Data Ingestion from the RDS to HDFS using Sqoop

Sqoop command used for importing table from RDS to HDFS

————————————————————--SQOOP
———————————————————— - **Setting up the environment for sqoop**

sudo -i
wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector- java-8.0.25.tar.gz
tar -xvf mysql-connector-java-8.0.25.tar.gz
cd mysql-connector-java-8.0.25/
sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
cd

Command to see the list of imported data in HDFS

- **Importing data from RDS to hdfs using swoop for card member** sqoop import \

--connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \ --table card_member \
--username upgraduser --password upgraduser \ --null-string '\\N' --null-non-string '\\N' \

--target-dir /user/root/card_member \ -m 1 --as-textfile

```
[root@ip-172-31-21-211 ~]# sqoop import \
> --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --table card_member \
> --username upgraduser --password upgraduser \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/card_member \
> -m 1 --as-textfile
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/05/13 19:16:35 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]
22/05/13 19:16:35 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/13 19:16:35 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/13 19:16:35 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 automatically registered via the SPI and manual loading of the driver class i
s generally unnecessary.
22/05/13 19:16:36 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/13 19:16:36 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/13 19:16:36 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/651ea94beaa0a7b06bf04ccfc0f6935d/card_member.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/13 19:16:39 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/651ea94beaa0a7b06bf04ccfc0f6935d/card_member.jar
22/05/13 19:16:39 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/13 19:16:39 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/13 19:16:39 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/13 19:16:39 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/13 19:16:39 INFO mapreduce.ImportJobBase: Beginning import of card_member
22/05/13 19:16:39 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/13 19:16:40 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/13 19:16:40 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-21-211.ec2.internal/172.31.21.211:8032
22/05/13 19:16:44 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/13 19:16:44 INFO mapreduce.JobSubmitter: number of splits:1
22/05/13 19:16:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652465715257_0002
22/05/13 19:16:45 INFO impl.YarnClientImpl: Submitted application application_1652465715257_0002
22/05/13 19:16:45 INFO mapreduce.Job: The url to track the job: http://ip-172-31-21-211.ec2.internal:20888/proxy/application_1652465715257_0002/
22/05/13 19:16:45 INFO mapreduce.Job: Running job: job_1652465715257_0002
22/05/13 19:16:55 INFO mapreduce.Job: Job job_1652465715257_0002 running in uber mode : false
22/05/13 19:16:55 INFO mapreduce.Job:  map 0% reduce 0%
22/05/13 19:17:03 INFO mapreduce.Job:  map 100% reduce 0%
22/05/13 19:17:04 INFO mapreduce.Job: Job job_1652465715257_0002 completed successfully
22/05/13 19:17:04 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189492
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

- **Importing data from RDS to hdfs using sqoop for member score** sqoop import \
--connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \

--table member_score \
--username upgraduser --password upgraduser \ --null-string '\\N' --null-non-string '\\N' \ --target-dir /user/root/member_score \
-m 1 --as-textfile

```
[root@ip-172-31-21-211 ~]# sqoop import \
> --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --table member_score \
> --username upgraduser --password upgraduser \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/member_score \
> -m 1 --as-textfile
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/05/13 19:17:38 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]
22/05/13 19:17:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/13 19:17:38 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/13 19:17:38 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 automatically registered via the SPI and manual loading of the driver class i
s generally unnecessary.
22/05/13 19:17:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/13 19:17:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/13 19:17:39 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/0617866c7b6995aa5f5b1040c02378c5/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/13 19:17:42 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/0617866c7b6995aa5f5b1040c02378c5/member_score.jar
22/05/13 19:17:42 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/13 19:17:42 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/13 19:17:42 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/13 19:17:42 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/13 19:17:42 INFO mapreduce.ImportJobBase: Beginning import of member_score
22/05/13 19:17:43 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/13 19:17:44 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/13 19:17:44 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-21-211.ec2.internal/172.31.21.211:8032
22/05/13 19:17:48 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/13 19:17:48 INFO mapreduce.JobSubmitter: number of splits:1
22/05/13 19:17:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652465715257_0003
22/05/13 19:17:49 INFO impl.YarnClientImpl: Submitted application application_1652465715257_0003
22/05/13 19:17:49 INFO mapreduce.Job: The url to track the job: http://ip-172-31-21-211.ec2.internal:20888/proxy/application_1652465715257_0003/
22/05/13 19:17:49 INFO mapreduce.Job: Running job: job_1652465715257_0003
22/05/13 19:17:58 INFO mapreduce.Job: Job job_1652465715257_0003 running in uber mode : false
22/05/13 19:17:58 INFO mapreduce.Job:  map 0% reduce 0%
22/05/13 19:18:04 INFO mapreduce.Job:  map 100% reduce 0%
22/05/13 19:18:04 INFO mapreduce.Job: Job job_1652465715257_0003 completed successfully
22/05/13 19:18:04 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189440
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

- **Checking whether importing is successful or not fro card_member?** hadoop fs -ls /user/root/card_member

- **Checking whether importing is successful or not for member_score?**

hadoop fs -ls /user/root/member_score

```
[root@ip-172-31-21-211 ~]# hadoop fs -ls /user/root/card_member
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-05-13 19:17 /user/root/card_member/_SUCCESS
-rw-r--r--   1 root hadoop      85081 2022-05-13 19:17 /user/root/card_member/part-m-00000
[root@ip-172-31-21-211 ~]#
```

```
[root@ip-172-31-21-211 ~]# hadoop fs -ls /user/root/member_score
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-05-13 19:18 /user/root/member_score/_SUCCESS
-rw-r--r--   1 root hadoop      19980 2022-05-13 19:18 /user/root/member_score/part-m-00000
[root@ip-172-31-21-211 ~]#
```

- **Count the number of rows for card member**
hadoop fs -cat /user/root/card_member/part-m-00000 | wc -l

- **Count the number of rows for member score**
hadoop fs -cat /user/root/member_score/part-m-00000 | wc -l

——————————————————HIVE ——————————————————

use capstone_project;

- **Creating table in hive for card member** 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/root/ card_member';

```
[root@ip-172-31-21-211 ~]# hadoop fs -cat /user/root/card_member/part-m-00000 | wc -l
999
[root@ip-172-31-21-211 ~]#
```

```
[root@ip-172-31-21-211 ~]# hadoop fs -cat /user/root/member_score/part-m-00000 | wc -l
999
[root@ip-172-31-21-211 ~]#
```

```
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/root/card_member';
OK
Time taken: 0.471 seconds
```

- **Creating table in hive for member score** CREATE EXTERNAL TABLE IF NOT EXISTS MEMBER_SCORE_EXT(`MEMBER_ID` STRING, `SCORE` INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION '/user/root/ member_score';

- **Creating table in hive for card member orc** 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");

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS MEMBER_SCORE_EXT( `MEMBER_ID` STRING,
    > `SCORE` INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/root/member_score';
OK
Time taken: 0.053 seconds
```

```
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: 0.331 seconds
```

- **Creating table in hive for member score orc** 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 MEMBER_SCORE_ORC(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.067 seconds
```

- **Inserting into card member orc**

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 = root_20220513192104_2dcf48de-a67a-43ac-8d06-22e4a6689275
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0004)

--------------------------------------------------------------------------------
        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: 6.20 s
--------------------------------------------------------------------------------
Loading data to table capstone_project.card_member_orc
OK
Time taken: 12.335 seconds
```

- **Inserting into member score orc**

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 = root_20220513192121_24fffe79-492b-4c91-ad1f-194d82b33cb5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0004)

--------------------------------------------------------------------------------
        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: 0.56 s
--------------------------------------------------------------------------------
Loading data to table capstone_project.member_score_orc
OK
Time taken: 1.945 seconds
```

- **Printing top 10 rows from card member orc** SELECT * FROM CARD_MEMBER_ORC LIMIT 10;

```
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: 0.16 seconds, Fetched: 10 row(s)
```

- **Printing top 10 rows from member score orc** SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;

```
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: 0.133 seconds, Fetched: 10 row(s)
```

- **Counting rows from card member orc**
SELECT count(*) FROM CARD_MEMBER_ORC LIMIT 10;

```
hive> SELECT count(*) FROM CARD_MEMBER_ORC LIMIT 10;
Query ID = root_20220513192319_93f614b8-eb0b-480d-9694-fdb010d3f29b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.76 s
--------------------------------------------------------------------------------
OK
999
Time taken: 1.366 seconds, Fetched: 1 row(s)
```

- **Counting rows from member score orc**
SELECT count(*) FROM MEMBER_SCORE_ORC LIMIT 10;