

- First, log in into your EMR instance (using hadoop, then switch user to root using `sudo -i`) and complete the initial steps of setup. Now you need to run the following command to install the MySQL connector jar file.

```
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/
```

- To ingest data from mySQL RDS to HBase table:

```
sqoop import --connect jdbc:mysql://mapreduceassignment.cnx2n9cjsi3r.us-east-1.rds.amazonaws.com/mapreduceDB --username admin --password password --table task1 --hbase-table task1_hbase --column-family cf1 --hbase-create-table --hbase-row-key tpep_pickup_datetime,tpep_dropoff_datetime --hbase-bulkload --split-by payment_type
```

command explanation:

This is a `sqoop` command that imports data from a MySQL database table `task1` into an HBase table `task1_hbase`. Here is what each option in the command does:

- `--connect`: specifies the JDBC connection string for the MySQL database.
- `--username`: specifies the username to use when connecting to the MySQL database.
- `--password`: specifies the password to use when connecting to the MySQL database.
- `--table`: specifies the name of the MySQL table to import data from.
- `--hbase-table`: specifies the name of the HBase table to import data into.
- `--column-family`: specifies the name of the column family in HBase where the

imported data will be stored.

- `--hbase-create-table``: creates an HBase table if it does not exist.
- `--hbase-row-key``: specifies one or more columns from the MySQL table that will be used as the row key in HBase.
- `--hbase-bulkload``: uses HBase bulk load feature for faster data loading.
- `--split-by``: specifies a column from the MySQL table that will be used to split data into multiple HBase regions.

In this command, data is imported from the MySQL table `task1`` into an HBase table named `task1_hbase``. The column family in HBase where the imported data will be stored is named `cf1``. The row key in HBase is composed of two columns from the MySQL table: `tpep_pickup_datetime`` and `tpep_dropoff_datetime``. The data is loaded into HBase using bulk load feature for faster loading. The data is split into multiple regions based on the column `payment_type``. Let me know if you have any other questions.

```
Map output bytes=41181317811
Map output materialized bytes=4892837443
Input split bytes=591
Combine input records=0
Combine output records=0
Reduce input groups=18842048
Reduce shuffle bytes=4892837443
Reduce input records=302089520
Reduce output records=301472768
Spilled Records=1104697432
Shuffled Maps =5
Failed Shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=15178
CPU time spent (ms)=2876050
Physical memory (bytes) snapshot=4982210560
Virtual memory (bytes) snapshot=21423403008
Total committed heap usage (bytes)=4391960576

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=23981007255
23/07/09 12:15:58 INFO mapreduce.ImportJobBase: Transferred 22.3341 GB in 2,314.3603 seconds (9.8818 MB/sec)
23/07/09 12:15:58 INFO mapreduce.ImportJobBase: Retrieved 302089520 records.
23/07/09 12:15:58 WARN mapreduce.LoadIncrementalHFiles: managed connection cannot be used for bulkload. Creat
23/07/09 12:15:58 WARN mapreduce.LoadIncrementalHFiles: Skipping non-directory hdfs://ip-172-31-36-93.ec2.in
23/07/09 12:15:59 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2-hbase.properties
23/07/09 12:15:59 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
23/07/09 12:15:59 INFO impl.MetricsSystemImpl: HBase metrics system started
23/07/09 12:15:59 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-36-93.ec2
1344eddadbda162b054be99 with size: 11191064341 bytes can be problematic as it may lead to oversplitting.
23/07/09 12:15:59 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-36-93.ec2
f624901a53077bdbc3190c90 with size: 11191083018 bytes can be problematic as it may lead to oversplitting.
23/07/09 12:15:59 INFO Configuration.deprecation: hbase.offheapcache.minblocksize is deprecated. Instead, us
[hadoop@ip-172-31-36-93 ~]$
```