MapReduce Programing Assignment

Taks 2:

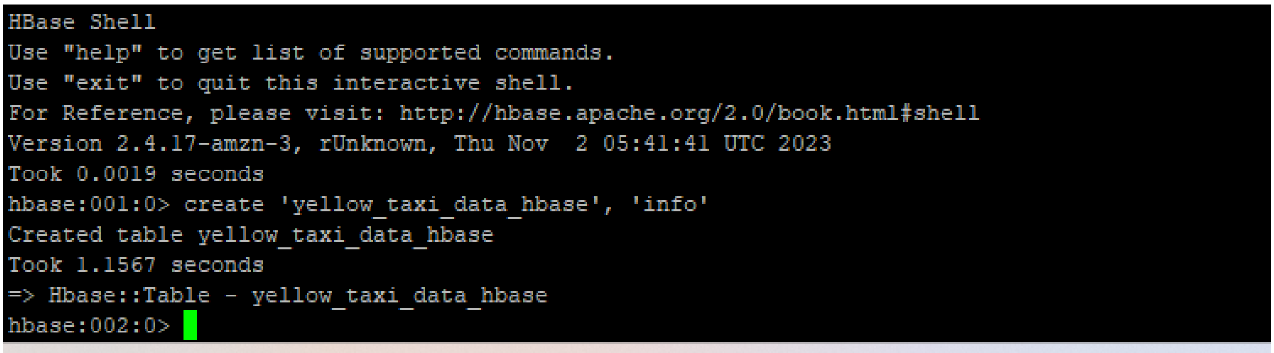
Connect to EMR cluster from local machine using putty.

Give command: **sudo su -l** to enter root user.

In root user give command: **hbase shell** to enter hbase shell.

In hbase shell give below command to create hbase table taxi_data_hbase.

create 'yellow_taxi_data_hbase', 'info'



```
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.4.17-amzn-3, rUnknown, Thu Nov  2 05:41:41 UTC 2023
Took 0.0019 seconds
hbase:001:0> create 'yellow_taxi_data_hbase', 'info'
Created table yellow_taxi_data_hbase
Took 1.1567 seconds
=> Hbase::Table - yellow_taxi_data_hbase
hbase:002:0>
```

Go back to root user by giving command exit.

In Root user run below sqoop command to ingest data from RDS table to hbase table.

sqoop import \

--connect jdbc:mysql://rdsinstance.ckadi12cwuvi.us-east-1.rds.amazonaws.com:3306/demoDB \

--username admin \

--password GURUkula123! \

--query "SELECT VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, trip_distance, total_amount, passenger_count, VendorID || '_' || tpep_pickup_datetime || '_' || tpep_dropoff_datetime || '_' || PULocationID || '_' || DOLocationID || '_' || trip_distance || '_' || total_amount AS row_key FROM yellow_taxi_data WHERE \\$CONDITIONS" \

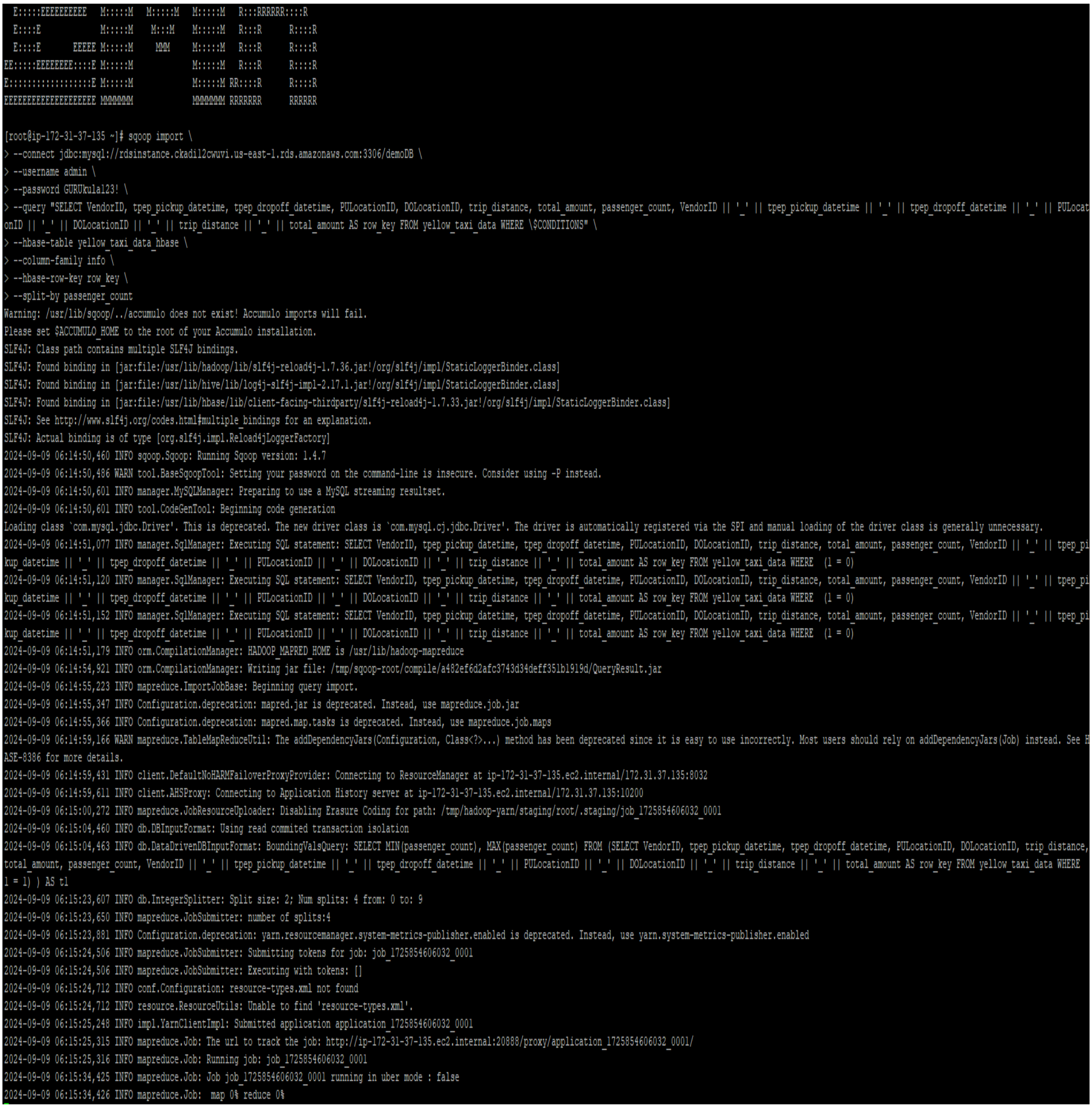
--hbase-table yellow_taxi_data_hbase \

--column-family info \

--hbase-row-key row_key \

--split-by passenger_count

We are using combination of VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, trip_distance, total_amount to get unique values for row_key for hbase table. We are using passenger_count as split-by column.



```
  E:::::EEEEEEEEE   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-172-31-37-135 ~]# sqoop import \
> --connect jdbc:mysql://rdsinstance.ckadil2cwuvi.us-east-1.rds.amazonaws.com:3306/demoDB \
> --username admin \
> --password GURUkula123! \
> --query "SELECT VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, trip_distance, total_amount, passenger_count, VendorID || '_' || tpep_pickup_datetime || '_' || tpep_dropoff_datetime || '_' || PULocat
onID || '_' || DOLocationID || '_' || trip_distance || '_' || total_amount AS row_key FROM yellow_taxi_data WHERE \$CONDITIONS" \
> --hbase-table yellow_taxi_data_hbase \
> --column-family info \
> --hbase-row-key row_key \
> --split-by passenger_count
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-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-reload4j-1.7.33.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.Reload4jLoggerFactory]
2024-09-09 06:14:50,460 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-09-09 06:14:50,486 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-09-09 06:14:50,601 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-09-09 06:14:50,601 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 generally unnecessary.
2024-09-09 06:14:51,077 INFO manager.SqlManager: Executing SQL statement: SELECT VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, trip_distance, total_amount, passenger_count, VendorID || '_' || tpep_pic
kup_datetime || '_' || tpep_dropoff_datetime || '_' || PULocationID || '_' || DOLocationID || '_' || trip_distance || '_' || total_amount AS row_key FROM yellow_taxi_data WHERE  (1 = 0)
2024-09-09 06:14:51,120 INFO manager.SqlManager: Executing SQL statement: SELECT VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, trip_distance, total_amount, passenger_count, VendorID || '_' || tpep_pic
kup_datetime || '_' || tpep_dropoff_datetime || '_' || PULocationID || '_' || DOLocationID || '_' || trip_distance || '_' || total_amount AS row_key FROM yellow_taxi_data WHERE  (1 = 0)
2024-09-09 06:14:51,152 INFO manager.SqlManager: Executing SQL statement: SELECT VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, trip_distance, total_amount, passenger_count, VendorID || '_' || tpep_pic
kup_datetime || '_' || tpep_dropoff_datetime || '_' || PULocationID || '_' || DOLocationID || '_' || trip_distance || '_' || total_amount AS row_key FROM yellow_taxi_data WHERE  (1 = 0)
2024-09-09 06:14:51,179 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2024-09-09 06:14:54,921 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/a482ef6d2afc3743d34deff351b1919d/QueryResult.jar
2024-09-09 06:14:55,223 INFO mapreduce.ImportJobBase: Beginning query import.
2024-09-09 06:14:55,347 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2024-09-09 06:14:55,366 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-09-09 06:14:59,166 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use incorrectly. Most users should rely on addDependencyJars(Job) instead. See H
ASE-8386 for more details.
2024-09-09 06:14:59,431 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-37-135.ec2.internal/172.31.37.135:8032
2024-09-09 06:14:59,611 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-37-135.ec2.internal/172.31.37.135:10200
2024-09-09 06:15:00,272 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1725854606032_0001
2024-09-09 06:15:04,460 INFO db.DBInputFormat: Using read commited transaction isolation
2024-09-09 06:15:04,463 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(passenger_count), MAX(passenger_count) FROM (SELECT VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, trip_distance,
total_amount, passenger_count, VendorID || '_' || tpep_pickup_datetime || '_' || tpep_dropoff_datetime || '_' || PULocationID || '_' || DOLocationID || '_' || trip_distance || '_' || total_amount AS row_key FROM yellow_taxi_data WHERE
1 = 1) ) AS t1
2024-09-09 06:15:23,607 INFO db.IntegerSplitter: Split size: 2; Num splits: 4 from: 0 to: 9
2024-09-09 06:15:23,650 INFO mapreduce.JobSubmitter: number of splits:4
2024-09-09 06:15:23,881 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2024-09-09 06:15:24,506 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725854606032_0001
2024-09-09 06:15:24,506 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-09 06:15:24,712 INFO conf.Configuration: resource-types.xml not found
2024-09-09 06:15:24,712 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-09 06:15:25,248 INFO impl.YarnClientImpl: Submitted application application_1725854606032_0001
2024-09-09 06:15:25,315 INFO mapreduce.Job: The url to track the job: http://ip-172-31-37-135.ec2.internal:20888/proxy/application_1725854606032_0001/
2024-09-09 06:15:25,316 INFO mapreduce.Job: Running job: job_1725854606032_0001
2024-09-09 06:15:34,425 INFO mapreduce.Job: Job job_1725854606032_0001 running in uber mode : false
2024-09-09 06:15:34,426 INFO mapreduce.Job:  map 0% reduce 0%
```