Queries for Hive Case study

Tasks:

1. Create a table named taxidata . Required ddl script is given below.

Create database taxi;

Use taxi;

CREATE TABLE IF NOT EXISTS taxidata

(vendor\_id string, pickup\_datetime string,

dropoff\_datetime string, passenger\_count int, trip\_distance DECIMAL(9,6),

pickup\_longitude DECIMAL(9,6), pickup\_latitude DECIMAL(9,6), rate\_code int,

store\_and\_fwd\_flag string, dropoff\_longitude DECIMAL(9,6), dropoff\_latitude

DECIMAL(9,6),

payment\_type string, fare\_amount DECIMAL(9,6), extra DECIMAL(9,6),

mta\_tax DECIMAL(9,6), tip\_amount DECIMAL(9,6), tolls\_amount DECIMAL(9,6),

total\_amount DECIMAL(9,6), trip\_time\_in\_secs int )

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED as TEXTFILE

TBLPROPERTIES ("skip.header.line.count"="1");

2. Load data from the csv file - yellow\_tripdata\_2015-01-06.csv

LOAD DATA INPATH '/user/paras/taxi' OVERWRITE INTO TABLE taxidata;

3. Run some basic queries to check the data is loaded properly.

Query: Select \* from taxidata;

4. Run the queries required to answer the following questions.

Problem statement:

Use the above data to come up with answers to these questions:

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

Query: Select count(\*) from taxidata;

2. What is the total revenue generated by all the trips ? Fare is stored in the column

total\_amount.

Select sum(total\_amount) as total\_revenue from taxidata;

3. What fraction of the total is paid for tolls? Toll is stored in tolls\_amount.

Select sum(tolls\_amount)/sum(total\_amount) as toll\_pct from taxidata;

4. What fraction of it is driver tips? Tip is stored in tip\_amount.

Select sum(tip\_amount)/sum(total\_amount) as tip\_pct from taxidata;

5. What is the average trip amount?

Select avg(total\_amount) as avg\_tripamount from taxidata;

6. For each payment type, display the following details

i. Average fare generated – fare amount is stored in fare\_amount

ii. Average tip

iii. Average tax – tax is stored in column mta\_tax

query : select payment\_type,

avg(fare\_amount) as average\_fare,

avg(tip\_amount) as average\_tip,

avg(mta\_tax) as average\_tax,

from taxidata

group by payment\_type;

7. On an average which hour of the day generates the highest revenue?

select h24 as hour,

avg(total\_amount) as avg\_revenue

from (select hour(pickup\_datetime) as h24,

total\_amount

from taxidata)

ff

group by h24

order by avg\_revenue desc;

8. What is the average distance of the trips? Distance is stored in the column trip\_distance.

select

avg(trip\_distance) as avg\_distance

from trips4;

9. How many different payment types are used? Column name – payment\_type.

select distinct payment\_type from taxidata;