MapReduce - Programming Assignment

Data Ingestion Task

BY

Dharan R
Sanjana Dutta
Punith Vadla

Ingestion_Task

```
>>sudo -i

>>mkdir HBase

>>ls

>>cd HBase/

>>hbase shell

>>status 'detailed'

>>version

>>table_help


>>create "trip_data","TLC"
```

Sqoop requires MySQL’s JDBC driver to be installed in order to talk to the MySQL database engine. Download the driver and install it in /usr/lib/sqoop/lib.

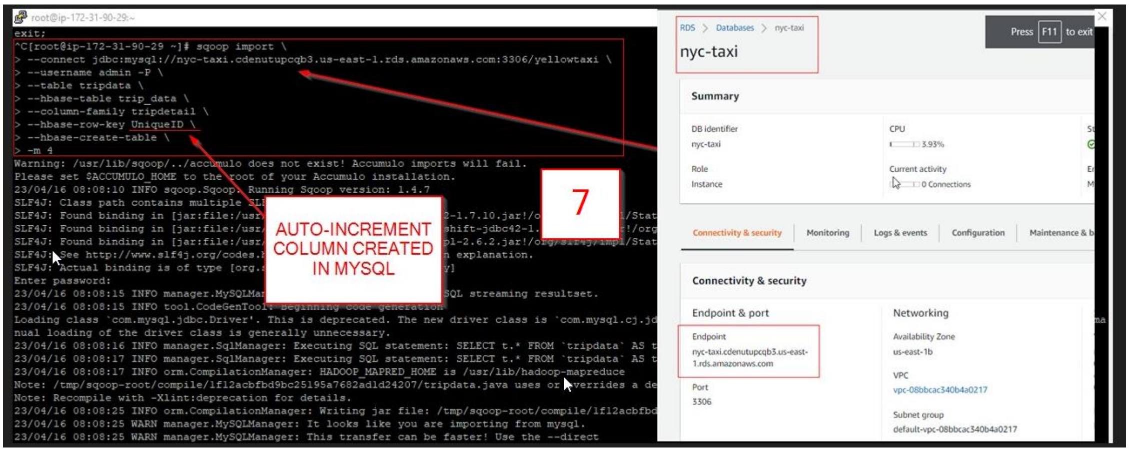
at root run following commands:

```
>>wget http://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-java-5.1.38.tar.gz

>>tar -xvzf mysql-connector-java-5.1.38.tar.gz

>>sudo cp mysql-connector-java-5.1.38/mysql-connector-java-5.1.38-bin.jar /usr/lib/sqoop/lib/
```

Get the tables available

Now that you have a table created, use Sqoop to get a list of tables that are available in the database.

In the following command, use the –P option to be prompted to enter a password. Replace “db-instance-endpoint” with the RDS endpoint retrieved in Step 7, and replace “yourdatabase” with the name of your database in which the “pv_aggregates” table was created.



root@ip-172-31-90-29:~

```
exit;
^C[root@ip-172-31-90-29 ~]# sqoop import \
> --connect jdbc:mysql://nyc-taxi.cdenutupcqb3.us-east-1.rds.amazonaws.com:3306/yellowtaxi \
> --username admin -P \
> --table tripdata \
> --hbase-table trip_data \
> --column-family tripdetail \
> --hbase-row-key UniqueID \
> --hbase-create-table \
> -m 4
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/04/16 08:08:10 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF
SLF4J: Found binding in [jar:file:/usr/                          2-1.7.10.jar!/o              /Stat
SLF4J: Found binding in [jar:file:/usr/                          shift-jdbc42-1.              !/org
SLF4J: Found binding in [jar:file:/usr/                          pl-2.6.2.jar!/org/slf4j/impl/Stat
SLF4J: See http://www.slf4j.org/codes.h                          n explanation.
SLF4J: Actual binding is of type [org.s                          y]
Enter password:
23/04/16 08:08:15 INFO manager.MySQLMan                          SQL streaming resultset.
23/04/16 08:08:15 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jd
nual loading of the driver class is generally unnecessary.
23/04/16 08:08:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `tripdata` AS t
23/04/16 08:08:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `tripdata` AS t
23/04/16 08:08:17 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/1f12acbfbd9bc25195a7682ad1d24207/tripdata.java uses or overrides a de
Note: Recompile with -Xlint:deprecation for details.
23/04/16 08:08:25 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/1f12acbfbd
23/04/16 08:08:25 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/04/16 08:08:25 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
```

7

AUTO-INCREMENT COLUMN CREATED IN MYSQL

nyc-taxi

Summary

DB identifier: nyc-taxi

CPU: 3.93%

Role: Instance

Current activity: 0 Connections

Connectivity & security

Endpoint & port

Endpoint: nyc-taxi.cdenutupcqb3.us-east-1.rds.amazonaws.com

Port: 3306

Networking

Availability Zone: us-east-1b

VPC: vpc-08bbcac340b4a0217

Subnet group: default-vpc-08bbcac340b4a0217

>>sqoop list-tables --connect jdbc:mysql://assignmentdb.cqwodvedrqu4.us-east-1.rds.amazonaws.com:3306/assignmentdb --username admin -P

sqoop import \

> --connect "jdbc:mysql://assignmentdb.cqwodvedrqu4.us-east-1.rds.amazonaws.com:3306/assignmentdb" \

> --table "trip_data" \

> --hbase-table "trip_data" \

> --hbase-row-key "ID" \

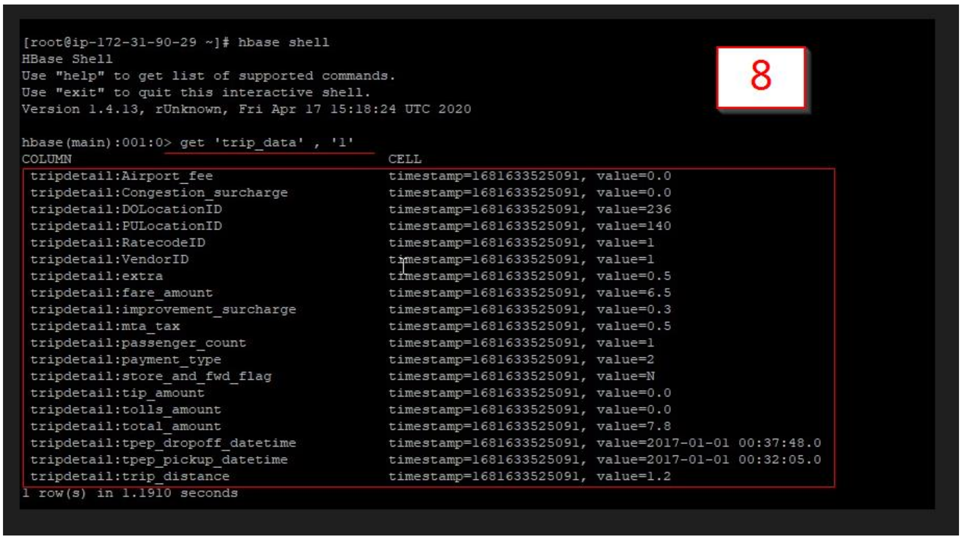
> --column-family "TLC" \

> --hbase-create-table \

> --username admin -P

in hbase run this command to check count of rows:

count 'trip_data', INTERVAL => 100000



```
[root@ip-172-31-90-29 ~]# hbase shell
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

hbase(main):001:0> get 'trip_data' , '1'
COLUMN                                      CELL
 tripdetail:Airport_fee                     timestamp=1681633525091, value=0.0
 tripdetail:Congestion_surcharge            timestamp=1681633525091, value=0.0
 tripdetail:DOLocationID                    timestamp=1681633525091, value=236
 tripdetail:PULocationID                    timestamp=1681633525091, value=140
 tripdetail:RatecodeID                      timestamp=1681633525091, value=1
 tripdetail:VendorID                        timestamp=1681633525091, value=1
 tripdetail:extra                           timestamp=1681633525091, value=0.5
 tripdetail:fare_amount                     timestamp=1681633525091, value=6.5
 tripdetail:improvement_surcharge           timestamp=1681633525091, value=0.3
 tripdetail:mta_tax                         timestamp=1681633525091, value=0.5
 tripdetail:passenger_count                 timestamp=1681633525091, value=1
 tripdetail:payment_type                    timestamp=1681633525091, value=2
 tripdetail:store_and_fwd_flag              timestamp=1681633525091, value=N
 tripdetail:tip_amount                      timestamp=1681633525091, value=0.0
 tripdetail:tolls_amount                    timestamp=1681633525091, value=0.0
 tripdetail:total_amount                    timestamp=1681633525091, value=7.8
 tripdetail:tpep_dropoff_datetime           timestamp=1681633525091, value=2017-01-01 00:37:48.0
 tripdetail:tpep_pickup_datetime            timestamp=1681633525091, value=2017-01-01 00:32:05.0
 tripdetail:trip_distance                   timestamp=1681633525091, value=1.2
1 row(s) in 1.1910 seconds
```

8