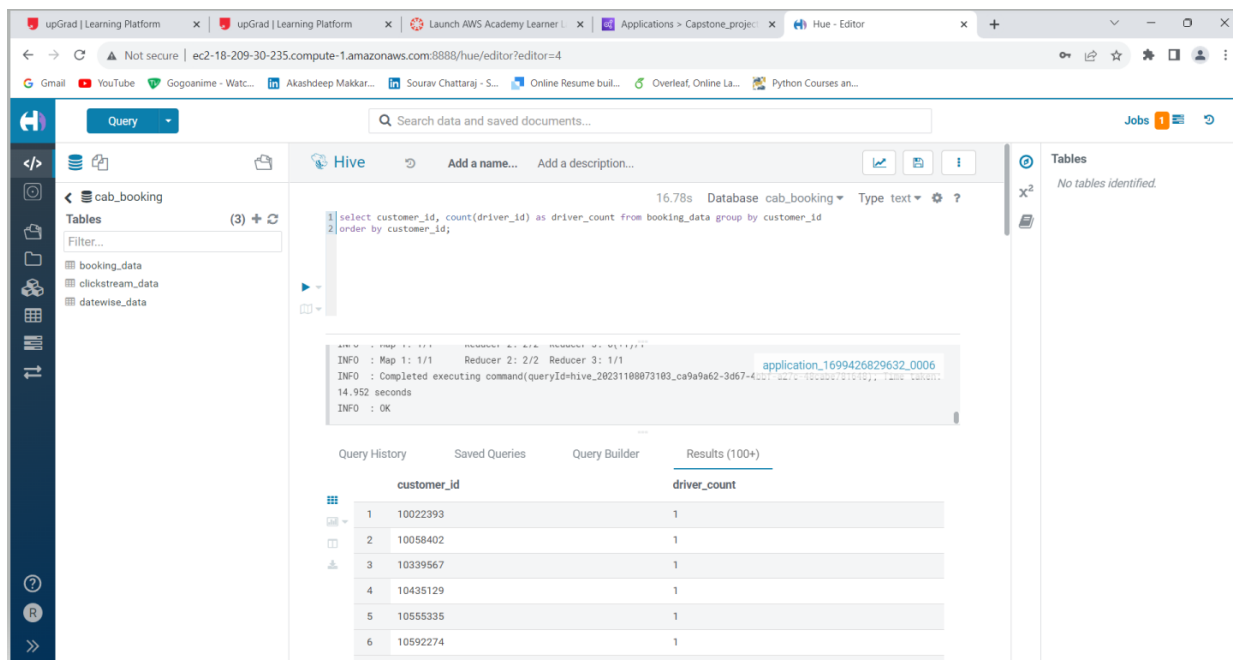


Queries

1. Hive Query for Task 5:- Calculate the total different drivers for each customer.

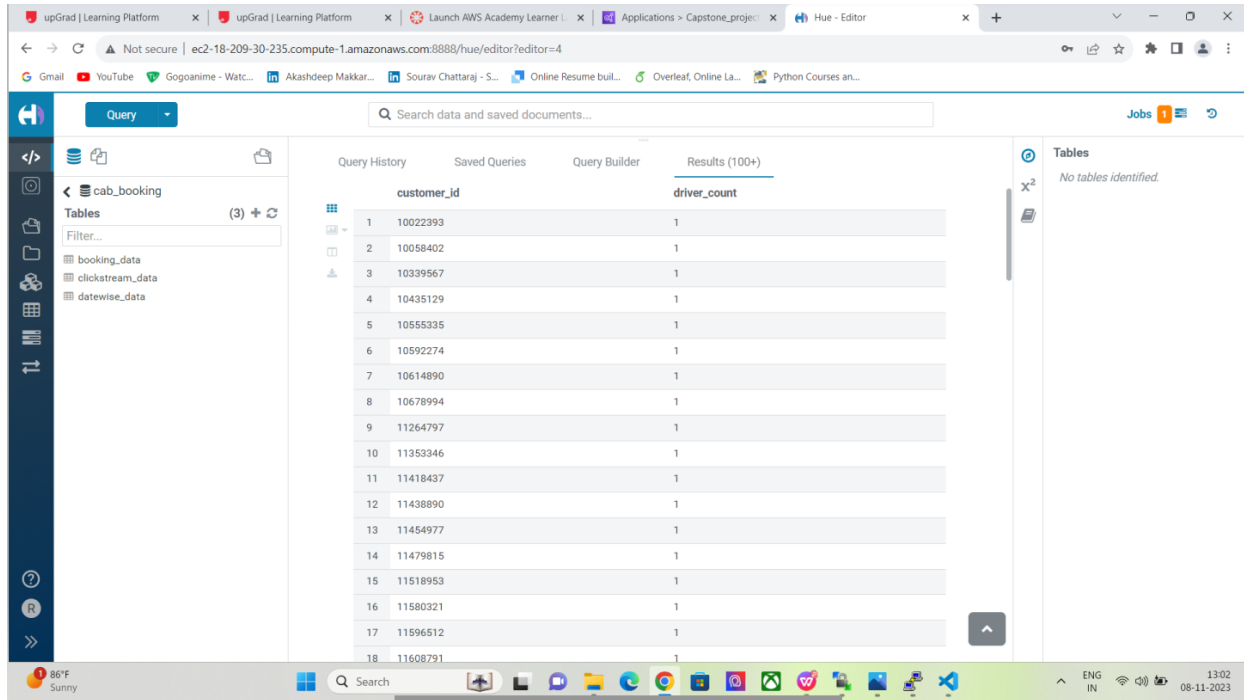
```
select customer_id, count(driver_id) as driver_count from booking_data group by customer_id
order by customer_id;
```

Screenshot after executing Query:-



The screenshot shows the Hue Editor interface with a Hive query executed. The query is: `select customer_id, count(driver_id) as driver_count from booking_data group by customer_id order by customer_id;` The results are displayed in a table with 6 rows.

customer_id	driver_count
10022393	1
10058402	1
10339567	1
10435129	1
10555335	1
10592274	1



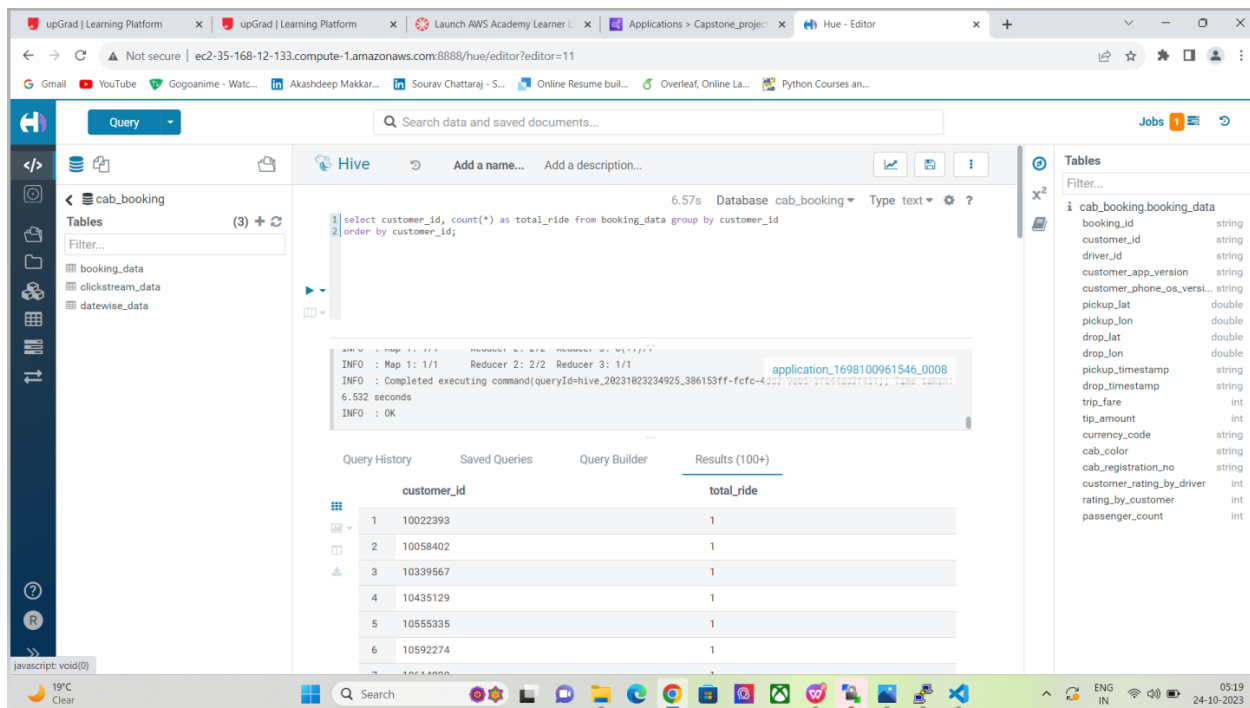
The screenshot shows the Hue Editor interface with a query result table. The table has two columns: **customer_id** and **driver_count**. The results are as follows:

customer_id	driver_count
10022393	1
10058402	1
10339567	1
10435129	1
10555335	1
10592274	1
10614890	1
10678994	1
11264797	1
11353346	1
11418437	1
11438890	1
11454977	1
11479815	1
11518953	1
11580321	1
11596512	1
11608791	1

2. Hive Query for Task 6:- Calculate the total rides taken by each customer.

```
select customer_id, count(*) as total_ride from booking_data group by customer_id
order by customer_id;
```

Screenshot after executing Query:-



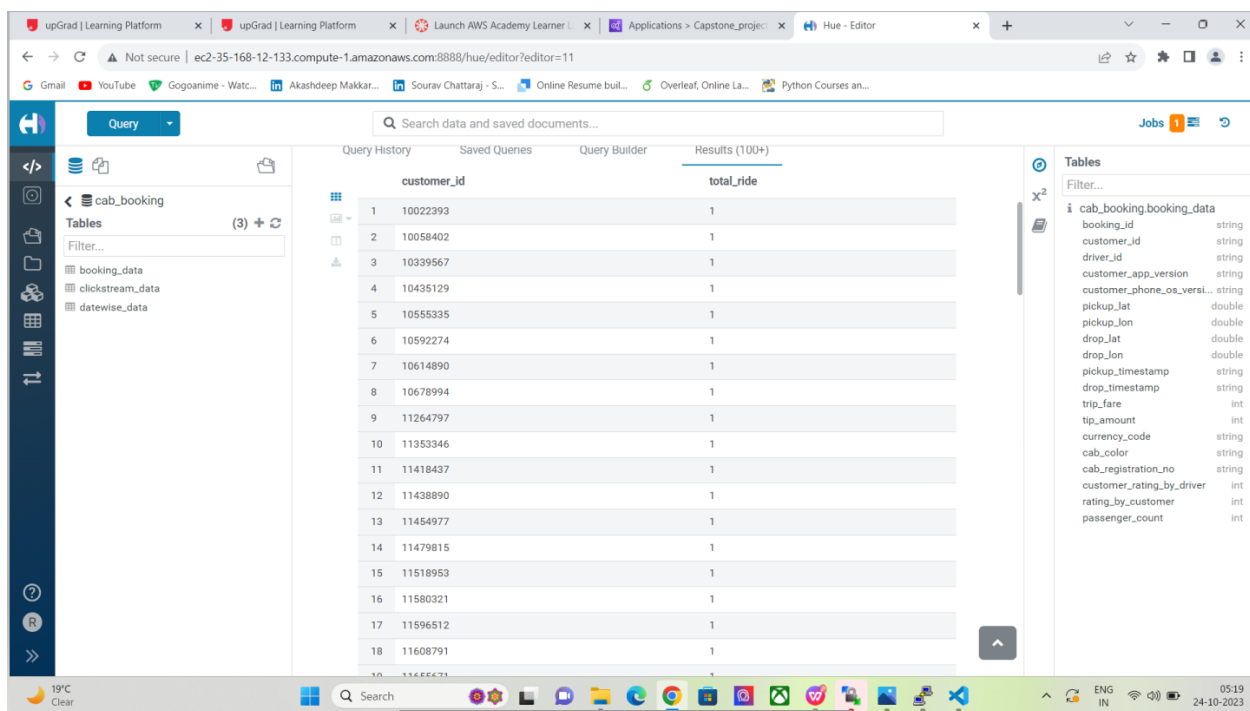
The screenshot shows the Hue Editor interface with a Hive query executed. The query is:

```
select customer_id, count(*) as total_ride from booking_data group by customer_id
order by customer_id;
```

The results are displayed in a table with 6 rows:

customer_id	total_ride
10022393	1
10058402	1
10339567	1
10435129	1
10555335	1
10592274	1

The interface also shows a sidebar with tables (booking_data, clickstream_data, datewise_data) and a right panel with a table schema for cab_booking.booking_data.



The screenshot shows the Hue Editor interface with a Hive query executed. The query is:

```
select customer_id, count(*) as total_ride from booking_data group by customer_id
order by customer_id;
```

The results are displayed in a table with 18 rows:

customer_id	total_ride
10022393	1
10058402	1
10339567	1
10435129	1
10555335	1
10592274	1
10614890	1
10678994	1
11264797	1
11353346	1
11418437	1
11438890	1
11454977	1
11479815	1
11518953	1
11580321	1
11596512	1
11608791	1

The interface also shows a sidebar with tables (booking_data, clickstream_data, datewise_data) and a right panel with a table schema for cab_booking.booking_data.

3. Hive Query for Task 7:-

Find total visits by each customer on the booking page and total 'Book Now' button press. This can show the conversion-ratio.

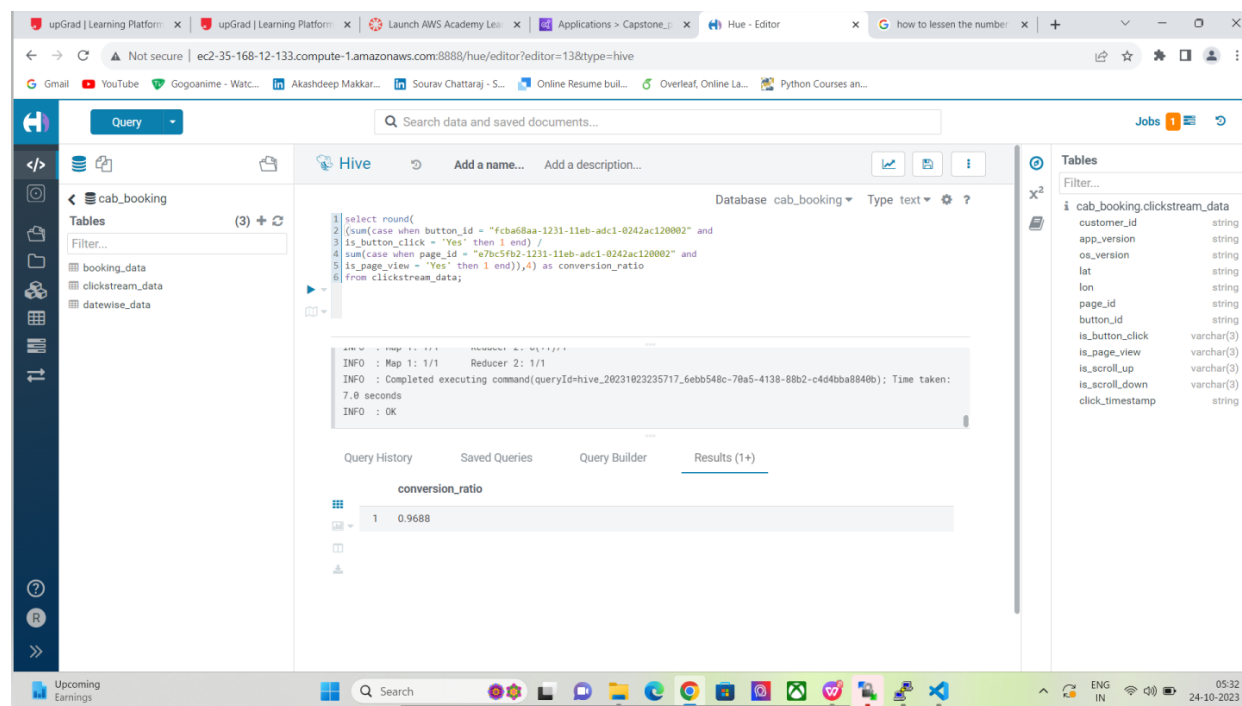
Booking page id is 'e7bc5fb2-1231-11eb-adc1-0242ac120002'

Book Now button id is 'fcba68aa-1231-11eb-adc1-0242ac120002'.

Also, calculate the conversion ratio as described in the Tasks segment.

```
select round(
(sum(case when button_id = "fcba68aa-1231-11eb-adc1-0242ac120002" and
is_button_click = 'Yes' then 1 end) /
sum(case when page_id = "e7bc5fb2-1231-11eb-adc1-0242ac120002" and
is_page_view = 'Yes' then 1 end)),4) as conversion_ratio
from clickstream_data;
```

Screenshot after executing Query:-



The screenshot shows the Hue interface with the following details:

- Query:**

```
select round(
(sum(case when button_id = "fcba68aa-1231-11eb-adc1-0242ac120002" and
is_button_click = 'Yes' then 1 end) /
sum(case when page_id = "e7bc5fb2-1231-11eb-adc1-0242ac120002" and
is_page_view = 'Yes' then 1 end)),4) as conversion_ratio
from clickstream_data;
```
- Execution Log:**

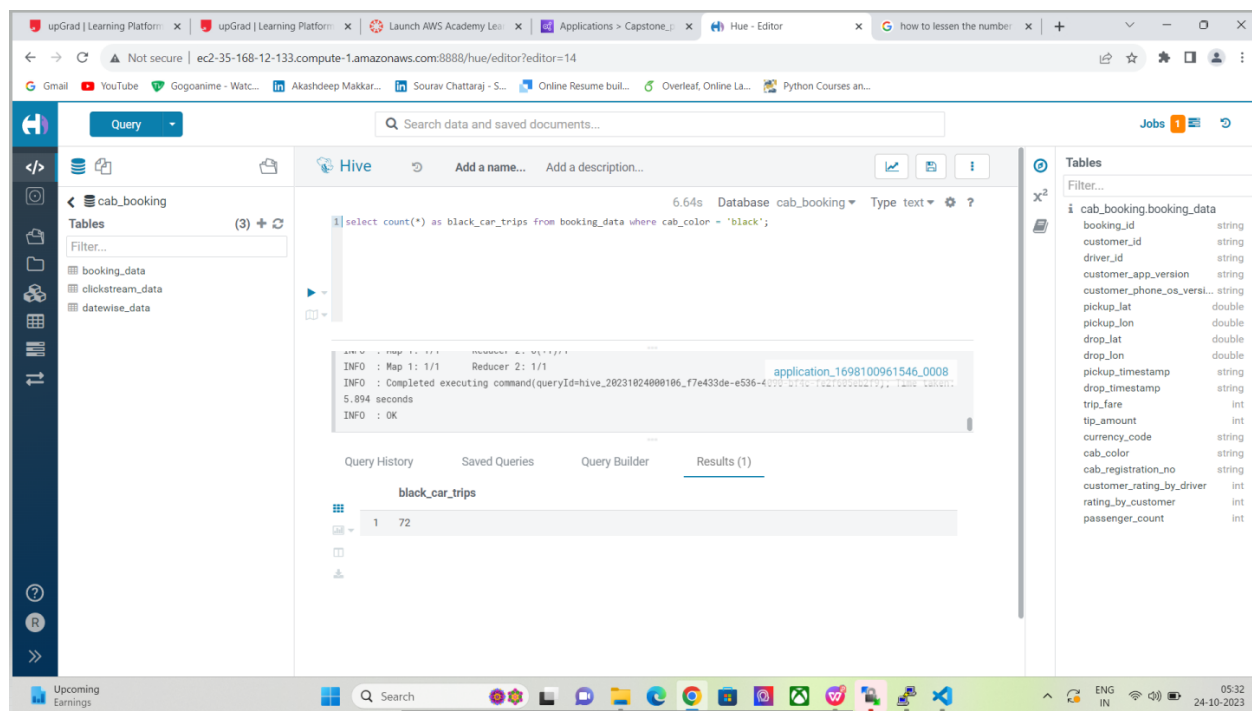
```
INFO : Map 1: 1/1 Reducer 2: 1/1
INFO : Completed executing command(queryId=hive_20231923235717_6ebb548c-78a5-4138-88b2-c4d4bba8840b); Time taken: 7.0 seconds
INFO : OK
```
- Results (1+):**

conversion_ratio
1 0.9688
- Tables:**
 - cab_booking.clickstream_data
 - customer_id string
 - app_version string
 - os_version string
 - lat string
 - lon string
 - page_id string
 - button_id string
 - is_button_click varchar(3)
 - is_page_view varchar(3)
 - is_scroll_up varchar(3)
 - is_scroll_down varchar(3)
 - click_timestamp string

4. Hive Query for Task 8:- Calculate the count of all trips done on Black cabs.

```
select count(*) as black_car_trips from booking_data where cab_color = 'black';
```

Screenshot after executing Query:-



The screenshot shows the Hue interface with a Hive query executed. The query is:

```
select count(*) as black_car_trips from booking_data where cab_color = 'black';
```

The execution log shows the query completed successfully in 5.894 seconds. The result is displayed in a table named 'black_car_trips' with one row containing the value 72.

5. Hive Query for Task 9:- Calculate the total tip amount on a given date to all drivers by customers.

```
select date_format(pickup_timestamp,'YYYY-MM-dd') as datewise, sum(tip_amount) as total_tip
from booking_data
group by date_format(pickup_timestamp,'YYYY-MM-dd')
order by datewise;
```

Screenshot after executing Query:-

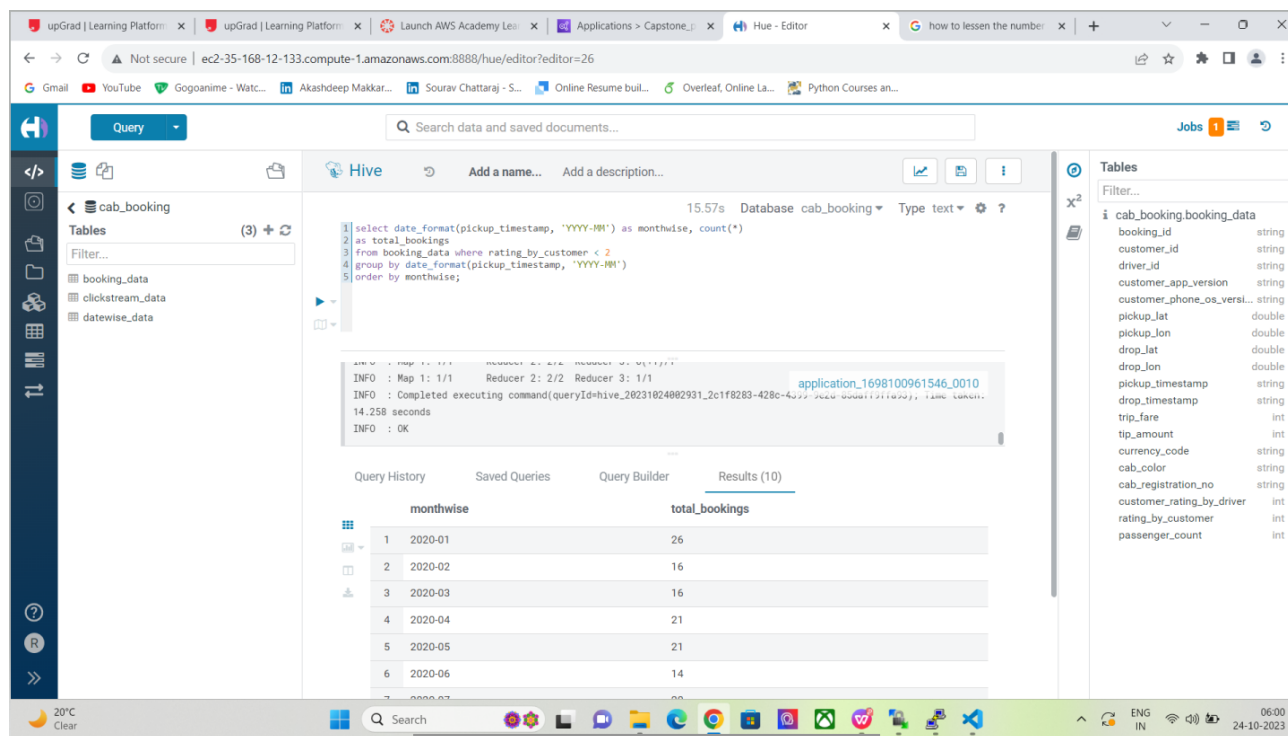
The screenshot shows the Databricks workspace interface. The top navigation bar includes the 'Query' tab. The main area displays a table of query results for the 'cab_booking' table, with columns 'datewise' and 'total_tip'. The table contains 17 rows of data. On the right, the 'Tables' sidebar shows the 'cab_booking' table structure with columns like 'booking_id', 'customer_id', 'driver_id', etc.

Query History	Saved Queries	Query Builder	Results (289)																																				
			<table border="1"> <thead> <tr> <th>datewise</th> <th>total_tip</th> </tr> </thead> <tbody> <tr><td>1 2020-01-01</td><td>59</td></tr> <tr><td>2 2020-01-02</td><td>95</td></tr> <tr><td>3 2020-01-03</td><td>11</td></tr> <tr><td>4 2020-01-04</td><td>123</td></tr> <tr><td>5 2020-01-05</td><td>134</td></tr> <tr><td>6 2020-01-06</td><td>189</td></tr> <tr><td>7 2020-01-07</td><td>148</td></tr> <tr><td>8 2020-01-08</td><td>111</td></tr> <tr><td>9 2020-01-09</td><td>48</td></tr> <tr><td>10 2020-01-10</td><td>77</td></tr> <tr><td>11 2020-01-11</td><td>81</td></tr> <tr><td>12 2020-01-12</td><td>109</td></tr> <tr><td>13 2020-01-14</td><td>142</td></tr> <tr><td>14 2020-01-15</td><td>338</td></tr> <tr><td>15 2020-01-16</td><td>155</td></tr> <tr><td>16 2020-01-17</td><td>296</td></tr> <tr><td>17 2020-01-18</td><td>240</td></tr> </tbody> </table>	datewise	total_tip	1 2020-01-01	59	2 2020-01-02	95	3 2020-01-03	11	4 2020-01-04	123	5 2020-01-05	134	6 2020-01-06	189	7 2020-01-07	148	8 2020-01-08	111	9 2020-01-09	48	10 2020-01-10	77	11 2020-01-11	81	12 2020-01-12	109	13 2020-01-14	142	14 2020-01-15	338	15 2020-01-16	155	16 2020-01-17	296	17 2020-01-18	240
datewise	total_tip																																						
1 2020-01-01	59																																						
2 2020-01-02	95																																						
3 2020-01-03	11																																						
4 2020-01-04	123																																						
5 2020-01-05	134																																						
6 2020-01-06	189																																						
7 2020-01-07	148																																						
8 2020-01-08	111																																						
9 2020-01-09	48																																						
10 2020-01-10	77																																						
11 2020-01-11	81																																						
12 2020-01-12	109																																						
13 2020-01-14	142																																						
14 2020-01-15	338																																						
15 2020-01-16	155																																						
16 2020-01-17	296																																						
17 2020-01-18	240																																						

6. Hive Query for Task 10:- Calculate the count of all the bookings with a rating below 2 in a particular month.

```
select date_format(pickup_timestamp, 'YYYY-MM') as monthwise, count(*)
as total_bookings
from booking_data where rating_by_customer < 2
group by date_format(pickup_timestamp, 'YYYY-MM')
order by monthwise;
```

Screenshot after executing Query:-



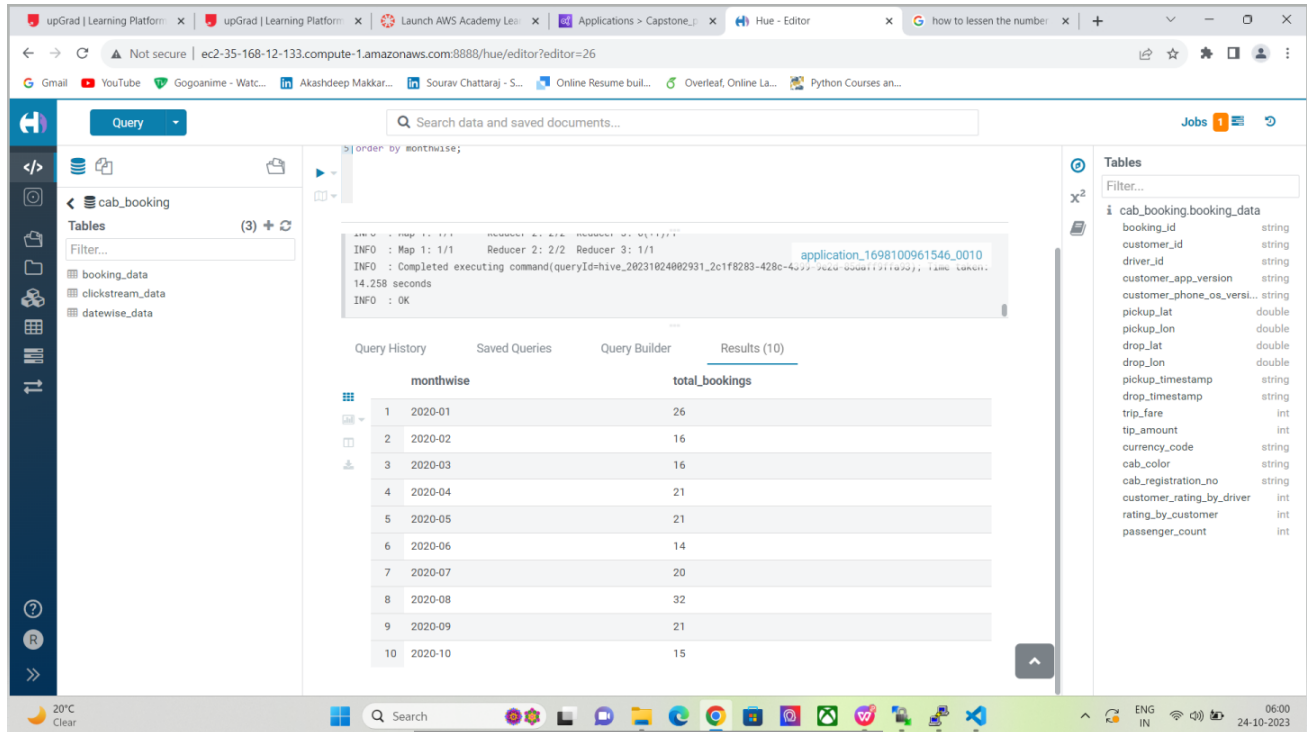
The screenshot shows the Hue interface with the following components:

- Query Editor:** Contains the Hive query:


```
select date_format(pickup_timestamp, 'YYYY-MM') as monthwise, count(*)
as total_bookings
from booking_data where rating_by_customer < 2
group by date_format(pickup_timestamp, 'YYYY-MM')
order by monthwise;
```
- Query Execution Log:** Shows the command execution details:


```
INFO : Map 1: 1/1 Reducer 2: 2/2 Reducer 3: 1/1
INFO : Completed executing command(queryId=hive_20231024062931_2c1f8283-428c-4599-9600-000111111111, time taken: 14.258 seconds)
INFO : OK
```
- Results Table:** Displays the query results with 10 rows (shown as 6 in the image):

monthwise	total_bookings
1 2020-01	26
2 2020-02	16
3 2020-03	16
4 2020-04	21
5 2020-05	21
6 2020-06	14
- Tables List:** Shows the schema for the 'cab_booking.booking_data' table, including columns like booking_id, customer_id, driver_id, customer_app_version, customer_phone_os_version, pickup_latitude, pickup_longitude, drop_latitude, drop_longitude, pickup_timestamp, drop_timestamp, trip_fare, tip_amount, currency_code, cab_color, cab_registration_no, customer_rating_by_driver, rating_by_customer, and passenger_count.



Query: `select count(*) AS TOTAL_USERS from clickstream_data where os_version = 'iOS';`

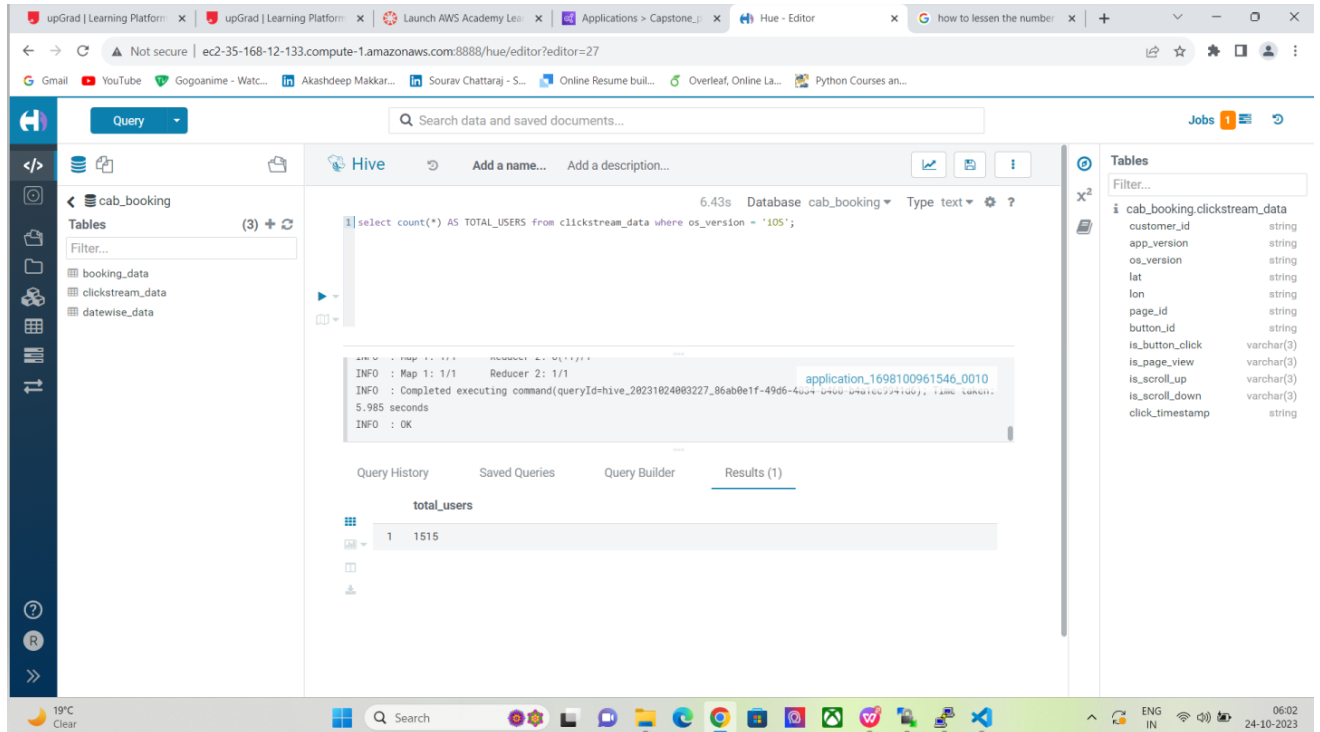
Results (10):

monthwise	total_bookings
1 2020-01	26
2 2020-02	16
3 2020-03	16
4 2020-04	21
5 2020-05	21
6 2020-06	14
7 2020-07	20
8 2020-08	32
9 2020-09	21
10 2020-10	15

7. Hive Query for Task 11:- Calculate the count of total iOS users.

`select count(*) AS TOTAL_USERS from clickstream_data where os_version = 'iOS';`

Screenshot after executing Query:-



The screenshot displays the Hue web interface for managing Hive data. The browser tabs at the top include 'upGrad | Learning Platform', 'Launch AWS Academy Les...', 'Applications > Capstone...', 'Hue - Editor', and 'how to lessen the number...'. The address bar shows the URL 'ec2-35-168-12-133.compute-1.amazonaws.com:8888/hue/editor?editor=27'.

On the left sidebar, under the 'cab_booking' database, the following tables are listed: booking_data, clickstream_data, and datewise_data. The main editor area shows a Hive query: `select count(*) AS TOTAL_USERS from clickstream_data where os_version = 'IOS';`. The query execution log indicates it took 5.985 seconds to complete. The results pane shows a single row with the value 1515 for the column 'total_users'.

On the right, the 'Tables' pane shows the schema for 'cab_booking.clickstream_data', listing columns such as customer_id, app_version, os_version, lat, lon, page_id, button_id, is_button_click, is_page_view, is_scroll_up, is_scroll_down, and click_timestamp.