- 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.