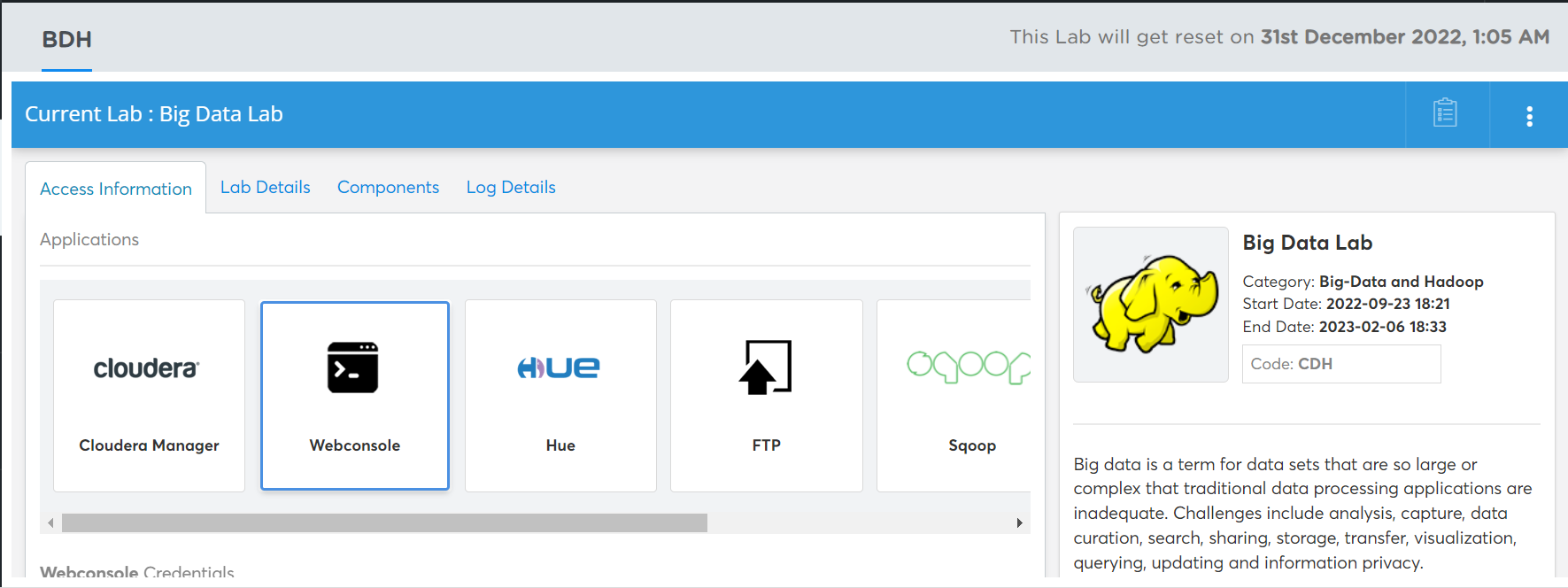
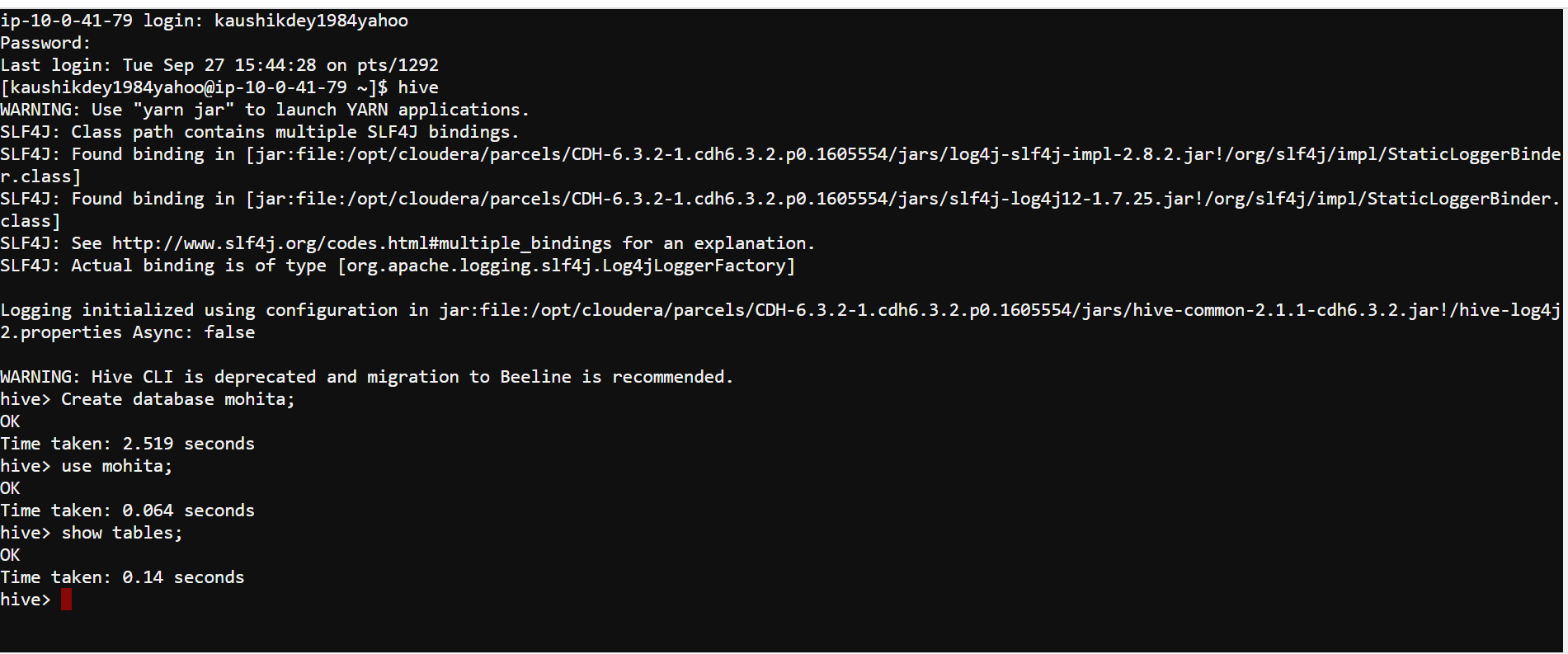
**First-Step:** Before Executing any queries inside HDFS, HIVE first we must prepare the Lab. In this project I have used Simplilearn big data cloudera distribution lab.



So from web-Console first login into the web console with proper user name and password.

Username: kaushikdey1984yahoo

Password : kaushikdey1984yahooodeqx

After that from web-console go to the hive shell with following screen-shot commands.

**Second-Step:** Load Yellow Taxi trip csv data from local system to hdfs via FTP.

Graphical user interface

Description automatically generated

Now the tick checkbox data is from yellow\_tripdata\_2015-01-06.csv.

Graphical user interface

Description automatically generated with medium confidence

We have to confirm that from the webconsole shell also. This is the following screen-shot.

A picture containing text

Description automatically generated

Now move this csv file from userpath to hdfs with following commands.

1) Hadoop fs -mkdir newDataFlair

2) Hadoop fs -put yellow\_tripdata\_2015-01-06.csv newDataFlair

3) Hadoop fs -ls newDataFlair

Text

Description automatically generated

**Third-Step:** Now change the shell ( that is Hive) and execute the following commands:

1. Create database mohita;
2. Use mohita;
3. Hive>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");
4. Hive>Show tables;

Text

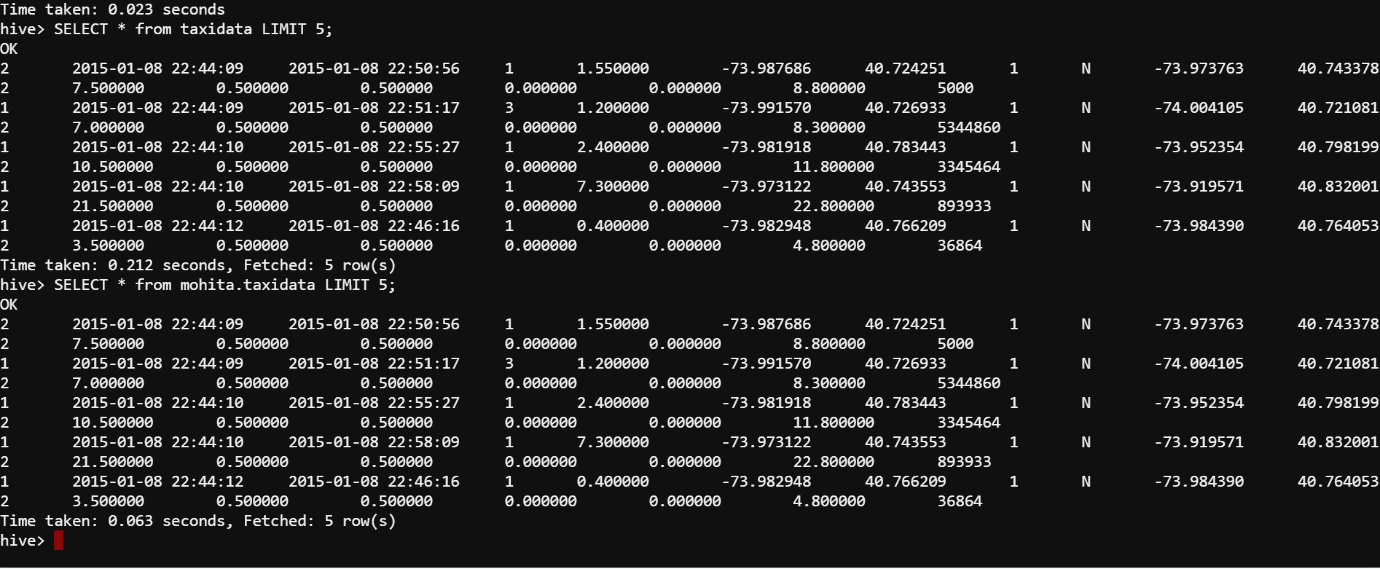
Description automatically generated

Now we have to load data from hdfs to hive shell.

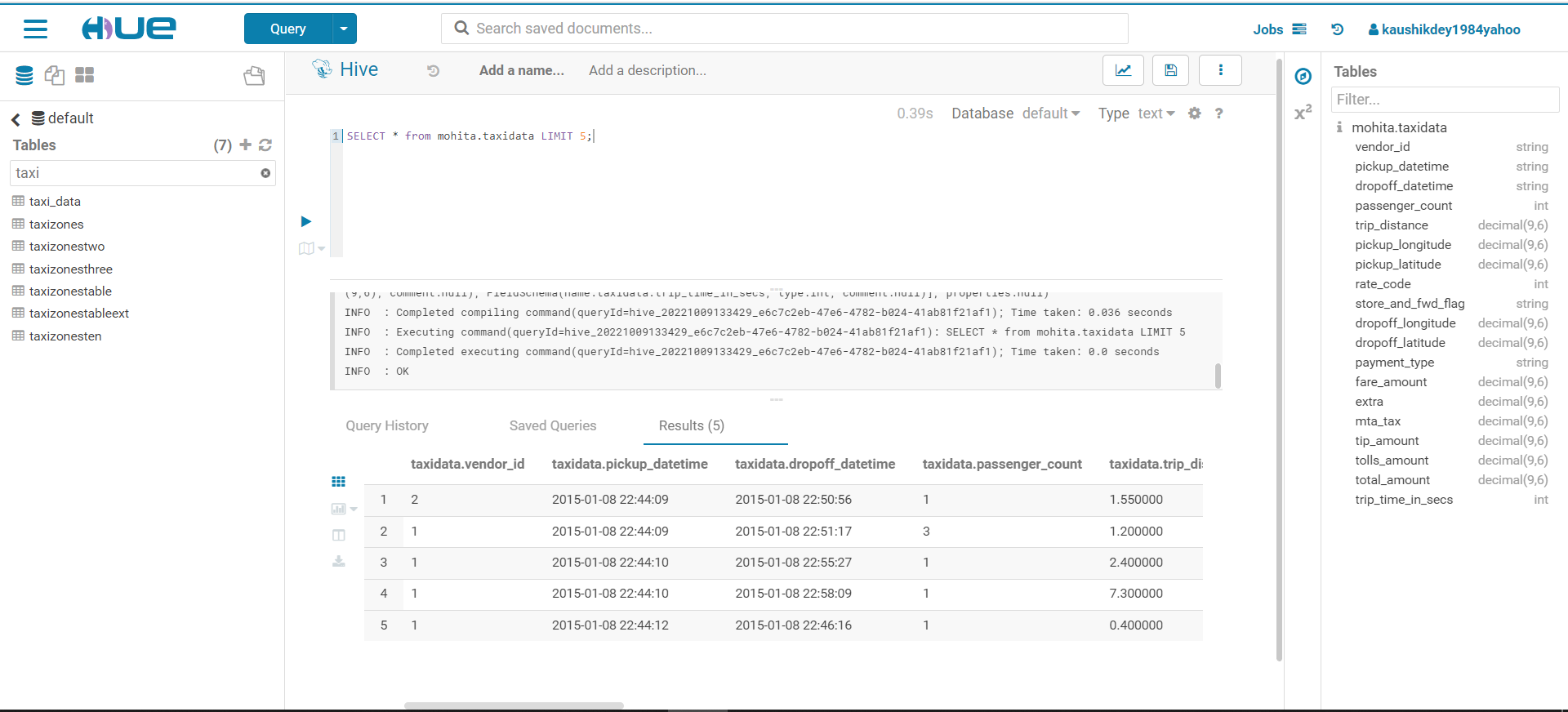
1. Hive> LOAD DATA INPATH 'newDataFlair/yellow\_tripdata\_2015-01-06.csv' INTO TABLE taxidata;

Now we have to check that from Hive shell & Hue Editor. ( It is from UI SIDE)

1. Hive> select \* from mohita.taxidata LIMIT 5;

ScreenShot from Hive Shell;

ScreenShot from Hue Editor;



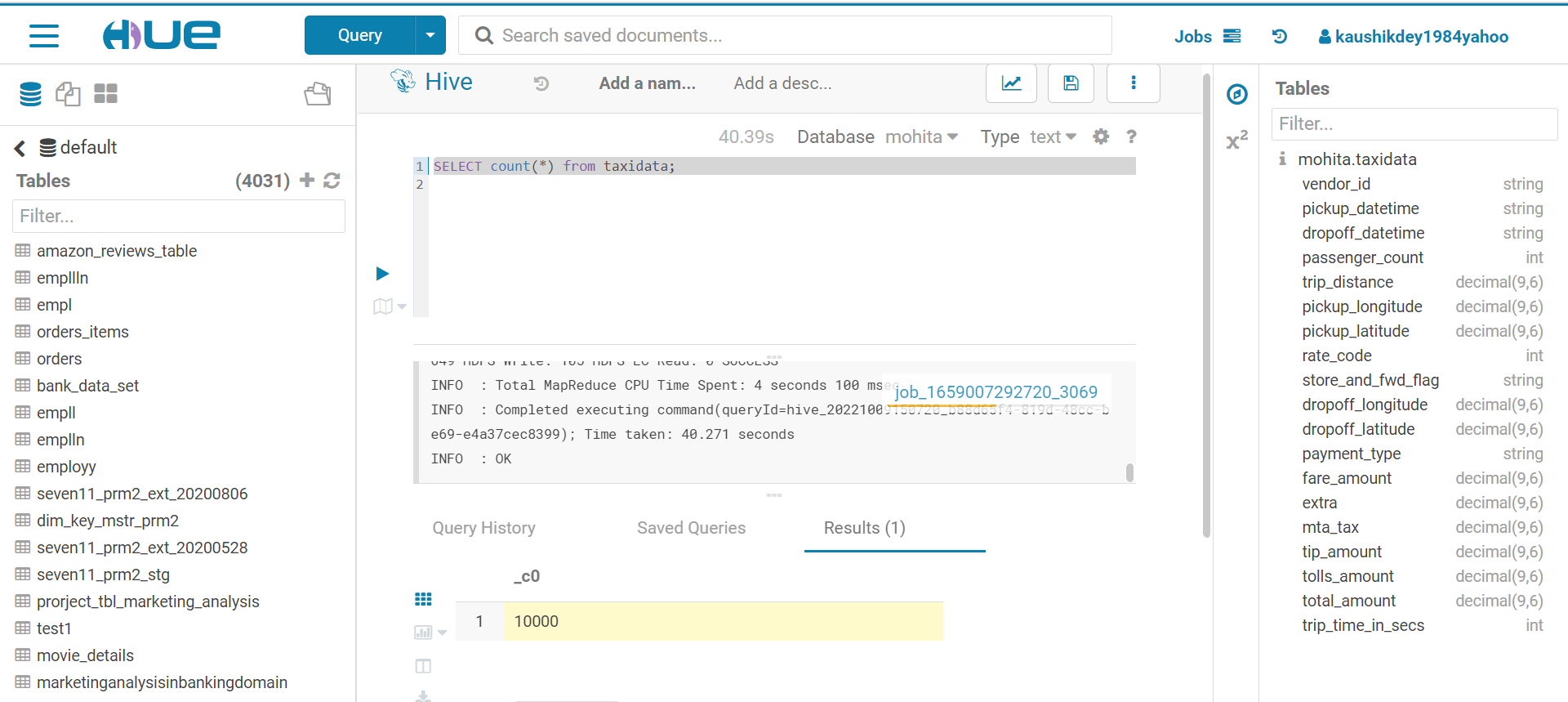
**Fourth-Step:** in Fourth step we have to analysis the following questions & answers.

**Perform taxi trip analysis by solving the questions below:**

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

SELECT count(\*) from mohita.taxidata;

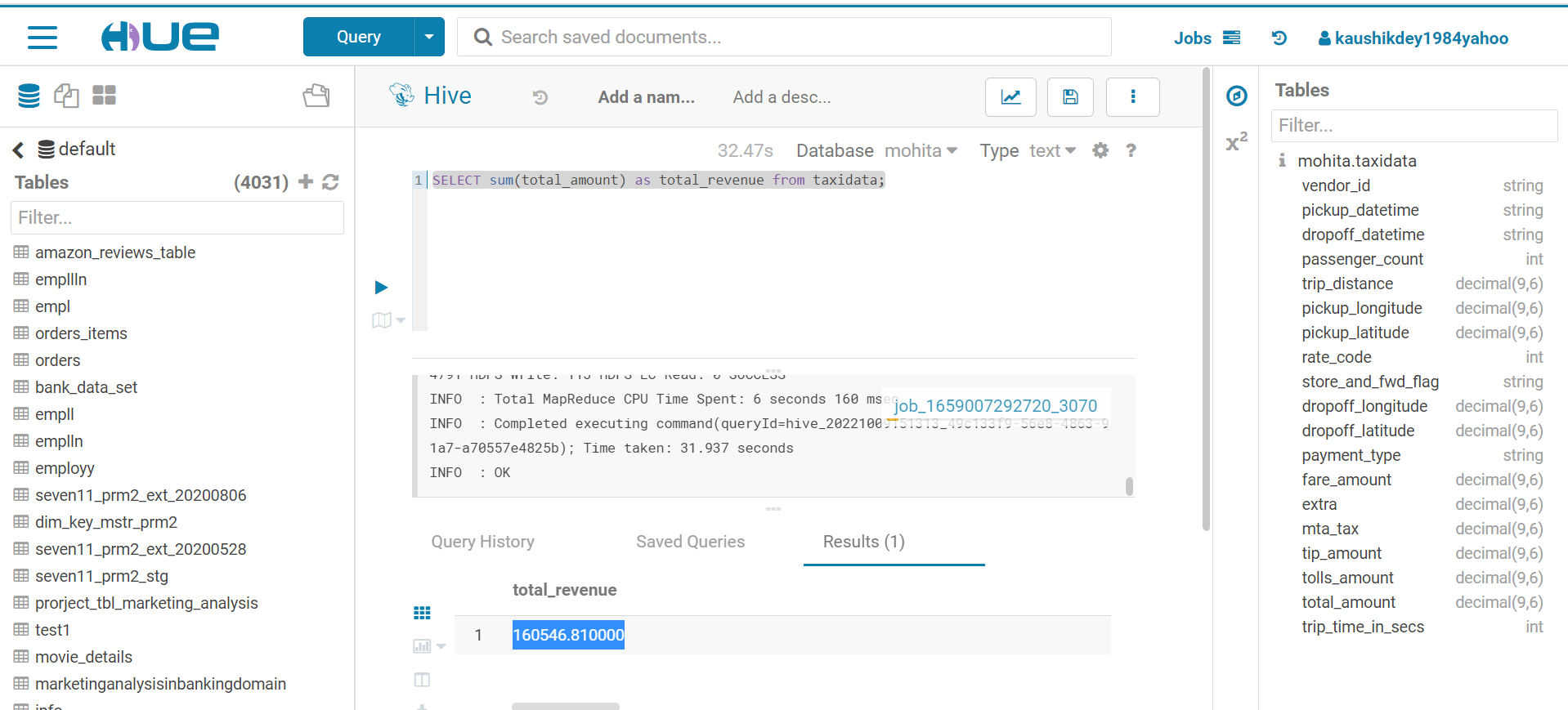
Output : 10000



1. What is the total revenue generated by all the trips? The fare is stored in the column total\_amount.

Ans : SELECT sum(total\_amount) as total\_revenue from taxidata;

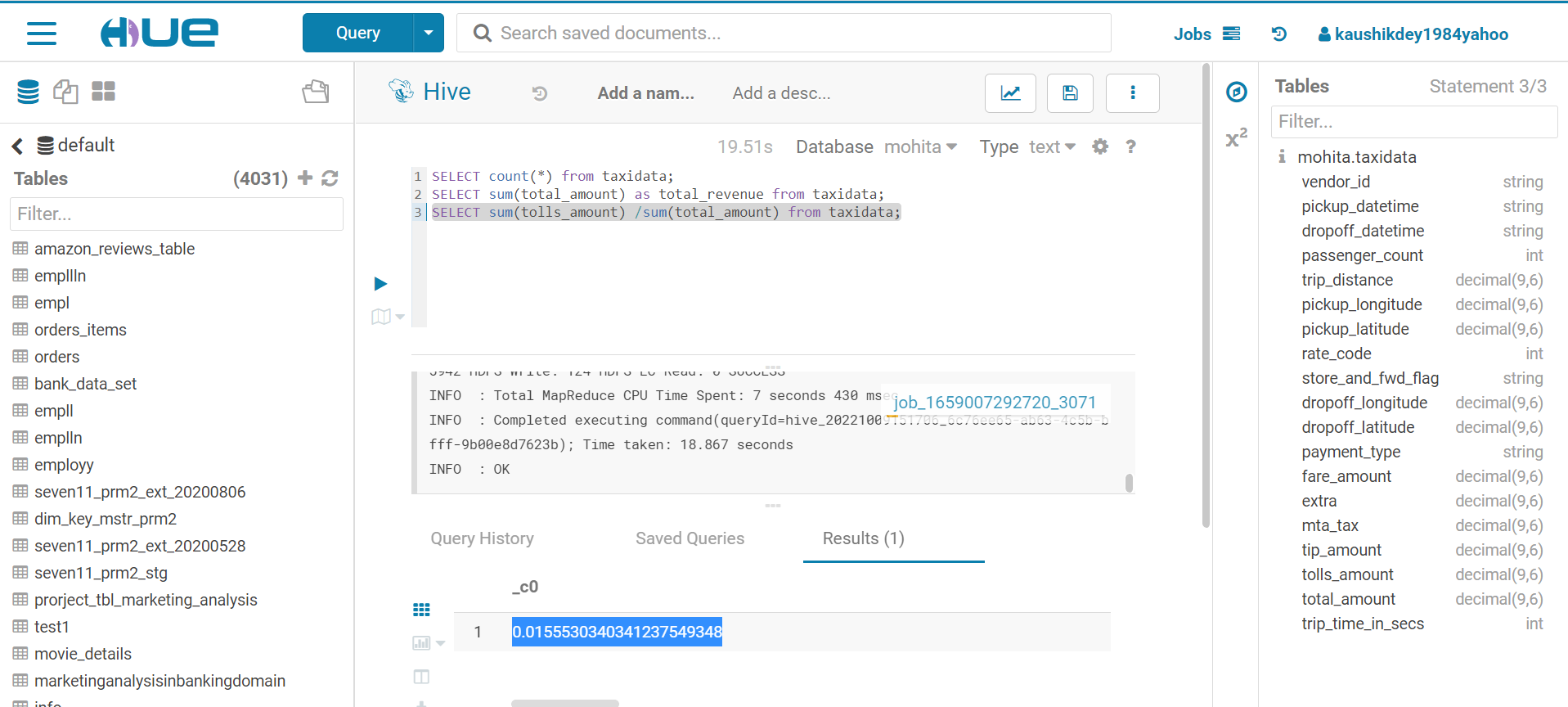
Output : 160546.810000



1. What fraction of the total is paid for tolls? The toll is stored in tolls\_amount.

Ans : SELECT sum(tolls\_amount) /sum(total\_amount) from taxidata;

Output : 0.0155530340341237549348



1. What fraction of it is driver tips? The tip is stored in tip\_amount.

Ans : SELECT sum(tip\_amount) / sum (total\_amount) from taxidata;

Output : 0.1078520339332808917225

Graphical user interface, text, application

Description automatically generated

1. What is the average trip total amount?

Ans : SELECT avg(total\_amount) as trip\_amount from taxidata;

Output : 16.0546810000

Graphical user interface, text, application

Description automatically generated

1. For each payment type, display the following details?
2. Average fare generated
3. Average tip
4. Average tax – tax is stored in column mta\_tax

Ans : 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;

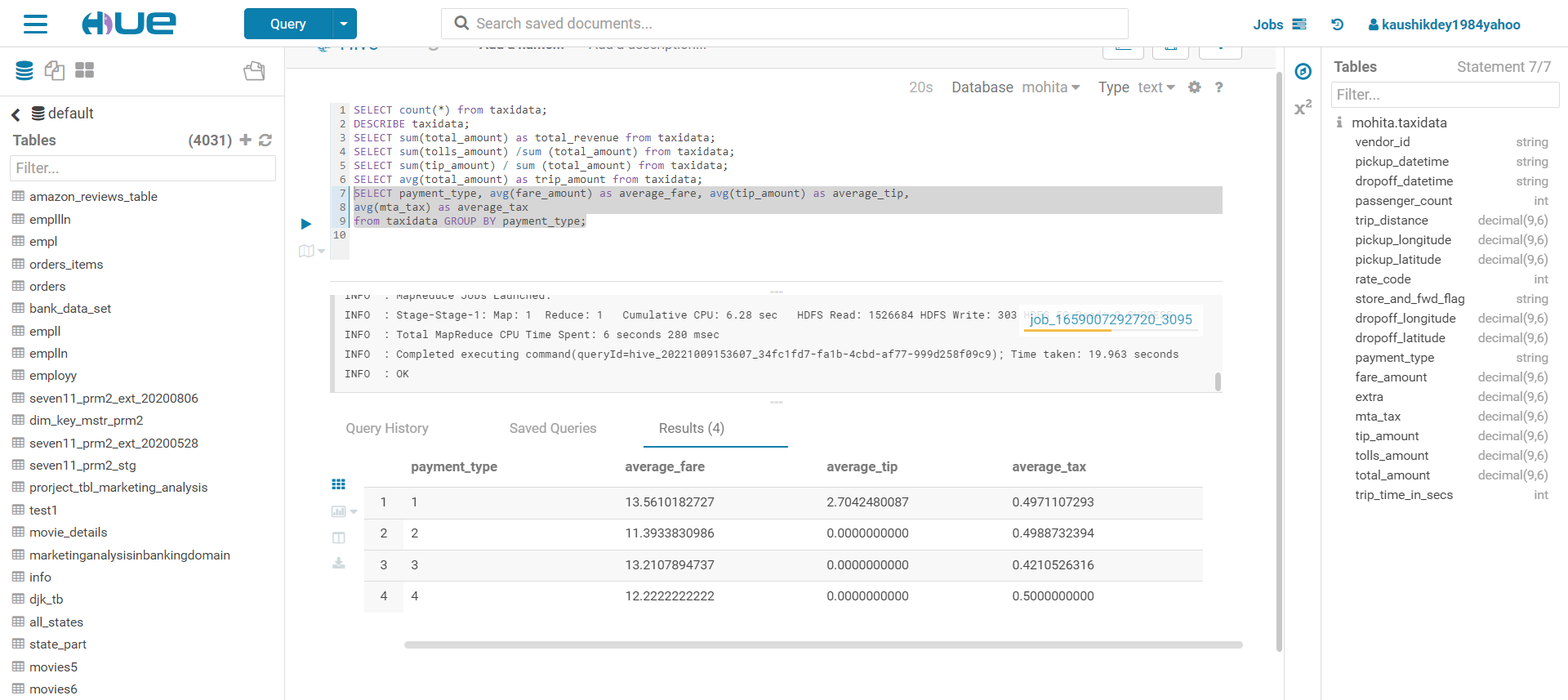
output :

Top of Form

Bottom of Form

|  | **payment\_type** | **average\_fare** | **average\_tip** | **average\_tax** |
| --- | --- | --- | --- | --- |

|  | |
| --- | --- |
|  | **payment\_type** | | **average\_fare** | **average\_tip** | **average\_tax** |
| 1 | 1 | | 13.5610182727 | 2.7042480087 | 0.4971107293 |
| 2 | 2 | | 11.3933830986 | 0.0000000000 | 0.4988732394 |
| 3 | 3 | | 13.2107894737 | 0.0000000000 | 0.4210526316 |
| 4 | 4 | | 12.2222222222 | 0.0000000000 | 0.5000000000 |



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

Ans : 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;

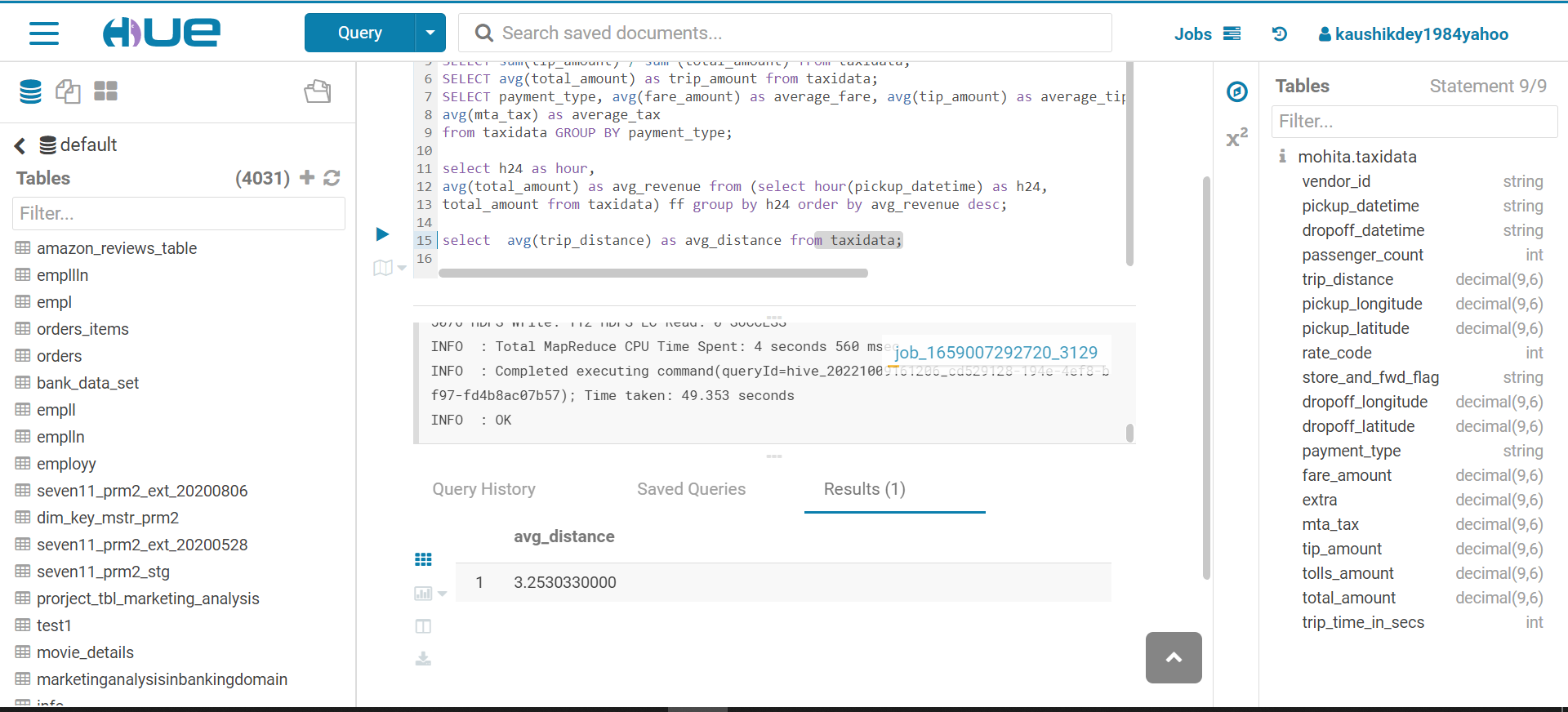
Graphical user interface, text, application, email

Description automatically generated

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

Ans : SELECT avg(trip\_distance) as avg\_distance from taxidata;

Output: 3.2530330000



1. How many different payment types are used? Column name – payment\_type. select distinct payment\_type from taxidata

Ans : SELECT distinct payment\_type from taxidata;

Output : Graphical user interface, text, application, email

Description automatically generated