

Data Ingestion from the RDS to HDFS using Sqoop

- **Sqoop command used for importing table from RDS to HDFS**

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

```
sqoop import --connect jdbc:mysql://upgradawsrds1.cyaie1c9bmnf.us-east1.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
```

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

```
sqoop import --connect jdbc:mysql://upgradawsrds1.cyaie1c9bmnf.us-east1.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
```

----- Sqoop Operations: Ends Here-----

- **Command to see the list of imported data in HDFS**

1. Start hive and 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';
```

2. 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';
```

3. 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");
```

4. 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");
```

5. 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;
```

6. 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;
```

7. Verify data in "**card_member_orc**" table.

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

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

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

- Screenshot of the imported data

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

```
[ec2-user@ip-172-31-11-78 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaie1c9bmnf.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.
23/01/07 12:07:30 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]
23/01/07 12:07:30 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/01/07 12:07:30 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/01/07 12:07:30 INFO tool.CodeGenTool: Beginning code generation
23/01/07 12:07:31 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'member_score' AS t LIMIT 1
23/01/07 12:07:31 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'member_score' AS t LIMIT 1
23/01/07 12:07:31 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/ab2ea66eb3392e4fec28041cc70727f3/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/01/07 12:07:33 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/ab2ea66eb3392e4fec28041cc70727f3/member_score.jar
23/01/07 12:07:34 INFO tool.ImportTool: Destination directory /ccfd_capstone_project/member_score is not present, hence not deleting.
23/01/07 12:07:34 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/01/07 12:07:34 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/01/07 12:07:34 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/01/07 12:07:34 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/01/07 12:07:34 INFO mapreduce.ImportJobBase: Beginning import of member_score
23/01/07 12:07:34 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/01/07 12:07:34 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/01/07 12:07:34 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-11-78.ec2.internal/172.31.11.78:8032
23/01/07 12:07:37 INFO db.DBInputFormat: Using read committed transaction isolation
23/01/07 12:07:37 INFO mapreduce.JobSubmitter: number of splits:1
23/01/07 12:07:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1673092517399_0002
23/01/07 12:07:37 INFO impl.YarnClientImpl: Submitted application application_1673092517399_0002
23/01/07 12:07:37 INFO mapreduce.Job: The url to track the job: http://ip-172-31-11-78.ec2.internal:20888/proxy/application_1673092517399_0002/
23/01/07 12:07:37 INFO mapreduce.Job: Running job: job_1673092517399_0002
23/01/07 12:07:44 INFO mapreduce.Job: Job job_1673092517399_0002 running in uber mode : false
23/01/07 12:07:44 INFO mapreduce.Job: map 0% reduce 0%
23/01/07 12:07:49 INFO mapreduce.Job: map 100% reduce 0%
23/01/07 12:07:50 INFO mapreduce.Job: Job job_1673092517399_0002 completed successfully
23/01/07 12:07:50 INFO mapreduce.Job: Counters: 30
```

```
23/01/07 12:07:50 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=189845
    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)=265152
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=2762
    Total vcore-milliseconds taken by all map tasks=2762
    Total megabyte-milliseconds taken by all map tasks=8484864
  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)=68
    CPU time spent (ms)=1830
    Physical memory (bytes) snapshot=322482176
    Virtual memory (bytes) snapshot=4621459456
    Total committed heap usage (bytes)=321912832
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=19980
23/01/07 12:07:50 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 15.6914 seconds (1.2435 KB/sec)
23/01/07 12:07:50 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-11-78 ~]$
```

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

```
[ec2-user@ip-172-31-11-78 ~]$ sqoop import --connect jdbc:mysql://upgradawsrds1.cyaiclc9bmnf.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.
23/01/07 12:10:15 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]
23/01/07 12:10:15 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/01/07 12:10:15 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/01/07 12:10:15 INFO tool.CodeGenTool: Beginning code generation
23/01/07 12:10:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'card_member' AS t LIMIT 1
23/01/07 12:10:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'card_member' AS t LIMIT 1
23/01/07 12:10:16 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/7350904e840073dd9e4e0303b0f9ab2c/card_member.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/01/07 12:10:18 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/7350904e840073dd9e4e0303b0f9ab2c/card_member.jar
23/01/07 12:10:19 INFO tool.ImportTool: Destination directory /ccfd_capstone_project/card_member is not present, hence not deleting.
23/01/07 12:10:19 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/01/07 12:10:19 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/01/07 12:10:19 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/01/07 12:10:19 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/01/07 12:10:19 INFO mapreduce.ImportJobBase: Beginning import of card_member
23/01/07 12:10:19 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/01/07 12:10:19 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/01/07 12:10:19 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-11-78.ec2.internal/172.31.11.78:8032
23/01/07 12:10:23 INFO db.DBInputFormat: Using read committed transaction isolation
23/01/07 12:10:23 INFO mapreduce.JobSubmitter: number of splits:1
23/01/07 12:10:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1673092517399_0003
23/01/07 12:10:24 INFO impl.YarnClientImpl: Submitted application application_1673092517399_0003
23/01/07 12:10:24 INFO mapreduce.Job: The url to track the job: http://ip-172-31-11-78.ec2.internal:20888/proxy/application_1673092517399_0003/
23/01/07 12:10:24 INFO mapreduce.Job: Running job: job_1673092517399_0003
23/01/07 12:10:30 INFO mapreduce.Job: Job job_1673092517399_0003 running in uber mode : false
23/01/07 12:10:30 INFO mapreduce.Job: map 0% reduce 0%
23/01/07 12:10:36 INFO mapreduce.Job: map 100% reduce 0%
23/01/07 12:10:37 INFO mapreduce.Job: Job job_1673092517399_0003 completed successfully
23/01/07 12:10:37 INFO mapreduce.Job: Counters: 30
```

```
23/01/07 12:10:37 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=189901
    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)=368448
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=3838
    Total vcore-milliseconds taken by all map tasks=3838
    Total megabyte-milliseconds taken by all map tasks=11790336
  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)=67
    CPU time spent (ms)=2400
    Physical memory (bytes) snapshot=281751552
    Virtual memory (bytes) snapshot=4638691328
    Total committed heap usage (bytes)=245366784
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=85081
23/01/07 12:10:37 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 18.2374 seconds (4.5559 KB/sec)
23/01/07 12:10:37 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[ec2-user@ip-172-31-11-78 ~]$
```

----- Sqoop Operations: Ends Here-----

1. 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: 0.148 seconds, Fetched: 10 row(s)
hive>
```

2. 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: 0.096 seconds, Fetched: 10 row(s)
hive> █
```