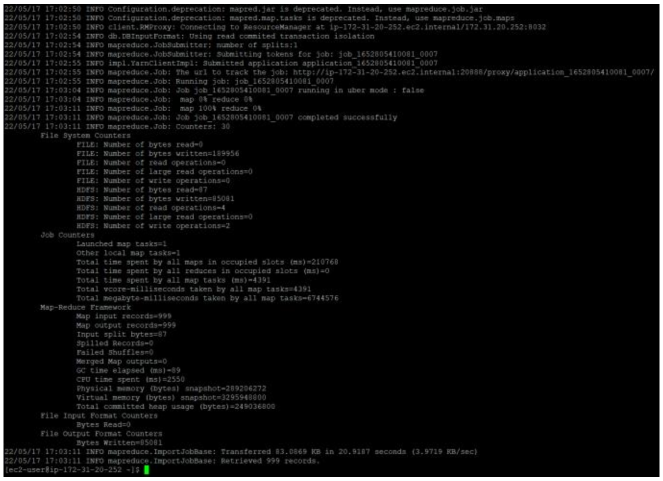


```
22/05/17 16:59:30 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189904
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                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)=175344
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3653
                Total vcore-milliseconds taken by all map tasks=3653
                Total megabyte-milliseconds taken by all map tasks=5611008
        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)=2170
                Physical memory (bytes) snapshot=264937472
                Virtual memory (bytes) snapshot=3279212544
                Total committed heap usage (bytes)=246415360
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
22/05/17 16:59:30 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 19.4451 seconds (1.0034 KB/sec)
22/05/17 16:59:30 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-20-252 ~]$
```

- Run Sqoop command to import "**card_member**" table from RDS to HDFS.

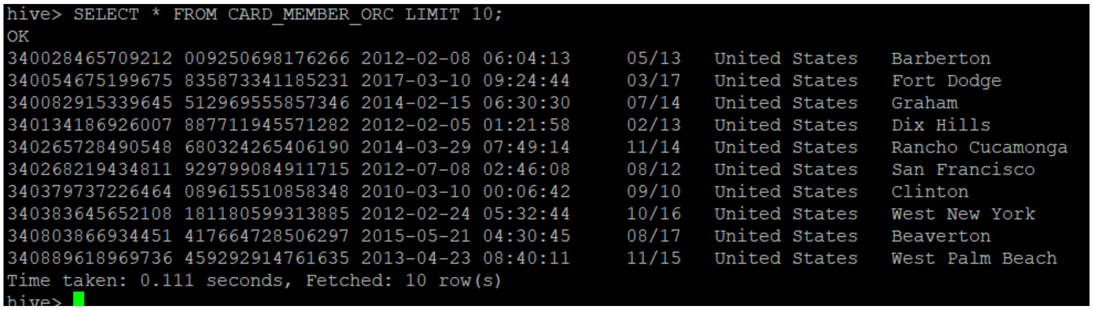


```
[ec2-user@ip-172-31-20-252 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table card_member \
> --null-string 'NA' \
> --null-non-string '\\N' \
> --delete-target-dir \
> --target-dir '/ccfd_capstone_project/card_member' \
> -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.
22/05/17 17:02:44 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]
22/05/17 17:02:44 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/17 17:02:44 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/17 17:02:44 INFO tool.CodeGenTool: Beginning code generation
22/05/17 17:02:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/17 17:02:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/17 17:02:45 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/0e95bac5c783ab7efc79e710f80dd68e/card_member.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/17 17:02:49 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/0e95bac5c783ab7efc79e710f80dd68e/card_member.jar
22/05/17 17:02:50 INFO tool.ImportTool: Destination directory /ccfd_capstone_project/card_member is not present, hence not deleting.
22/05/17 17:02:50 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/17 17:02:50 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/17 17:02:50 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/17 17:02:50 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/17 17:02:50 INFO mapreduce.ImportJobBase: Beginning import of card_member
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/17 17:02:50 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-20-252.ec2.internal/172.31.20.252:8032
22/05/17 17:02:54 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: number of splits:1
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652805410081_0007
22/05/17 17:02:55 INFO impl.YarnClientImpl: Submitted application application_1652805410081_0007
22/05/17 17:02:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-20-252.ec2.internal:20888/proxy/application_1652805410081_0007/
22/05/17 17:02:55 INFO mapreduce.Job: Running job: job_1652805410081_0007
22/05/17 17:03:04 INFO mapreduce.Job: Job job_1652805410081_0007 running in uber mode : false
22/05/17 17:03:04 INFO mapreduce.Job:  map 0% reduce 0%
```



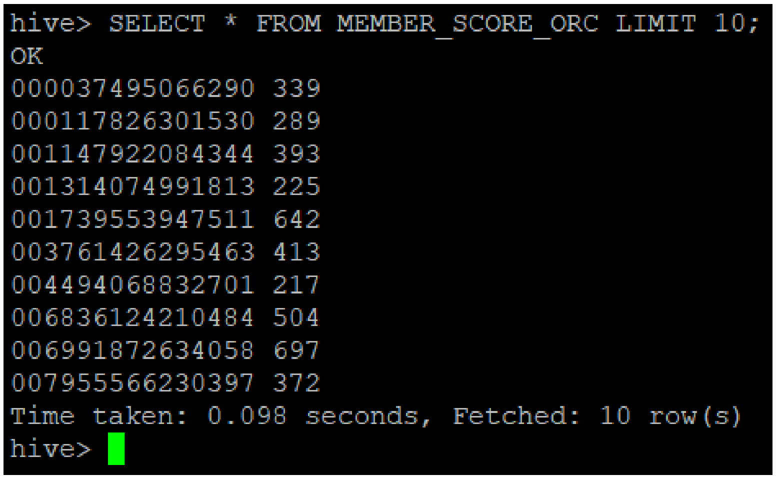
```
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/17 17:02:50 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/17 17:02:50 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-20-252.ec2.internal/172.31.20.252:8032
22/05/17 17:02:54 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: number of splits:1
22/05/17 17:02:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652805410081_0007
22/05/17 17:02:55 INFO impl.YarnClientImpl: Submitted application application_1652805410081_0007
22/05/17 17:02:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-20-252.ec2.internal:20888/proxy/application_1652805410081_0007/
22/05/17 17:02:55 INFO mapreduce.Job: Running job: job_1652805410081_0007
22/05/17 17:03:04 INFO mapreduce.Job: Job job_1652805410081_0007 running in uber mode : false
22/05/17 17:03:04 INFO mapreduce.Job:  map 0% reduce 0%
22/05/17 17:03:11 INFO mapreduce.Job:  map 100% reduce 0%
22/05/17 17:03:11 INFO mapreduce.Job: Job job_1652805410081_0007 completed successfully
22/05/17 17:03:11 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189956
                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)=210768
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4391
                Total vcore-milliseconds taken by all map tasks=4391
                Total megabyte-milliseconds taken by all map tasks=6744576
        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)=89
                CPU time spent (ms)=2550
                Physical memory (bytes) snapshot=289206272
                Virtual memory (bytes) snapshot=3295948800
                Total committed heap usage (bytes)=249036800
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85081
22/05/17 17:03:11 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 20.9187 seconds (3.9719 KB/sec)
22/05/17 17:03:11 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-20-252 ~]$
```

- Verify data in “**card_member_orc**" table.



```
hive> SELECT * FROM CARD_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: 0.111 seconds, Fetched: 10 row(s)
hive>
```

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



```
hive> SELECT * FROM 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: 0.098 seconds, Fetched: 10 row(s)
hive>
```