Data Ingestion from the RDS to HDFS using Sqoop

Sqoop command used for importing table from RDS to HDFS

1. Run below Sqoop command to import member_score table from RDS into HDFS, from command prompt.

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 '/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.
23/05/21 07:21:43 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]
23/05/21 07:21:44 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/05/21 07:21:44 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/05/21 07:21:44 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatical
ly registered via the SPI and manual loading of the driver class is generally unnecessary.
23/05/21 07:21:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/05/21 07:21:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/05/21 07:21:45 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/d8ce8213316f36df6c5fa5f56a4c7384/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/05/21 07:21:49 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/d8ce8213316f36df6c5fa5f56a4c7384/member_score
.jar
23/05/21 07:21:51 INFO tool.ImportTool: Destination directory /capstone_project/member_score is not present, hence not deleting.
23/05/21 07:21:51 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/05/21 07:21:51 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/05/21 07:21:51 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/05/21 07:21:51 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/05/21 07:21:51 INFO mapreduce.ImportJobBase: Beginning import of member_score
23/05/21 07:21:51 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/05/21 07:21:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/05/21 07:21:52 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-92-66.ec2.internal/172.31.92.66:8032
23/05/21 07:21:55 INFO db.DBInputFormat: Using read commited transaction isolation
23/05/21 07:21:55 INFO mapreduce.JobSubmitter: number of splits:1
23/05/21 07:21:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1684652299850_0002
23/05/21 07:21:56 INFO impl.YarnClientImpl: Submitted application application_1684652299850_0002
23/05/21 07:21:56 INFO mapreduce.Job: The url to track the job: http://ip-172-31-92-66.ec2.internal:20888/proxy/application_168465229985
0_0002/
23/05/21 07:21:56 INFO mapreduce.Job: Running job: job_1684652299850_0002
23/05/21 07:22:04 INFO mapreduce.Job: Job job_1684652299850_0002 running in uber mode : false
23/05/21 07:22:04 INFO mapreduce.Job:  map 0% reduce 0%
23/05/21 07:22:11 INFO mapreduce.Job:  map 100% reduce 0%
23/05/21 07:22:11 INFO mapreduce.Job: Job job_1684652299850_0002 completed successfully
23/05/21 07:22:11 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189631
                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)=213456
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4447
                Total vcore-milliseconds taken by all map tasks=4447
                Total megabyte-milliseconds taken by all map tasks=6830592
        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)=75
                CPU time spent (ms)=2170
                Physical memory (bytes) snapshot=266850304
                Virtual memory (bytes) snapshot=3281862656
                Total committed heap usage (bytes)=242745344
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
23/05/21 07:22:11 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 19.6785 seconds (1,015.3227 bytes/sec)
23/05/21 07:22:11 INFO mapreduce.ImportJobBase: Retrieved 999 records.
```

2. Run below Sqoop command to import card_member table from RDS into HDFS, from command prompt.

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 '/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.
23/05/21 07:28:19 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]
23/05/21 07:28:19 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/05/21 07:28:19 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/05/21 07:28:19 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatical
ly registered via the SPI and manual loading of the driver class is generally unnecessary.
23/05/21 07:28:19 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/05/21 07:28:20 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
23/05/21 07:28:20 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/93e78845b87762a455efc0b8dd316de7/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/05/21 07:28:22 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/93e78845b87762a455efc0b8dd316de7/member_score
.jar
23/05/21 07:28:24 INFO tool.ImportTool: Destination directory /capstone_project/card_member is not present, hence not deleting.
23/05/21 07:28:24 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/05/21 07:28:24 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/05/21 07:28:24 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/05/21 07:28:24 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/05/21 07:28:24 INFO mapreduce.ImportJobBase: Beginning import of member_score
23/05/21 07:28:24 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/05/21 07:28:24 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/05/21 07:28:24 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-92-66.ec2.internal/172.31.92.66:8032
23/05/21 07:28:26 INFO db.DBInputFormat: Using read commited transaction isolation
23/05/21 07:28:26 INFO mapreduce.JobSubmitter: number of splits:1
23/05/21 07:28:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1684652299850_0003
23/05/21 07:28:27 INFO impl.YarnClientImpl: Submitted application application_1684652299850_0003
23/05/21 07:28:27 INFO mapreduce.Job: The url to track the job: http://ip-172-31-92-66.ec2.internal:20888/proxy/application_168465229985
0_0003/
23/05/21 07:28:27 INFO mapreduce.Job: Running job: job_1684652299850_0003
23/05/21 07:28:35 INFO mapreduce.Job: Job job_1684652299850_0003 running in uber mode : false
23/05/21 07:28:35 INFO mapreduce.Job:  map 0% reduce 0%
23/05/21 07:28:41 INFO mapreduce.Job:  map 100% reduce 0%
23/05/21 07:28:41 INFO mapreduce.Job: Job job_1684652299850_0003 completed successfully
23/05/21 07:28:41 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189630
                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)=174816
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3642
                Total vcore-milliseconds taken by all map tasks=3642
                Total megabyte-milliseconds taken by all map tasks=5594112
        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)=74
                CPU time spent (ms)=2160
                Physical memory (bytes) snapshot=269864960
                Virtual memory (bytes) snapshot=3286827008
                Total committed heap usage (bytes)=244842496
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19980
23/05/21 07:28:41 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 17.3427 seconds (1.1251 KB/sec)
23/05/21 07:28:41 INFO mapreduce.ImportJobBase: Retrieved 999 records.
```

1. Start hive from command prompt. Create external table card_member_ext which will point to HDFS location to hold data from card_member table in RDS. Sqoop command will write in this location.

```
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 '/capstone_project/card_member';
```

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
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 '/capstone_project/card_member';
OK
Time taken: 0.723 seconds
```

2. Create external table member_score_ext which will point to HDFS location to hold data from member_score table in RDS. Sqoop command will write in this location.

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

```
hive>
    > CREATE EXTERNAL TABLE IF NOT EXISTS MEMBER_SCORE_EXT(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/capstone_project/member_score';
OK
Time taken: 0.058 seconds
```

3. Create card_member_orc table. *Please note ORC format will help in 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");
```

4. Create member_score_orc table. *Please note ORC format will help in better performance.*

```
CREATE TABLE IF NOT EXISTS MEMBER_SCORE_ORC(
```

`MEMBER_ID` STRING,
`SCORE` INT)
STORED AS ORC
TBLPROPERTIES ("orc.compress"="SNAPPY");

```
hive>
    >
    > 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");
OK
Time taken: 0.426 seconds
hive>
    > CREATE TABLE IF NOT EXISTS MEMBER_SCORE_ORC(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.252 seconds
```

5. Load data into card_member_orc from card_member_ext.

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

```
hive>
    > INSERT OVERWRITE TABLE CARD_MEMBER_ORC
    > SELECT CARD_ID, MEMBER_ID, MEMBER_JOINING_DT, CARD_PURCHASE_DT, COUNTRY, CITY FROM CARD_MEMBER_EXT;
Query ID = hadoop_20230521074040_ee2e72f5-76e9-4f39-9de8-572bf6f5fcf8
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_1684652299850_0005)

--------------------------------------------------------------------------------
        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: 5.12 s
--------------------------------------------------------------------------------
Loading data to table default.card_member_orc
OK
Time taken: 16.71 seconds
hive>
    >
```

6. Load data into member_score_orc from member_score_ext.

INSERT OVERWRITE TABLE MEMBER_SCORE_ORC
SELECT MEMBER_ID, SCORE FROM MEMBER_SCORE_EXT;

```
hive>
    > INSERT OVERWRITE TABLE MEMBER_SCORE_ORC
    > SELECT MEMBER_ID, SCORE FROM MEMBER_SCORE_EXT;
Query ID = hadoop_20230521074149_742d60b3-7099-43d8-8a4f-afbfdca2c817
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1684652299850_0005)

--------------------------------------------------------------------------------
        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: 5.63 s
--------------------------------------------------------------------------------
Loading data to table default.member_score_orc
OK
Time taken: 6.863 seconds
```

Screenshot of the imported data

7. Verify some data in card_member_orc table.

SELECT * FROM CARD_MEMBER_ORC LIMIT 10;

```
hive> SELECT * FROM CARD_MEMBER_ORC LIMIT 10;
OK
000037495066290 339     NULL    NULL    NULL    NULL
000117826301530 289     NULL    NULL    NULL    NULL
001147922084344 393     NULL    NULL    NULL    NULL
0013140749918l3 225     NULL    NULL    NULL    NULL
001739553947511 642     NULL    NULL    NULL    NULL
003761426295463 413     NULL    NULL    NULL    NULL
004494068832701 217     NULL    NULL    NULL    NULL
006836124210484 504     NULL    NULL    NULL    NULL
006991872634058 697     NULL    NULL    NULL    NULL
007955566230397 372     NULL    NULL    NULL    NULL
Time taken: 0.208 seconds, Fetched: 10 row(s)
hive>
```

8. Verify some data in member_score_orc table.

SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;

```
hive>
    > SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;
OK
000037495066290 339
000117826301530 289
001147922084344 393
0013140749918l3 225
001739553947511 642
003761426295463 413
004494068832701 217
006836124210484 504
006991872634058 697
007955566230397 372
Time taken: 0.125 seconds, Fetched: 10 row(s)
```