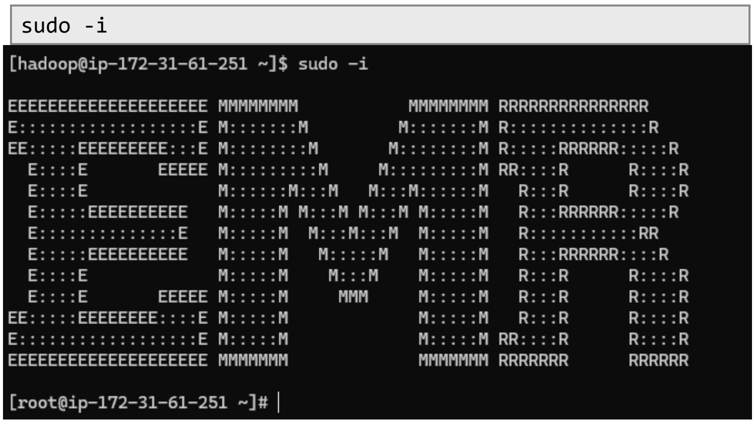
TASK-2:

Use Sqoop command to ingest the data from RDS into the HBase Table

1) First login into the EMR instance using Hadoop and switch to root user using the following command:

```
sudo -i
```



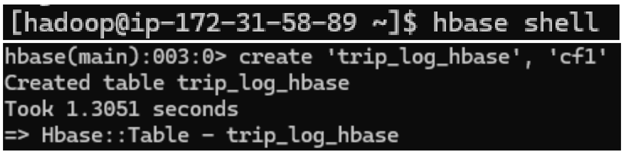
```
[hadoop@ip-172-31-61-251 ~]$ sudo -i

EEEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[root@ip-172-31-61-251 ~]#
```

2) Complete the initial steps of setup by running the following commands for setting up MySQL connector

```
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/
```

3) Start HBase shell to create a table named 'trip_data_hbase' with a column family 'cf1'

```
hbase shell
create 'trip_log_hbase', 'cf1'
exit
```

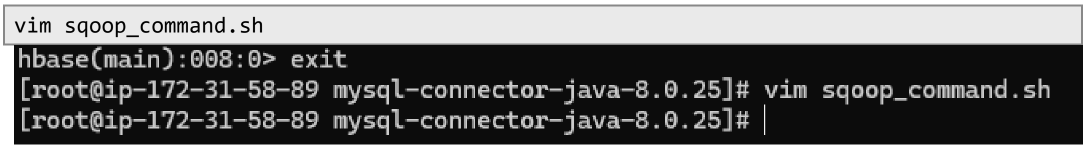


```
[hadoop@ip-172-31-58-89 ~]$ hbase shell
hbase(main):003:0> create 'trip_log_hbase', 'cf1'
Created table trip_log_hbase
Took 1.3051 seconds
=> Hbase::Table - trip_log_hbase
```

4) To ingest data from MySQL RDS to HBase table, we run the following commands:

   (a) Create a 'sqoop_command.sh' file and insert the following code in it:

```
vim sqoop_command.sh
```



```
hbase(main):008:0> exit
[root@ip-172-31-58-89 mysql-connector-java-8.0.25]# vim sqoop_command.sh
[root@ip-172-31-58-89 mysql-connector-java-8.0.25]#
```

```
sqoop import --connect jdbc:mysql://mydbinstance.c9zffzdmradr.us-east-1.rds.amazonaws.com:3306/taxi_records --username admin --password 20021960 --table trip_log --hbase-table trip_log_hbase --column-family cf1 --hbase-create-table --hbase-row-key tpep_pickup_datetime,tpep_dropoff_datetime --hbase-bulkload --split-by payment_type
```

This command facilitates efficient data migration from MySQL to HBase by:

Connecting to the specified MySQL database and table:

```
sqoop import --connect jdbc:mysql://mydbinstance.c9zffzdmradr.us-east-1.rds.amazonaws.com:3306/taxi_records
```

Creating the necessary HBase table and column family:

```
--username admin --password 20021960 --table trip_log --hbase-table trip_log_hbase --column-family cf1 --hbase-create-table
```

Using composite row keys for unique row identification:

```
--hbase-row-key tpep_pickup_datetime,tpep_dropoff_datetime
```

Employing bulk load for optimal performance:

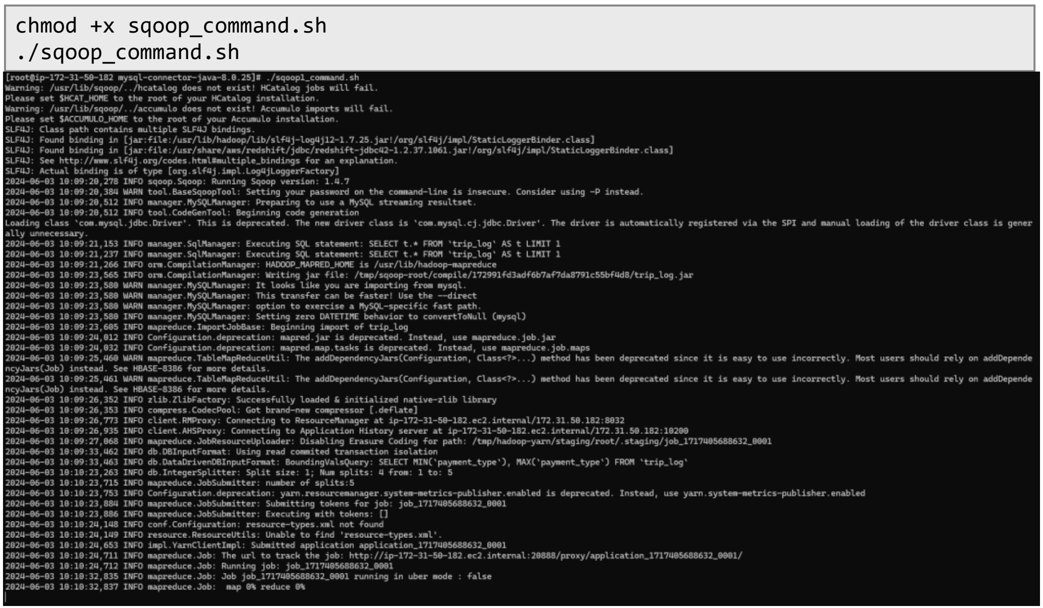
```
--hbase-bulkload
```

Utilizing parallel processing to speed up the import:

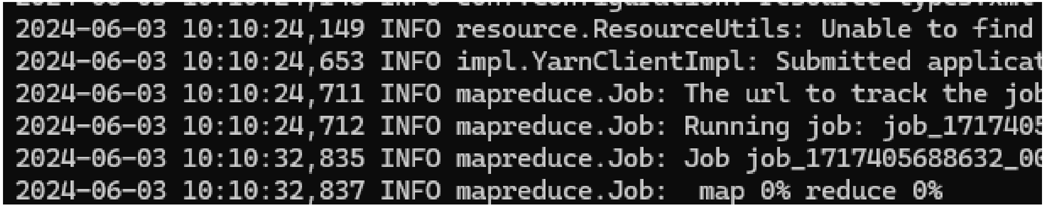
```
--split-by payment_type
```

(b) After saving the ‘sqoop_command.sh’ run the following command:

```
chmod +x sqoop_command.sh
./sqoop_command.sh
```



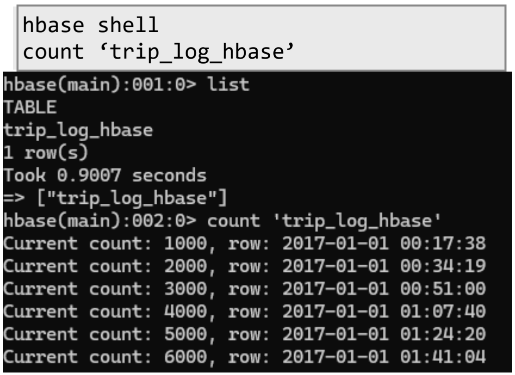
```
[root@ip-172-31-50-182 mysql-connector-java-8.0.25]# ./sqoop1_command.sh
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.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: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-06-03 10:09:20,278 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-03 10:09:20,384 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-03 10:09:20,512 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-03 10:09:20,512 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 is gener
ally unnecessary.
2024-06-03 10:09:21,153 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `trip_log` AS t LIMIT 1
2024-06-03 10:09:21,237 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `trip_log` AS t LIMIT 1
2024-06-03 10:09:21,266 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2024-06-03 10:09:23,565 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/172991fd3adf6b7af7da8791c55bf4d8/trip_log.jar
2024-06-03 10:09:23,580 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-06-03 10:09:23,580 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2024-06-03 10:09:23,580 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2024-06-03 10:09:23,580 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2024-06-03 10:09:23,605 INFO mapreduce.ImportJobBase: Beginning import of trip_log
2024-06-03 10:09:24,012 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2024-06-03 10:09:24,032 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-06-03 10:09:25,460 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use incorrectly. Most users should rely on addDepende
ncyJars(Job) instead. See HBASE-8386 for more details.
2024-06-03 10:09:25,461 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use incorrectly. Most users should rely on addDepende
ncyJars(Job) instead. See HBASE-8386 for more details.
2024-06-03 10:09:26,352 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
2024-06-03 10:09:26,353 INFO compress.CodecPool: Got brand-new compressor [.deflate]
2024-06-03 10:09:26,773 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-50-182.ec2.internal/172.31.50.182:8032
2024-06-03 10:09:26,935 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-50-182.ec2.internal/172.31.50.182:10200
2024-06-03 10:09:27,068 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1717405688632_0001
2024-06-03 10:09:33,462 INFO db.DBInputFormat: Using read commited transaction isolation
2024-06-03 10:09:33,463 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`payment_type`), MAX(`payment_type`) FROM `trip_log`
2024-06-03 10:10:23,263 INFO db.IntegerSplitter: Split size: 1; Num splits: 4 from: 1 to: 5
2024-06-03 10:10:23,715 INFO mapreduce.JobSubmitter: number of splits:5
2024-06-03 10:10:23,753 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2024-06-03 10:10:23,884 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1717405688632_0001
2024-06-03 10:10:23,886 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-06-03 10:10:24,148 INFO conf.Configuration: resource-types.xml not found
2024-06-03 10:10:24,149 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-06-03 10:10:24,653 INFO impl.YarnClientImpl: Submitted application application_1717405688632_0001
2024-06-03 10:10:24,711 INFO mapreduce.Job: The url to track the job: http://ip-172-31-50-182.ec2.internal:20888/proxy/application_1717405688632_0001/
2024-06-03 10:10:24,712 INFO mapreduce.Job: Running job: job_1717405688632_0001
2024-06-03 10:10:32,835 INFO mapreduce.Job: Job job_1717405688632_0001 running in uber mode : false
2024-06-03 10:10:32,837 INFO mapreduce.Job:  map 0% reduce 0%
```



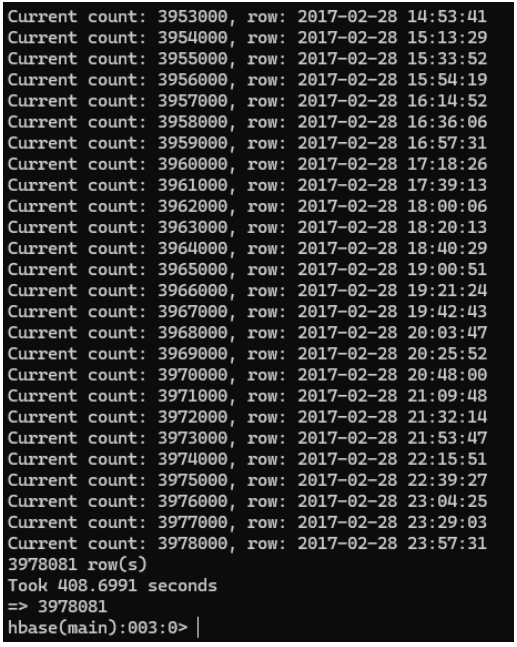
```
2024-06-03 10:10:24,149 INFO resource.ResourceUtils: Unable to find
2024-06-03 10:10:24,653 INFO impl.YarnClientImpl: Submitted applicat
2024-06-03 10:10:24,711 INFO mapreduce.Job: The url to track the job
2024-06-03 10:10:24,712 INFO mapreduce.Job: Running job: job_1717405
2024-06-03 10:10:32,835 INFO mapreduce.Job: Job job_1717405688632_00
2024-06-03 10:10:32,837 INFO mapreduce.Job:  map 0% reduce 0%
```

5) Check the count of the Hbase table

```
hbase shell
count ‘trip_log_hbase’
```



```
hbase(main):001:0> list
TABLE
trip_log_hbase
1 row(s)
Took 0.9007 seconds
=> ["trip_log_hbase"]
hbase(main):002:0> count 'trip_log_hbase'
Current count: 1000, row: 2017-01-01 00:17:38
Current count: 2000, row: 2017-01-01 00:34:19
Current count: 3000, row: 2017-01-01 00:51:00
Current count: 4000, row: 2017-01-01 01:07:40
Current count: 5000, row: 2017-01-01 01:24:20
Current count: 6000, row: 2017-01-01 01:41:04
```



```
Current count: 3953000, row: 2017-02-28 14:53:41
Current count: 3954000, row: 2017-02-28 15:13:29
Current count: 3955000, row: 2017-02-28 15:33:52
Current count: 3956000, row: 2017-02-28 15:54:19
Current count: 3957000, row: 2017-02-28 16:14:52
Current count: 3958000, row: 2017-02-28 16:36:06
Current count: 3959000, row: 2017-02-28 16:57:31
Current count: 3960000, row: 2017-02-28 17:18:26
Current count: 3961000, row: 2017-02-28 17:39:13
Current count: 3962000, row: 2017-02-28 18:00:06
Current count: 3963000, row: 2017-02-28 18:20:13
Current count: 3964000, row: 2017-02-28 18:40:29
Current count: 3965000, row: 2017-02-28 19:00:51
Current count: 3966000, row: 2017-02-28 19:21:24
Current count: 3967000, row: 2017-02-28 19:42:43
Current count: 3968000, row: 2017-02-28 20:03:47
Current count: 3969000, row: 2017-02-28 20:25:52
Current count: 3970000, row: 2017-02-28 20:48:00
Current count: 3971000, row: 2017-02-28 21:09:48
Current count: 3972000, row: 2017-02-28 21:32:14
Current count: 3973000, row: 2017-02-28 21:53:47
Current count: 3974000, row: 2017-02-28 22:15:51
Current count: 3975000, row: 2017-02-28 22:39:27
Current count: 3976000, row: 2017-02-28 23:04:25
Current count: 3977000, row: 2017-02-28 23:29:03
Current count: 3978000, row: 2017-02-28 23:57:31
3978081 row(s)
Took 408.6991 seconds
=> 3978081
hbase(main):003:0> |
```