**Hadoop-hdfs- scoop—hive Project**

**Creating a Data Base with MySQL engine using AWS RDS:**

**Create a Data Base->engine-MySQL**

Set

Templates to Free tier

Uncheck enable storage autoscaling

Publice access to yes

### Select the databse and see the Endpoint & port

Endpoint- database-2.c9dutbh8ufvz.us-east-1.rds.amazonaws.com

Save this endpoint information as this will be used as a hostname to connect to this data base instance(AWS RDS) from MYSQL workbench

Security

VPC security groups

Click the link below the VPC security groups and an inbound rule

Set Type to All Traffic and Source to Anywhere Ipv4

**MySQL Workbench: To add the table to AWSRDS database**

**Set up a new connection**

Give a connection name

In hostname: endpoint that we saved earlier

Password is the password of that AWS data base instance

->Testconnection

Once the connection is set up, we can the see the data base created and we can import our csv file to this DB using

->DB

->right click on table

->Table data import wizard

Once the table is imported, Open Virtual Box termiunal using localhost:4200

LOGIN TO ROOT USER

->sandbox-hdp login: root

->password:

We use scoop to import data from external RDBMS like SQL server to HIVE

**RUN THIS COMMAND ON root user (NOT ON HIVE):**

USING SCOOP TO IMPORT TABLE FROM RDBMS TO HIVE , TABLE IS AUTOMATICALLY SAVED ON DEFAULT DB ON HIVE

**sqoop import --connect 'jdbc:mysql://database-2.c9dutbh8ufvz.us-east 1.rds.amazonaws.com/AWSDB?autoReconnect=true&useSSL=false' --table yellow\_tripdata -m1 --username admin -P --hive-import**

**database-2.c9dutbh8ufvz.us-east 1.rds.amazonaws.com/AWSDB-** Endpoint of AWS RDS database

**yellow\_tripdata- Table name that we imported on MYSQL workbench**

**Running this command will further ask for password which will be the Amazon RDS DB password that we give while creating DB on RDS.**

Type hive to enter hive on Virtual Box liunux environment terminal:

Show databases;

Use default;

Show tables;

**Run the following queries:**

1. what is the total number of trips(equal to the number of rows)?

select count(\*) from yellowtripdata;

Graphical user interface, text, application, email

Description automatically generated

1. what is the total revenue generated by all the trips?

select sum(total\_amount) from yellowtripdata;

A screenshot of a computer

Description automatically generated

1. What fraction of the total is paid for tolls?

select sum(tolls\_amount)/sum(total\_amount) from yellowtripdata;

Graphical user interface, text, application, email

Description automatically generated

1. What fraction of it is driver tips?

select sum(tip\_amount)/sum(total\_amount) from yellowtripdata;

Graphical user interface, text, application, email

Description automatically generated

1. What is the average trip amount?

select sum(total\_amount)/count(\*) from yellowtripdata;

A screenshot of a computer

Description automatically generated

1. What is the average distance of the trips?

select sum(trip\_distance)/count(\*) from yellowtripdata;

Graphical user interface, text, application, email

Description automatically generated

1. How many different payment types are used?

Select count(distinct payment\_type) from yellowtripdata;

Graphical user interface, text, application, email

Description automatically generated

1. For each payment type display the following details:
2. average fare generated

select payment\_type,avg(fare\_amount) from yellowtripdata group by payment\_type order by payment\_type;

Graphical user interface, text, application

Description automatically generated

1. average tip

select payment\_type, avg(tip\_amount) from yellowtripdata group by payment\_type order by payment\_type;

A screenshot of a computer

Description automatically generated

1. average tax

select payment\_type, avg(mta\_tax) from yellowtripdata group by payment\_type order by payment\_type;

Graphical user interface, text, application, email

Description automatically generated

1. On average, which hour of the day generates the highest revenue?

select hour(tpep\_pickup\_datetime) from yellowtripdata group by hour(tpep\_pickup\_datetime) order by avg(total\_amount) desc limit 1;

Graphical user interface, text, application, email

Description automatically generated