Visualization of Yellow Trip Taxi (Task 5)

Table of Contents

Visualization of Yellow Trip Taxi (Task 5)

Creation of Schema on MySQL (RDS Instance) for storing Outputs

Below DDL is used to create the desired table to store data: -

```
CREATE TABLE mrtask_a
(
      VendorID                Varchar(20)      NOT NULL,
      Vendor_Revenue          DECIMAL(16,4)    NOT NULL
);
CREATE TABLE mrtask_b
(
      PULocationID            Varchar(20)      NOT NULL,
      PULocationID_Revenue    DECIMAL(16,4)    NOT NULL
);
CREATE TABLE mrtask_c
(
      payment_type            Varchar(20)      NOT NULL,
      number_of_transaction   DECIMAL(16,4)    NOT NULL
);
CREATE TABLE mrtask_d
(
      PULocationID            Varchar(20)      NOT NULL,
      avgtriptime_minutes     DECIMAL(16,4)    NOT NULL
);
CREATE TABLE mrtask_e
(
      PULocationID            Varchar(20)      NOT NULL,
      ratio_tiprevenue        DECIMAL(16,4)    NOT NULL
);
CREATE TABLE mrtask_f
(
      Month_Weekday_Day       Varchar(20)      NOT NULL,
      avgtriprevenue          DECIMAL(16,4)    NOT NULL
);
```

```
[root@ip-172-31-27-72 tripdata]# mysql -h $DNS_RDS -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 8.0.33 Source distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]> use tripdata
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [tripdata]> CREATE TABLE mrtask_a
    -> (
    ->       VendorID               Varchar(20)      NOT NULL,
    ->       Vendor_Revenue         DECIMAL(16,4)    NOT NULL
    -> );
);
CREATE TABLE mrtask_e
(
      PULocationID           Varchar(20)      NOT NULL,
      ratio_tiprevenue       DECIMAL(16,4)    NOT NULL
);
CREATE TABLE mrtask_f
(
      Month_Weekday_Day      Varchar(20)      NOT NULL,
      avgtriprevenue         DECIMAL(16,4)    NOT NULL
);
Query OK, 0 rows affected (0.03 sec)

MySQL [tripdata]> CREATE TABLE mrtask_b
    -> (
    ->       PULocationID           Varchar(20)      NOT NULL,
    ->       PULocationID_Revenue   DECIMAL(16,4)    NOT NULL
    -> );
Query OK, 0 rows affected (0.03 sec)

MySQL [tripdata]> CREATE TABLE mrtask_c
    -> (
    ->       payment_type           Varchar(20)      NOT NULL,
    ->       number_of_transaction  DECIMAL(16,4)    NOT NULL
    -> );
Query OK, 0 rows affected (0.03 sec)
```

```
MySQL [tripdata]> CREATE TABLE mrtask_d
    -> (
    ->       PULocationID           Varchar(20)      NOT NULL,
    ->       avgtriptime_minutes    DECIMAL(16,4)    NOT NULL
    -> );
Query OK, 0 rows affected (0.03 sec)

MySQL [tripdata]> CREATE TABLE mrtask_e
    -> (
    ->       PULocationID           Varchar(20)      NOT NULL,
    ->       ratio_tiprevenue       DECIMAL(16,4)    NOT NULL
    -> );
Query OK, 0 rows affected (0.03 sec)

MySQL [tripdata]> CREATE TABLE mrtask_f
    -> (
    ->       Month_Weekday_Day      Varchar(20)      NOT NULL,
    ->       avgtriprevenue         DECIMAL(16,4)    NOT NULL
    -> );
Query OK, 0 rows affected (0.03 sec)
```

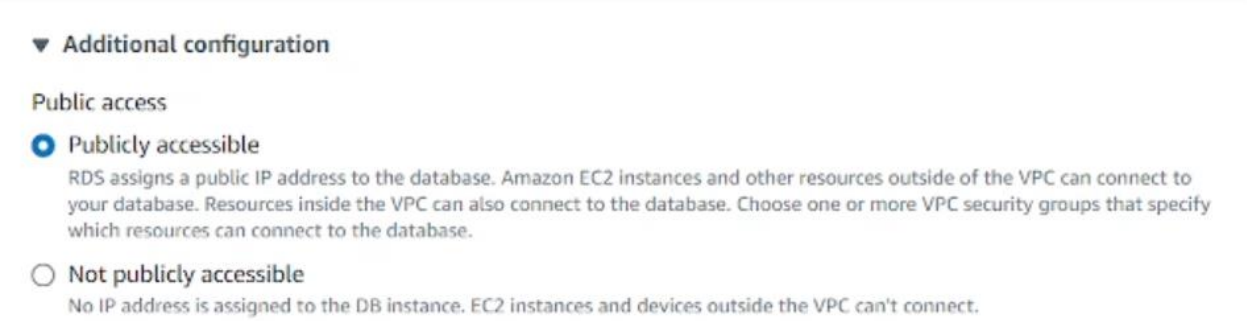
Export of Outputs of MR Tasks performed to MySQL (RDS Instance) using SQOOP

Below SQOOP commands are used to export outputs of MR Tasks (from 1 to 6) performed from HDFS to MySQL (RDS Instance):

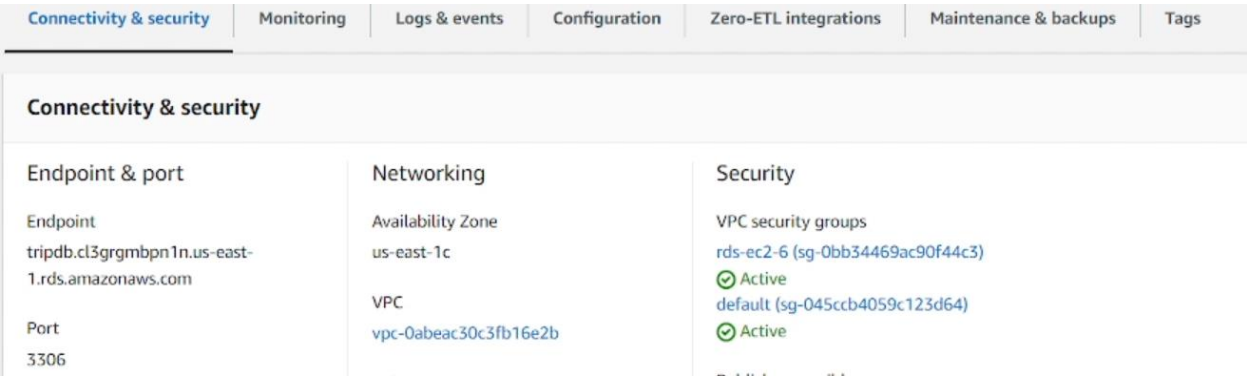
```
sqoop export -D org.apache.sqoop.splitter.allow_text_splitter=true \
--connect jdbc:mysql://$DNS_RDS:3306/tripdata \
--username admin --password-file /user/root/tripdata/password.txt \
--table mrtask_a \
--export-dir /user/root/tripdata/out_mrtask_a \
--fields-terminated-by '\t' \
--lines-terminated-by '\n'

sqoop export -D org.apache.sqoop.splitter.allow_text_splitter=true \
--connect jdbc:mysql://$DNS_RDS:3306/tripdata \
--username admin --password-file /user/root/tripdata/password.txt \
--table mrtask_b \
--export-dir /user/root/tripdata/out_mrtask_b \
--fields-terminated-by '\t' \
--lines-terminated-by '\n'

sqoop export -D org.apache.sqoop.splitter.allow_text_splitter=true \
--connect jdbc:mysql://$DNS_RDS:3306/tripdata \
--username admin --password-file /user/root/tripdata/password.txt \
--table mrtask_c \
--export-dir /user/root/tripdata/out_mrtask_c \
--fields-terminated-by '\t' \
--lines-terminated-by '\n'

sqoop export -D org.apache.sqoop.splitter.allow_text_splitter=true \
--connect jdbc:mysql://$DNS_RDS:3306/tripdata \
--username admin --password-file /user/root/tripdata/password.txt \
--table mrtask_d \
--export-dir /user/root/tripdata/out_mrtask_d \
--fields-terminated-by '\t' \
--lines-terminated-by '\n'

sqoop export -D org.apache.sqoop.splitter.allow_text_splitter=true \
--connect jdbc:mysql://$DNS_RDS:3306/tripdata \
--username admin --password-file /user/root/tripdata/password.txt \
--table mrtask_e \
--export-dir /user/root/tripdata/out_mrtask_e \
--fields-terminated-by '\t' \
--lines-terminated-by '\n'

sqoop export -D org.apache.sqoop.splitter.allow_text_splitter=true \
--connect jdbc:mysql://$DNS_RDS:3306/tripdata \
--username admin --password-file /user/root/tripdata/password.txt \
--table mrtask_f \
--export-dir /user/root/tripdata/out_mrtask_f \
--fields-terminated-by '\t' \
--lines-terminated-by '\n'
```

Changing Security Group Settings to allow access from Internet

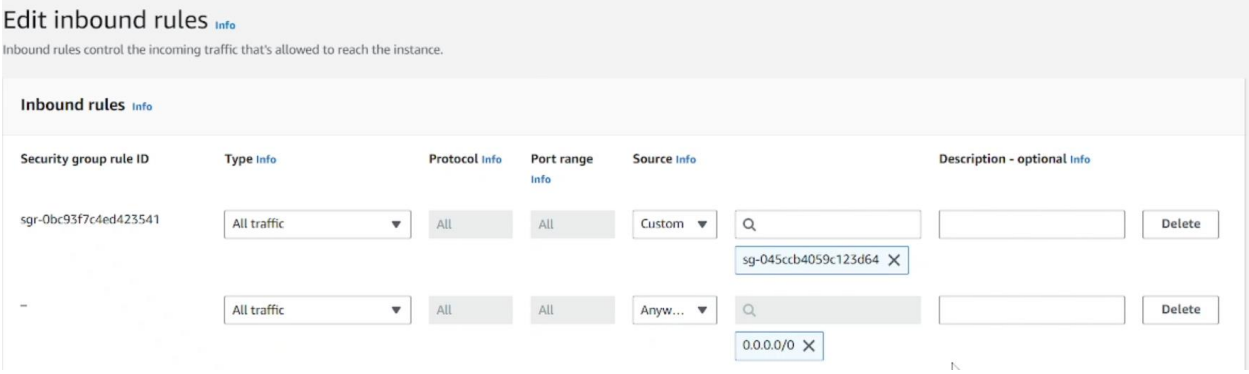
- Modify the RDS Instance setting using AWS console.
- Change Public access under Additional configuration in Connectivity haed:



- Choose option “Apply Immediately” to get it effective instantaneously.
- Additionally, add “Inbound Rule” in security group. For this select “default” security group under “Connectivity & security” on RDS database configuration page.

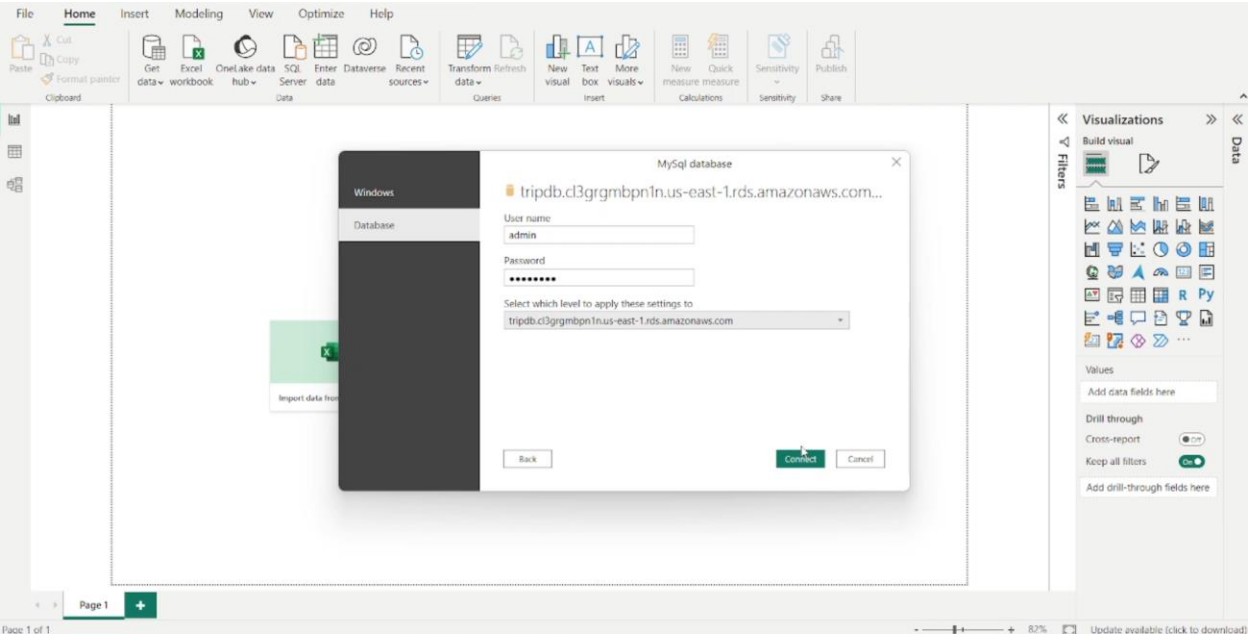


- Add inbound rule to allow “all traffic” from “Anywhere” under “Edit Inbound Rules”:

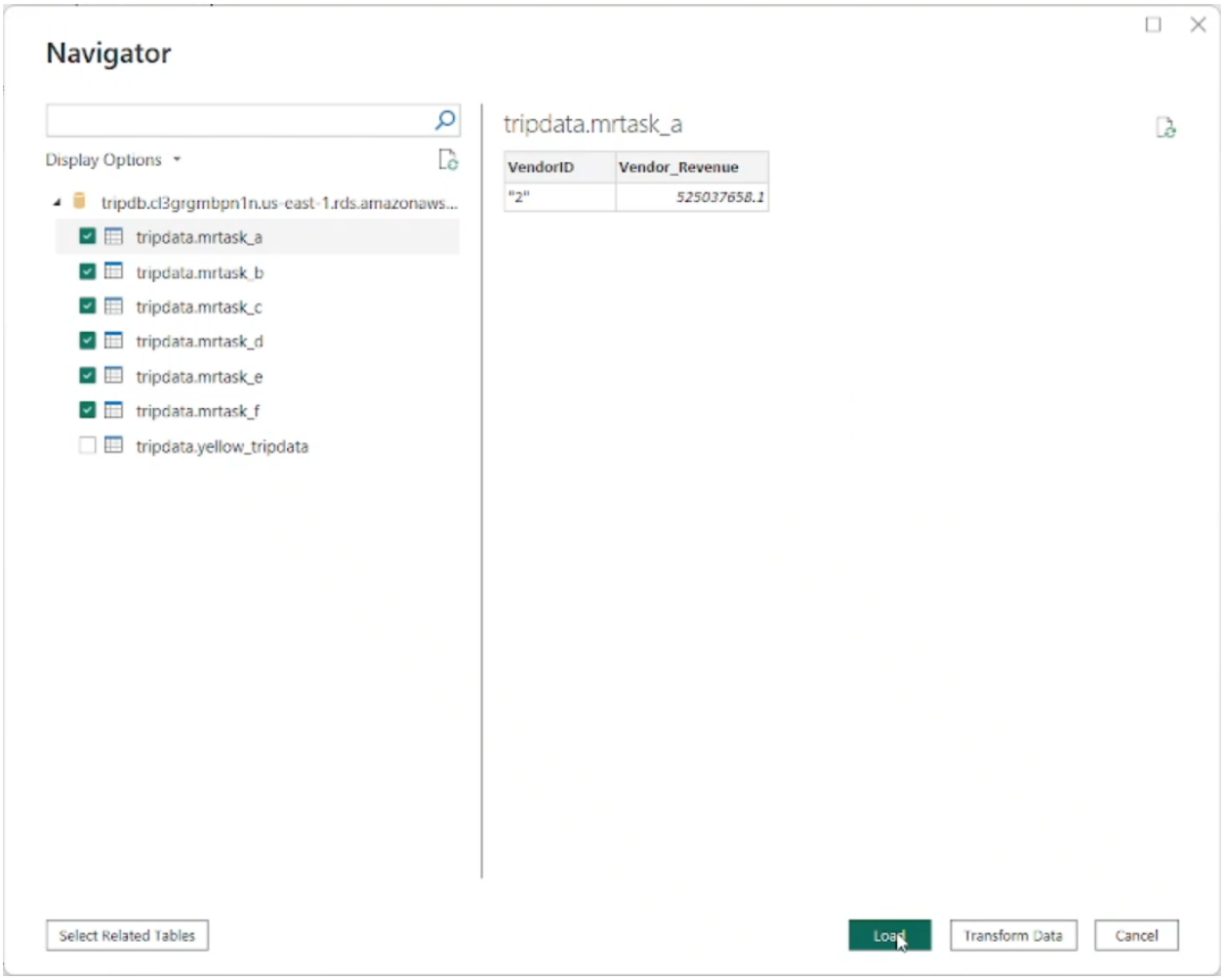


Establish connection between RDS Instance and Power BI (Desktop)

- On start of Power BI add credentials and connectivity url of RDS instance:



- Select the required tables and “Load”:



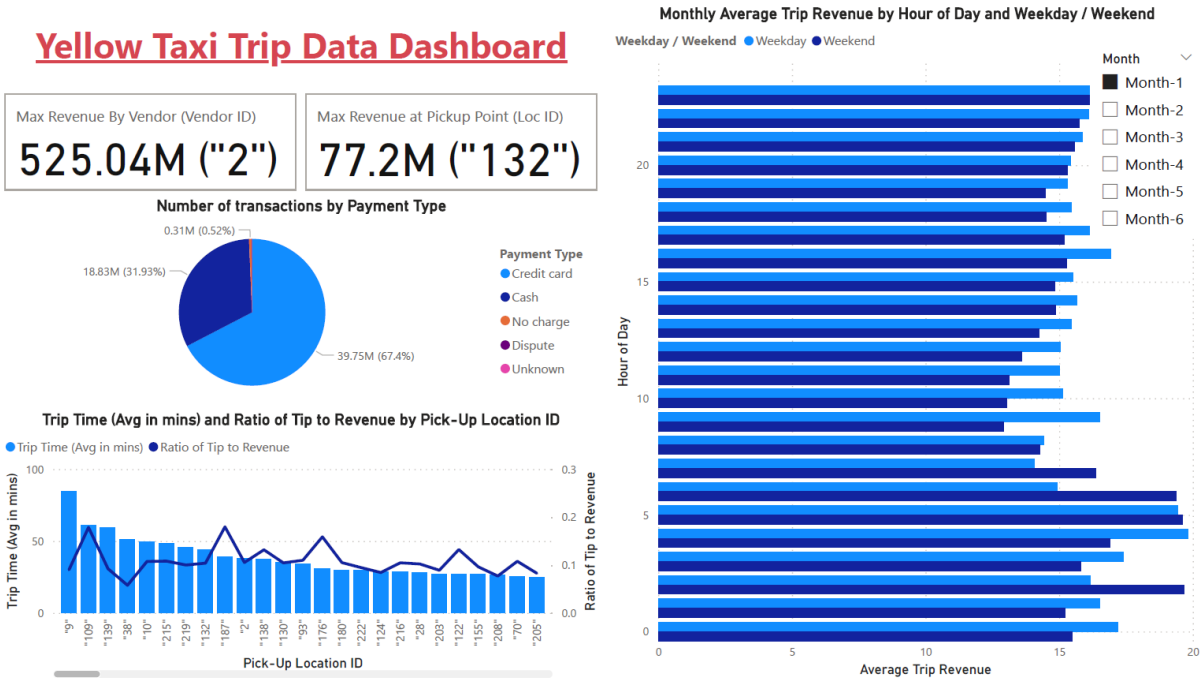
Transformations using "Power Query Editor"

Following small transformations are done using "Power Query Editor":

- For mrtask_a:
  - Add additional column that is concatenation of Vendor Revenue in millions with VendorID in bracket.
- For mrtask_b:
  - Add additional column that is concatenation of Revenue in millions with Pick-up location ID in bracket.
- For mrtask_c:
  - Mapping following code to values as mentioned in column description document:
    - "1" : Credit card
    - "2" : Cash
    - "3" : No charge
    - "4" : Dispute
    - "5" : Unknown

- For mrtask_f:
    - Split key to Month, Weekend/Weekday, and Hour of Day

Final Dashboard:



Final Observations:

Following observations are made through Dashboard:

- Vendor “Verifone Inc.” (code “2”) made maximum revenue among two vendors and total revenue made during first six months of 2017 is 525.04 million.
- Pickup point “132” generate maximum revenue of 77.2 million during first six months of 2017.
- Approx. 2/3 of transaction are done using “Credit card” as payment type. There is very less cases of “No Charge”, “Dispute”, and “Unknown”. There is no instance found for “Voided trip”.
- Average Trip Time is maximum for pick-up location ID “9” followed by “109”, and “139”.
- Among top 25 pick-up locations with average trip time, pick-up location ID “109”, “187”, and “176” have ‘Tip to Revenue’ ration higher than 50%.
- Average Trip Revenue is higher during 04-07 hour of days. Also, there was drop in average trip revenue from Jan-2017 to Feb-2017, and then start growing for day hours from 07 to 22 hour.