1. Install MYSQL connector

Login into the EMR instance and run following commands

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/

2. Access Hbase shell and create a table

Switch user to root using following command
sudo -i

Then access hbase shell using following command
hbase shell

Create a table ‘TaxiTrips’ with a column family ‘TripDetails’ using following statement

create table ‘TaxiTrips’, ‘TripDetails’

```
[[root@ip-172-31-16-22 ~]# sudo -i

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[[root@ip-172-31-16-22 ~]# hbase shell
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Sat Sep 19 02:15:00 UTC 2020

[hbase(main):001:0> create 'TaxiTrips', 'TripDetails'
0 row(s) in 1.4610 seconds

=> Hbase::Table - TaxiTrips
[hbase(main):002:0> list
TABLE
TaxiTrips
1 row(s) in 0.0160 seconds

=> ["TaxiTrips"]
hbase(main):003:0>
```

3. Import data from RDS to HBase table

```
sqoop import \
  --connect
jdbc:mysql://database-1.c10i7rbsadgr.us-east-1.rds.amazonaws.com:3306/YellowTaxiTr
ipsDB \
  --username admin \
  --password admin_mysql \
  --table tripdata \
  --hbase-table TaxiTrips \
  --column-family TripDetails \
  --hbase-create-table \
  --hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime \
  --split-by payment_type \
--hbase-bulkload
```

This command is a Sqoop command, which is a tool designed for efficiently transferring data between Apache Hadoop and relational databases. In this case, the command is importing data from a MySQL database into an HBase table.

Below is a breakdown of the command and its various options:

1. `sqoop import`: Initiates the Sqoop import process.

2. `--connect jdbc:mysql://database-1.c10i7rbsadgr.us-east-1.rds.amazonaws.com:3306/YellowTaxiTripsDB`: Specifies the JDBC connection string to the MySQL database. It includes the host address (`database-1.c10i7rbsadgr.us-east-1.rds.amazonaws.com`), port (`3306`), and the database name (`YellowTaxiTripsDB`).

3. `--username admin`: Specifies the MySQL username for authentication.

4. `--password admin_mysql`: Specifies the MySQL password for authentication.

5. `--table tripdata`: Specifies the name of the MySQL table (`tripdata`) from which data will be imported.

6. `--hbase-table TaxiTrips`: Specifies the name of the HBase table (`TaxiTrips`) where the data will be imported.

7. `--column-family TripDetails`: Specifies the HBase column family (`TripDetails`) where the data will be stored within the HBase table.

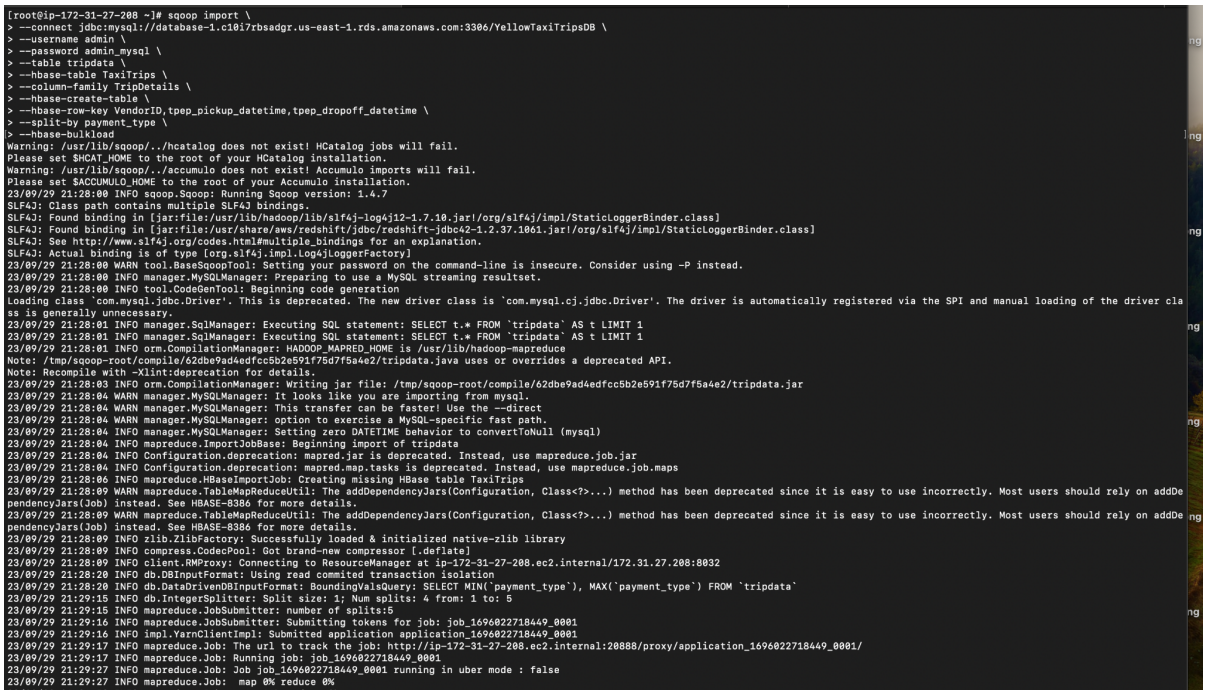
8. `--hbase-create-table`: Directs Sqoop to create the HBase table if it doesn't already exist.

9. `--hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime`: Specifies the columns from the source table that will be used to construct the row key in HBase.

10. `--split-by payment_type`: Specifies the column (`payment_type`) by which the data will be split into multiple HBase regions. This can help optimize data distribution.

11. `--hbase-bulkload`: Indicates that the import should use HBase bulk load. This can be more efficient for large datasets.

In summary, this Sqoop command is importing data from a MySQL table (`tripdata`) into an HBase table (`TaxiTrips`). It defines how the data should be structured in HBase, including the column family, row key, and the use of bulk loading for optimization.



```
23/09/29 21:29:17 INFO mapreduce.Job: Running job: job_1696022718449_0001
23/09/29 21:29:27 INFO mapreduce.Job: Job job_1696022718449_0001 running in uber mode : false
23/09/29 21:29:27 INFO mapreduce.Job:  map 0% reduce 0%
23/09/29 21:34:53 INFO mapreduce.Job:  map 14% reduce 0%
23/09/29 21:34:59 INFO mapreduce.Job:  map 18% reduce 0%
23/09/29 21:35:05 INFO mapreduce.Job:  map 20% reduce 0%
23/09/29 21:35:29 INFO mapreduce.Job:  map 13% reduce 0%
23/09/29 21:36:28 INFO mapreduce.Job:  map 20% reduce 0%
23/09/29 21:37:34 INFO mapreduce.Job:  map 40% reduce 0%
23/09/29 21:38:39 INFO mapreduce.Job:  map 60% reduce 0%
23/09/29 21:39:44 INFO mapreduce.Job:  map 80% reduce 0%
23/09/29 21:40:06 INFO mapreduce.Job:  map 94% reduce 0%
23/09/29 21:40:12 INFO mapreduce.Job:  map 97% reduce 0%
23/09/29 21:40:18 INFO mapreduce.Job:  map 99% reduce 0%
23/09/29 21:40:24 INFO mapreduce.Job:  map 100% reduce 0%
23/09/29 21:40:42 INFO mapreduce.Job:  map 93% reduce 0%
23/09/29 21:43:41 INFO mapreduce.Job:  map 100% reduce 0%
23/09/29 21:44:03 INFO mapreduce.Job:  map 100% reduce 27%
23/09/29 21:44:39 INFO mapreduce.Job:  map 100% reduce 67%
23/09/29 21:45:22 INFO mapreduce.Job:  map 100% reduce 68%
23/09/29 21:45:58 INFO mapreduce.Job:  map 100% reduce 69%
23/09/29 21:46:34 INFO mapreduce.Job:  map 100% reduce 70%
23/09/29 21:47:11 INFO mapreduce.Job:  map 100% reduce 71%
23/09/29 21:47:47 INFO mapreduce.Job:  map 100% reduce 72%
23/09/29 21:48:23 INFO mapreduce.Job:  map 100% reduce 73%
23/09/29 21:48:59 INFO mapreduce.Job:  map 100% reduce 74%
23/09/29 21:49:41 INFO mapreduce.Job:  map 100% reduce 75%
23/09/29 21:50:17 INFO mapreduce.Job:  map 100% reduce 76%
23/09/29 21:50:59 INFO mapreduce.Job:  map 100% reduce 77%
23/09/29 21:51:35 INFO mapreduce.Job:  map 100% reduce 78%
23/09/29 21:52:11 INFO mapreduce.Job:  map 100% reduce 79%
23/09/29 21:52:53 INFO mapreduce.Job:  map 100% reduce 80%
23/09/29 21:53:29 INFO mapreduce.Job:  map 100% reduce 81%
23/09/29 21:54:11 INFO mapreduce.Job:  map 100% reduce 82%
23/09/29 21:54:47 INFO mapreduce.Job:  map 100% reduce 83%
23/09/29 21:55:23 INFO mapreduce.Job:  map 100% reduce 84%
23/09/29 21:56:05 INFO mapreduce.Job:  map 100% reduce 85%
23/09/29 21:56:42 INFO mapreduce.Job:  map 100% reduce 86%
23/09/29 21:57:18 INFO mapreduce.Job:  map 100% reduce 87%
23/09/29 21:57:54 INFO mapreduce.Job:  map 100% reduce 88%
23/09/29 21:58:30 INFO mapreduce.Job:  map 100% reduce 89%
23/09/29 21:59:12 INFO mapreduce.Job:  map 100% reduce 90%
23/09/29 21:59:48 INFO mapreduce.Job:  map 100% reduce 91%
23/09/29 22:00:24 INFO mapreduce.Job:  map 100% reduce 92%
23/09/29 22:01:00 INFO mapreduce.Job:  map 100% reduce 93%
23/09/29 22:01:42 INFO mapreduce.Job:  map 100% reduce 94%
23/09/29 22:02:24 INFO mapreduce.Job:  map 100% reduce 95%
23/09/29 22:03:00 INFO mapreduce.Job:  map 100% reduce 96%
23/09/29 22:03:42 INFO mapreduce.Job:  map 100% reduce 97%
23/09/29 22:04:18 INFO mapreduce.Job:  map 100% reduce 98%
23/09/29 22:04:55 INFO mapreduce.Job:  map 100% reduce 99%
23/09/29 22:05:32 INFO mapreduce.Job:  map 100% reduce 100%
23/09/29 22:05:50 INFO mapreduce.Job: Job job_1696022718449_0001 completed successfully
23/09/29 22:05:50 INFO mapreduce.Job: Counters: 50
```

```
23/09/29 22:05:50 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=13788173381
                FILE: Number of bytes written=18919170535
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=591
                HDFS: Number of bytes written=27290827018
                HDFS: Number of read operations=19
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Launched reduce tasks=1
                Other local map tasks=5
                Total time spent by all maps in occupied slots (ms)=70213296
                Total time spent by all reduces in occupied slots (ms)=127223040
                Total time spent by all map tasks (ms)=1462777
                Total time spent by all reduce tasks (ms)=1325240
                Total vcore-milliseconds taken by all map tasks=1462777
                Total vcore-milliseconds taken by all reduce tasks=1325240
                Total megabyte-milliseconds taken by all map tasks=2246825472
                Total megabyte-milliseconds taken by all reduce tasks=4071137280
        Map-Reduce Framework
                Map input records=18880595
                Map output records=302089520
                Map output bytes=45070720381
                Map output materialized bytes=5129630195
                Input split bytes=591
                Combine input records=0
                Combine output records=0
                Reduce input groups=18856687
                Reduce shuffle bytes=5129630195
                Reduce input records=302089520
                Reduce output records=301706992
                Spilled Records=1113612712
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=14601
                CPU time spent (ms)=2822130
                Physical memory (bytes) snapshot=4907794432
                Virtual memory (bytes) snapshot=21257220096
                Total committed heap usage (bytes)=4120379392
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=27290827018
23/09/29 22:05:50 INFO mapreduce.ImportJobBase: Transferred 25.4166 GB in 2,260.7043 seconds (11.5126 MB/sec)
23/09/29 22:05:50 INFO mapreduce.ImportJobBase: Retrieved 302089520 records.
23/09/29 22:05:50 WARN mapreduce.LoadIncrementalHFiles: managed connection cannot be used for bulkload. Creating unmanaged connection.
23/09/29 22:05:50 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2-hbase.properties
```