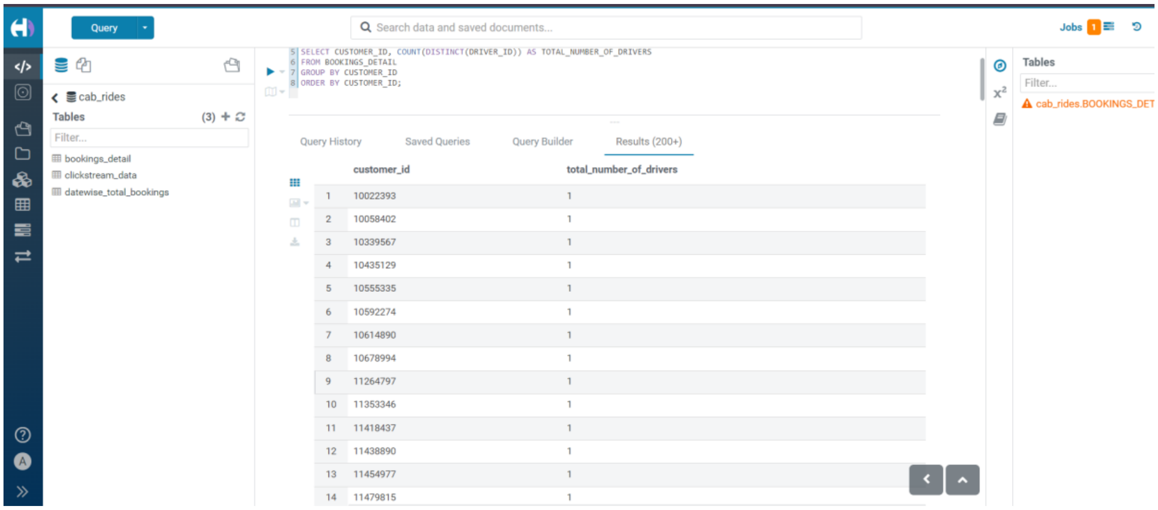
Queries

Task 5: Calculate the total number of different drivers for each customer.

Query:

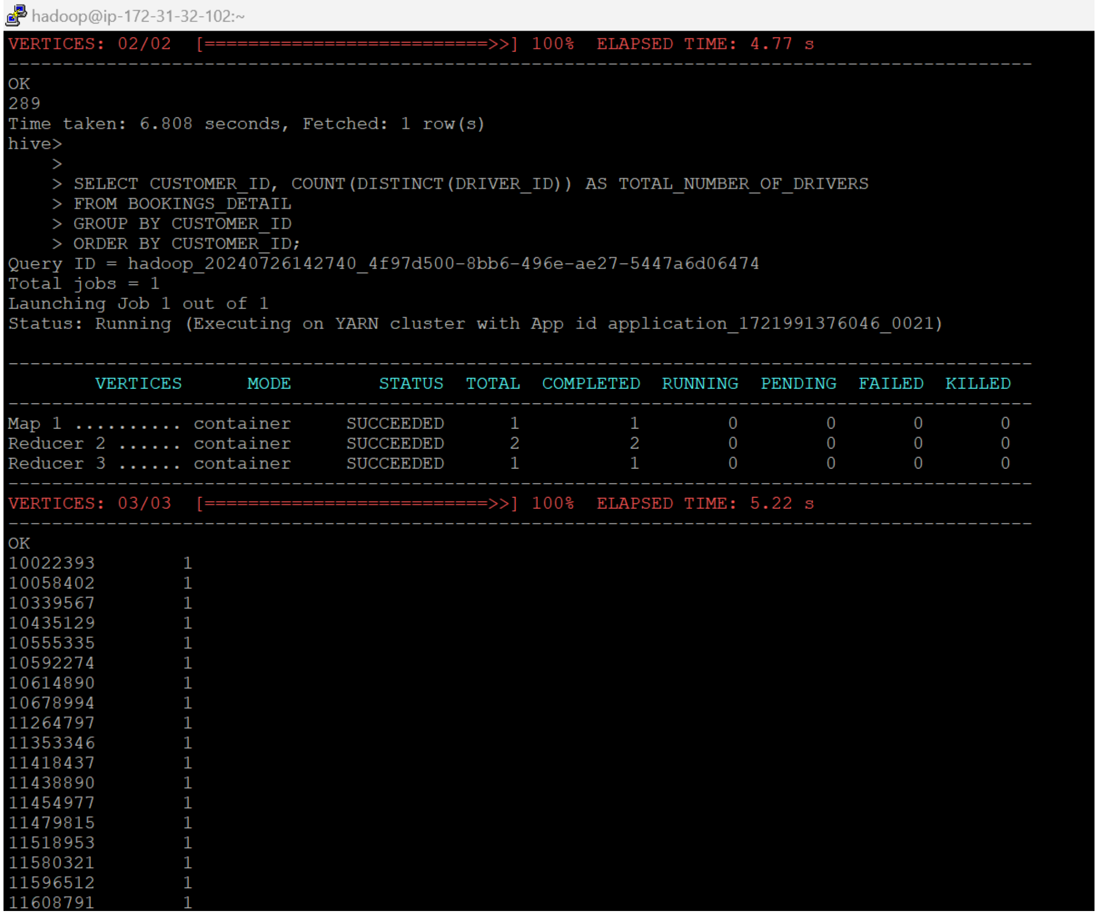
SELECT CUSTOMER_ID, COUNT(DISTINCT(DRIVER_ID)) AS TOTAL_NUMBER_OF_DRIVERS
FROM BOOKINGS_DETAIL
GROUP BY CUSTOMER_ID
ORDER BY CUSTOMER_ID;

Output:



```
SELECT CUSTOMER_ID, COUNT(DISTINCT(DRIVER_ID)) AS TOTAL_NUMBER_OF_DRIVERS
FROM BOOKINGS_DETAIL
GROUP BY CUSTOMER_ID
ORDER BY CUSTOMER_ID;
```

	customer_id	total_number_of_drivers
1	10022393	1
2	10058402	1
3	10339567	1
4	10435129	1
5	10555335	1
6	10592274	1
7	10614890	1
8	10678994	1
9	11264797	1
10	11353346	1
11	11418437	1
12	11438890	1
13	11454977	1
14	11479815	1



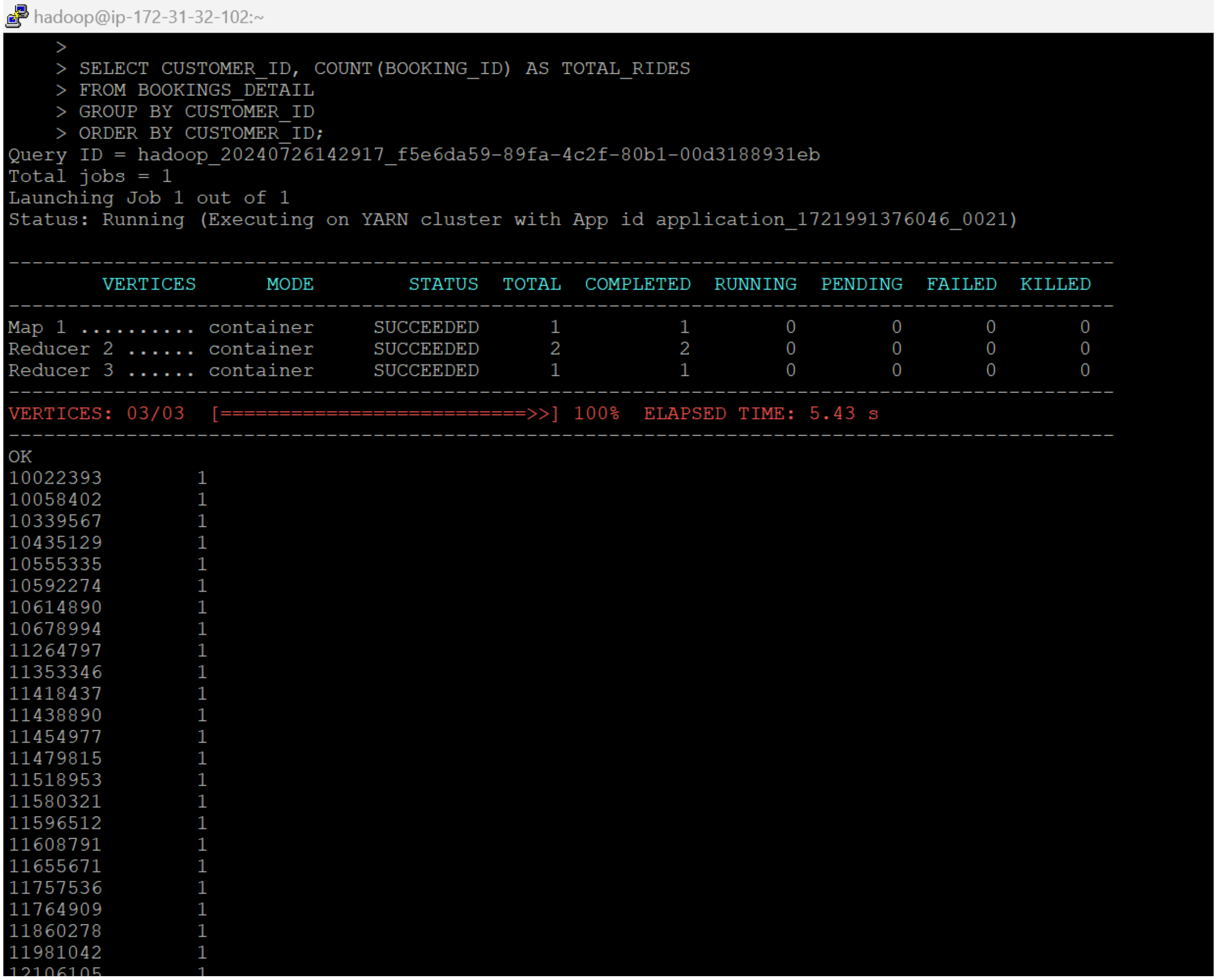
```
hadoop@ip-172-31-32-102:~
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.77 s
----------------------------------------------------------------------------------------------
OK
289
Time taken: 6.808 seconds, Fetched: 1 row(s)
hive>
    >
    > SELECT CUSTOMER_ID, COUNT(DISTINCT(DRIVER_ID)) AS TOTAL_NUMBER_OF_DRIVERS
    > FROM BOOKINGS_DETAIL
    > GROUP BY CUSTOMER_ID
    > ORDER BY CUSTOMER_ID;
Query ID = hadoop_20240726142740_4f97d500-8bb6-496e-ae27-5447a6d06474
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.22 s
----------------------------------------------------------------------------------------------
OK
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
```

Task 6: Calculate the total rides taken by each customer.

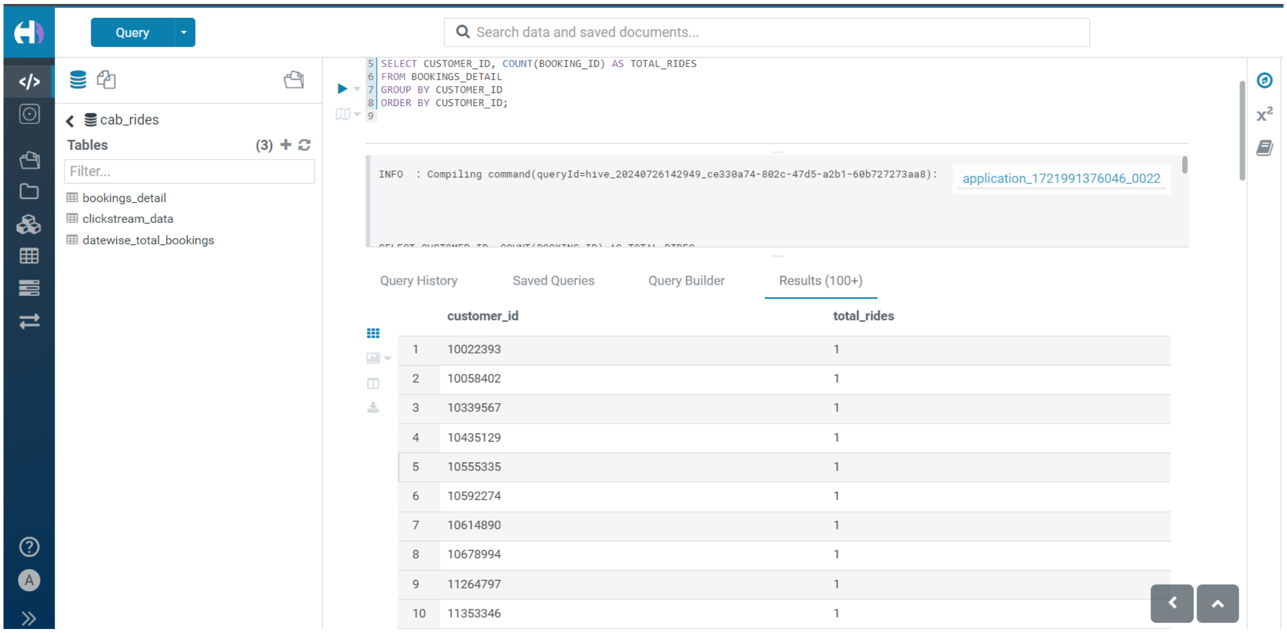
Query:-

SELECT CUSTOMER_ID, COUNT(BOOKING_ID) AS TOTAL_RIDES
FROM BOOKINGS_DETAIL
GROUP BY CUSTOMER_ID
ORDER BY CUSTOMER_ID;

Output:



```
    >
    > SELECT CUSTOMER_ID, COUNT(BOOKING_ID) AS TOTAL_RIDES
    > FROM BOOKINGS_DETAIL
    > GROUP BY CUSTOMER_ID
    > ORDER BY CUSTOMER_ID;
Query ID = hadoop_20240726142917_f5e6da59-89fa-4c2f-80b1-00d3188931eb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.43 s
----------------------------------------------------------------------------------------------
OK
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
11655671        1
11757536        1
11764909        1
11860278        1
11981042        1
12106105        1
```



```
SELECT CUSTOMER_ID, COUNT(BOOKING_ID) AS TOTAL_RIDES
FROM BOOKINGS_DETAIL
GROUP BY CUSTOMER_ID
ORDER BY CUSTOMER_ID;
```

```
INFO  : Compiling command(queryId=hive_20240726142949_ce330a74-802c-47d5-a2b1-60b727273aa8):
```

application_1721991376046_0022

	customer_id	total_rides
1	10022393	1
2	10058402	1
3	10339567	1
4	10435129	1
5	10555335	1
6	10592274	1
7	10614890	1
8	10678994	1
9	11264797	1
10	11353346	1

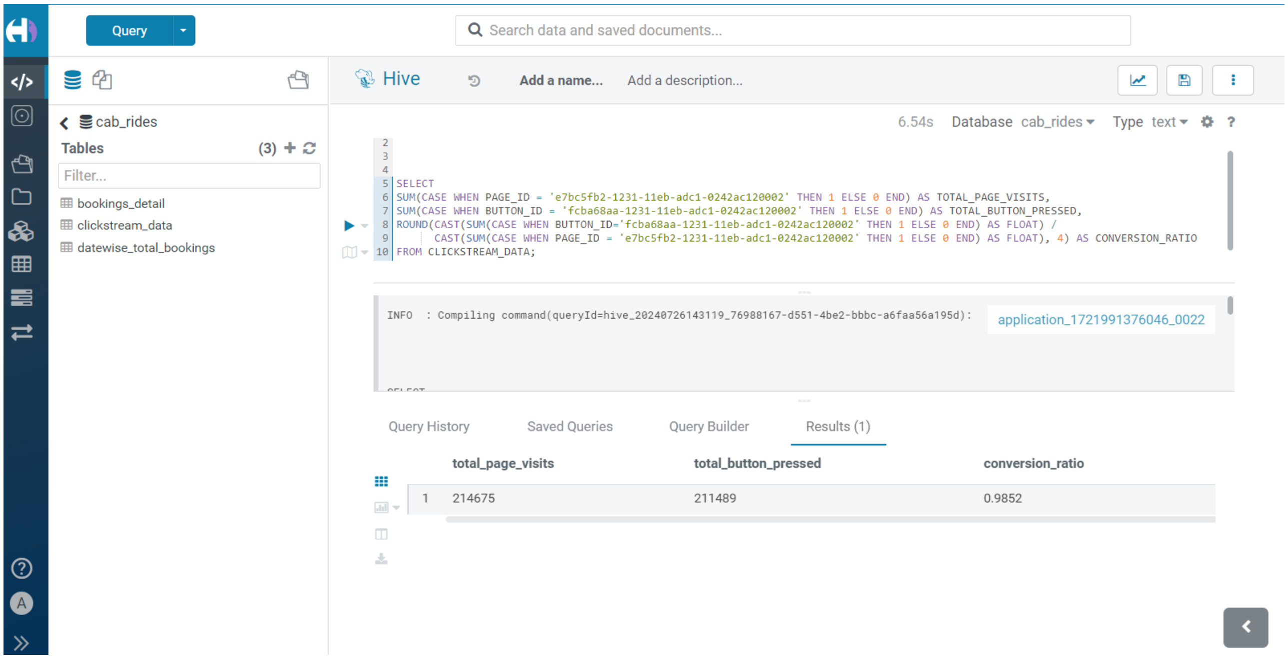
Task 7: Find the total visits made by each customer on the booking page and the total ‘Book Now’ button presses. This can show the conversion ratio.

The booking page id is ‘e7bc5fb2-1231-11eb-adc1-0242ac120002’. The Book Now button id is ‘fcba68aa-1231-11eb-adc1-0242ac120002’. You also need to calculate the conversion ratio as part of this task. Conversion ratio can be calculated as **Total 'Book Now' Button Press/Total Visits made by customer on the booking page**.

Query:-

```
SELECT
SUM(CASE WHEN PAGE_ID = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0
END) AS TOTAL_PAGE_VISITS,
SUM(CASE WHEN BUTTON_ID = 'fcba68aa-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0
END) AS TOTAL_BUTTON_PRESSED,
ROUND(CAST(SUM(CASE WHEN BUTTON_ID='fcba68aa-1231-11eb-adc1-0242ac120002'
THEN 1 ELSE 0 END) AS FLOAT) /
    CAST(SUM(CASE WHEN PAGE_ID = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' THEN 1
ELSE 0 END) AS FLOAT), 4) AS CONVERSION_RATIO
FROM CLICKSTREAM_DATA;
```

Output:-

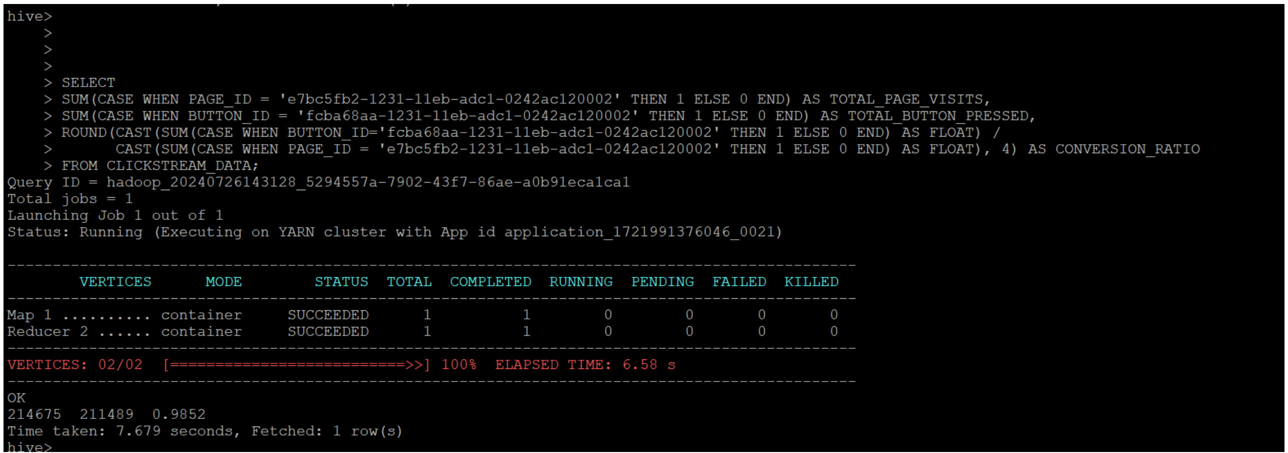


```
SELECT
SUM(CASE WHEN PAGE_ID = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS TOTAL_PAGE_VISITS,
SUM(CASE WHEN BUTTON_ID = 'fcba68aa-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS TOTAL_BUTTON_PRESSED,
ROUND(CAST(SUM(CASE WHEN BUTTON_ID='fcba68aa-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS FLOAT) /
      CAST(SUM(CASE WHEN PAGE_ID = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS FLOAT), 4) AS CONVERSION_RATIO
FROM CLICKSTREAM_DATA;
```

```
INFO  : Compiling command(queryId=hive_20240726143119_76988167-d551-4be2-bbbc-a6faa56a195d):
```

application_1721991376046_0022

	total_page_visits	total_button_pressed	conversion_ratio
1	214675	211489	0.9852



```
hive>
    >
    >
    >
    > SELECT
    > SUM(CASE WHEN PAGE_ID = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS TOTAL_PAGE_VISITS,
    > SUM(CASE WHEN BUTTON_ID = 'fcba68aa-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS TOTAL_BUTTON_PRESSED,
    > ROUND(CAST(SUM(CASE WHEN BUTTON_ID='fcba68aa-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS FLOAT) /
    >       CAST(SUM(CASE WHEN PAGE_ID = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0 END) AS FLOAT), 4) AS CONVERSION_RATIO
    > FROM CLICKSTREAM_DATA;
Query ID = hadoop_20240726143128_5294557a-7902-43f7-86ae-a0b91eca1ca1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.58 s
----------------------------------------------------------------------------------------------
OK
214675  211489  0.9852
Time taken: 7.679 seconds, Fetched: 1 row(s)
hive>
```

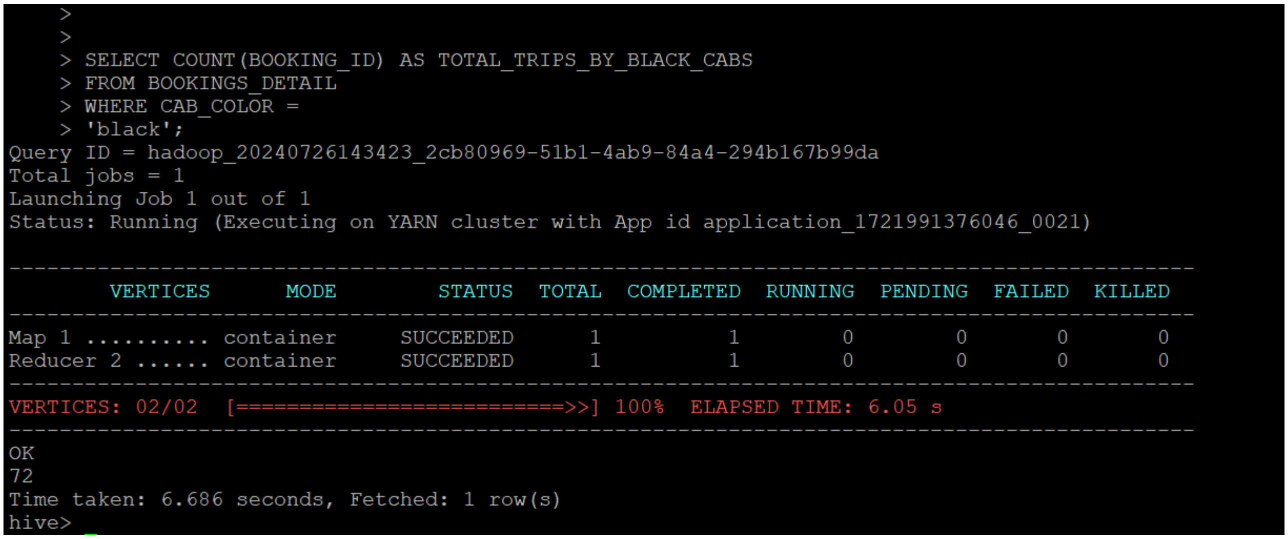
** There is slight difference (0.9852~0.9688=0.0164) ~=1.6% in the conversion ratio as per the validation document due to the additional 16 records present in the loaded clickstream data in HDFS.

Task 8: Calculate the count of all trips done on black cabs.

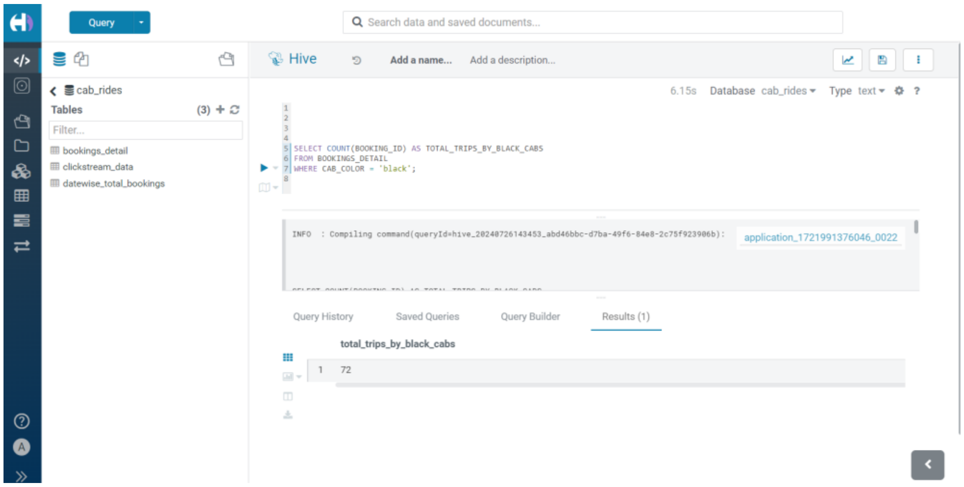
Query:-

```
 SELECT COUNT(BOOKING_ID) AS TOTAL_TRIPS_BY_BLACK_CABS
 FROM BOOKINGS_DETAIL
 WHERE CAB_COLOR = 'black';
```

Validation:



```
    >
    >
    > SELECT COUNT(BOOKING_ID) AS TOTAL_TRIPS_BY_BLACK_CABS
    > FROM BOOKINGS_DETAIL
    > WHERE CAB_COLOR =
    > 'black';
Query ID = hadoop_20240726143423_2cb80969-51b1-4ab9-84a4-294b167b99da
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.05 s
----------------------------------------------------------------------------------------------
OK
72
Time taken: 6.686 seconds, Fetched: 1 row(s)
hive>
```



```
SELECT COUNT(BOOKING_ID) AS TOTAL_TRIPS_BY_BLACK_CABS
FROM BOOKINGS_DETAIL
WHERE CAB_COLOR = 'black';
```

```
INFO  : Compiling command(queryId=hive_20240726143453_abd46bbc-d7ba-49f6-84e8-2c75f923906b):
```

application_1721991376046_0022

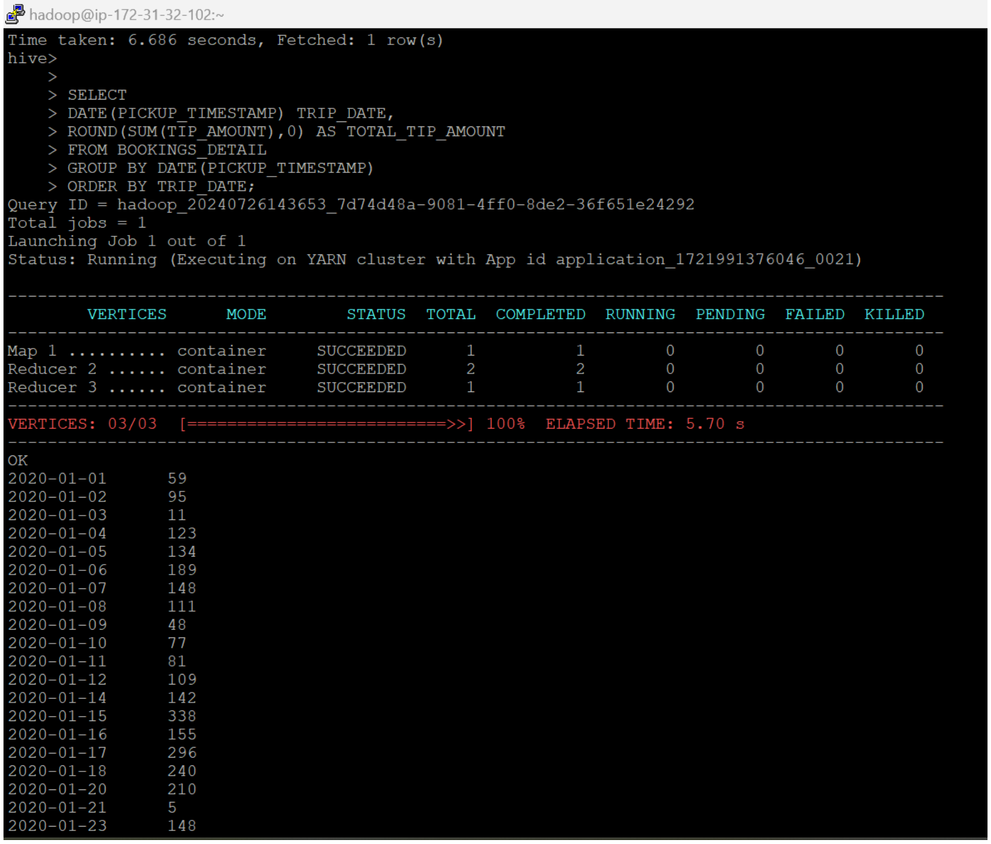
	total_trips_by_black_cabs
1	72

Task 9: Calculate the total amount of tips given date wise to all drivers by customers.

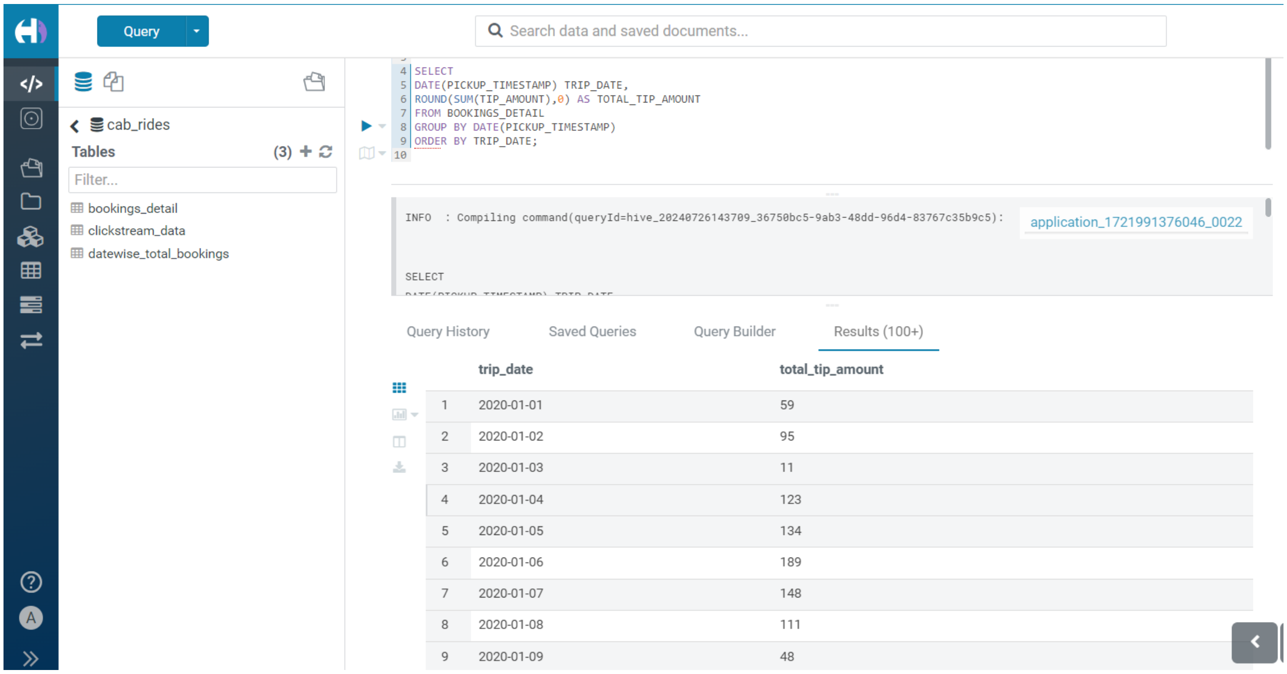
Query:-

SELECT
DATE(PICKUP_TIMESTAMP) TRIP_DATE,
ROUND(SUM(TIP_AMOUNT),0) AS TOTAL_TIP_AMOUNT
FROM BOOKINGS_DETAIL
GROUP BY DATE(PICKUP_TIMESTAMP)
ORDER BY TRIP_DATE;

Output:



```
Time taken: 6.686 seconds, Fetched: 1 row(s)
hive>
    >
    > SELECT
    > DATE(PICKUP_TIMESTAMP) TRIP_DATE,
    > ROUND(SUM(TIP_AMOUNT),0) AS TOTAL_TIP_AMOUNT
    > FROM BOOKINGS_DETAIL
    > GROUP BY DATE(PICKUP_TIMESTAMP)
    > ORDER BY TRIP_DATE;
Query ID = hadoop_20240726143653_7d74d48a-9081-4ff0-8de2-36f651e24292
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.70 s
----------------------------------------------------------------------------------------------
OK
2020-01-01      59
2020-01-02      95
2020-01-03      11
2020-01-04      123
2020-01-05      134
2020-01-06      189
2020-01-07      148
2020-01-08      111
2020-01-09      48
2020-01-10      77
2020-01-11      81
2020-01-12      109
2020-01-14      142
2020-01-15      338
2020-01-16      155
2020-01-17      296
2020-01-18      240
2020-01-20      210
2020-01-21      5
2020-01-23      148
```



```
SELECT
DATE(PICKUP_TIMESTAMP) TRIP_DATE,
ROUND(SUM(TIP_AMOUNT),0) AS TOTAL_TIP_AMOUNT
FROM BOOKINGS_DETAIL
GROUP BY DATE(PICKUP_TIMESTAMP)
ORDER BY TRIP_DATE;
```

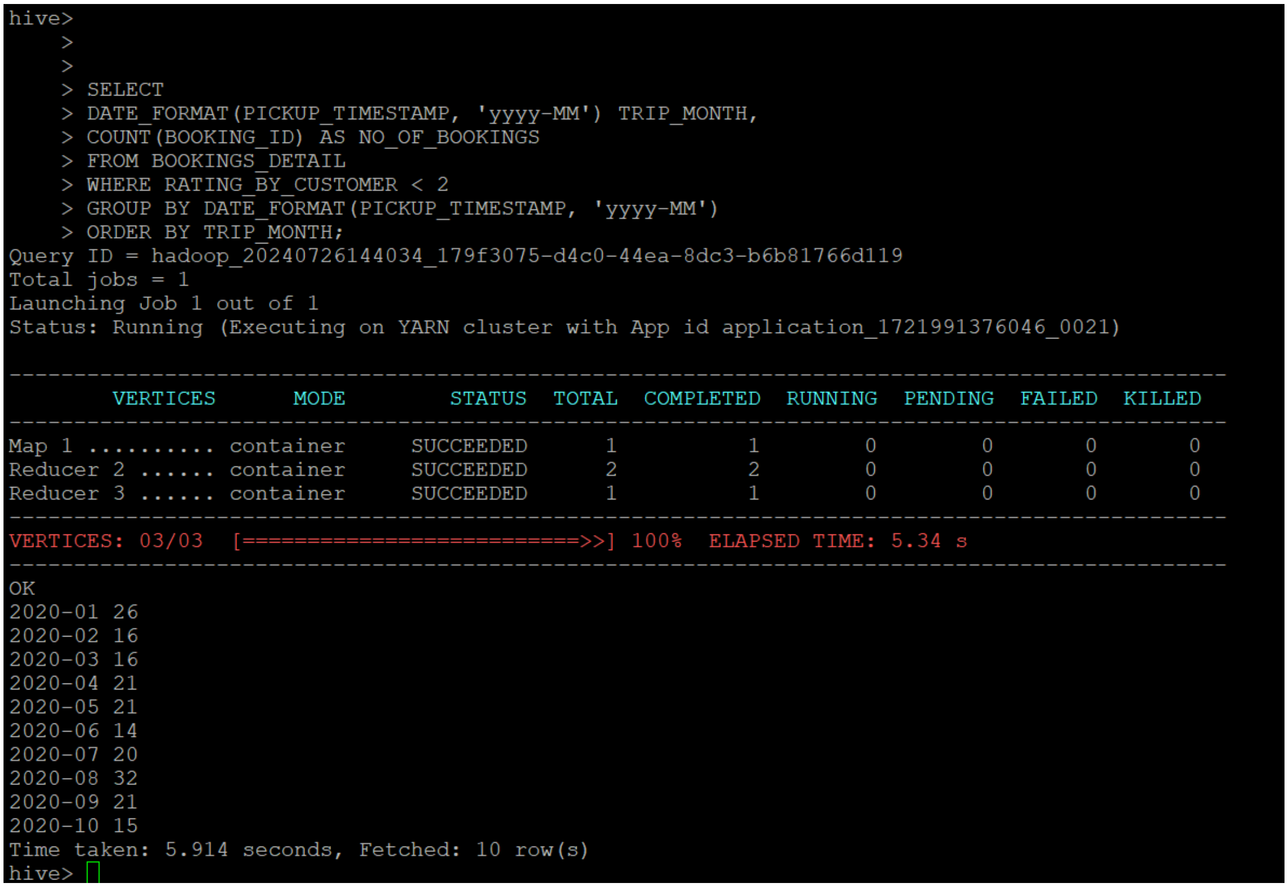
```
INFO  : Compiling command(queryId=hive_20240726143709_36750bc5-9ab3-48dd-96d4-83767c35b9c5):    application_1721991376046_0022


SELECT
```

	trip_date	total_tip_amount
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

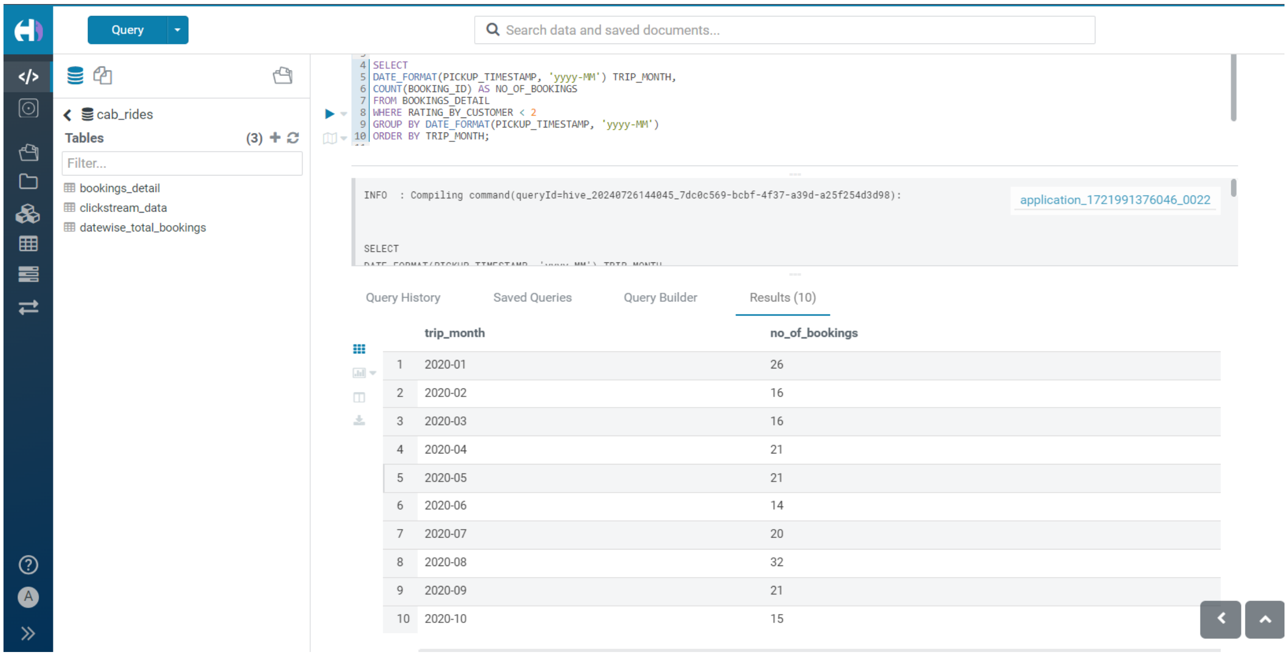
Task 10: Calculate the total count of all the bookings with ratings lower than 2 as given by customers in a particular month.

Query:-
SELECT
DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM') TRIP_MONTH,
COUNT(BOOKING_ID) AS NO_OF_BOOKINGS
FROM BOOKINGS_DETAIL
WHERE RATING_BY_CUSTOMER < 2
GROUP BY DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM')
ORDER BY TRIP_MONTH;

Output:



```
hive>
    >
    >
    > SELECT
    > DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM') TRIP_MONTH,
    > COUNT(BOOKING_ID) AS NO_OF_BOOKINGS
    > FROM BOOKINGS_DETAIL
    > WHERE RATING_BY_CUSTOMER < 2
    > GROUP BY DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM')
    > ORDER BY TRIP_MONTH;
Query ID = hadoop_20240726144034_179f3075-d4c0-44ea-8dc3-b6b81766d119
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.34 s
----------------------------------------------------------------------------------------------
OK
2020-01 26
2020-02 16
2020-03 16
2020-04 21
2020-05 21
2020-06 14
2020-07 20
2020-08 32
2020-09 21
2020-10 15
Time taken: 5.914 seconds, Fetched: 10 row(s)
hive>
```



```
SELECT
DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM') TRIP_MONTH,
COUNT(BOOKING_ID) AS NO_OF_BOOKINGS
FROM BOOKINGS_DETAIL
WHERE RATING_BY_CUSTOMER < 2
GROUP BY DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM')
ORDER BY TRIP_MONTH;
```

```
INFO  : Compiling command(queryId=hive_20240726144045_7dc0c569-bcbf-4f37-a39d-a25f254d3d98):

SELECT
```

application_1721991376046_0022

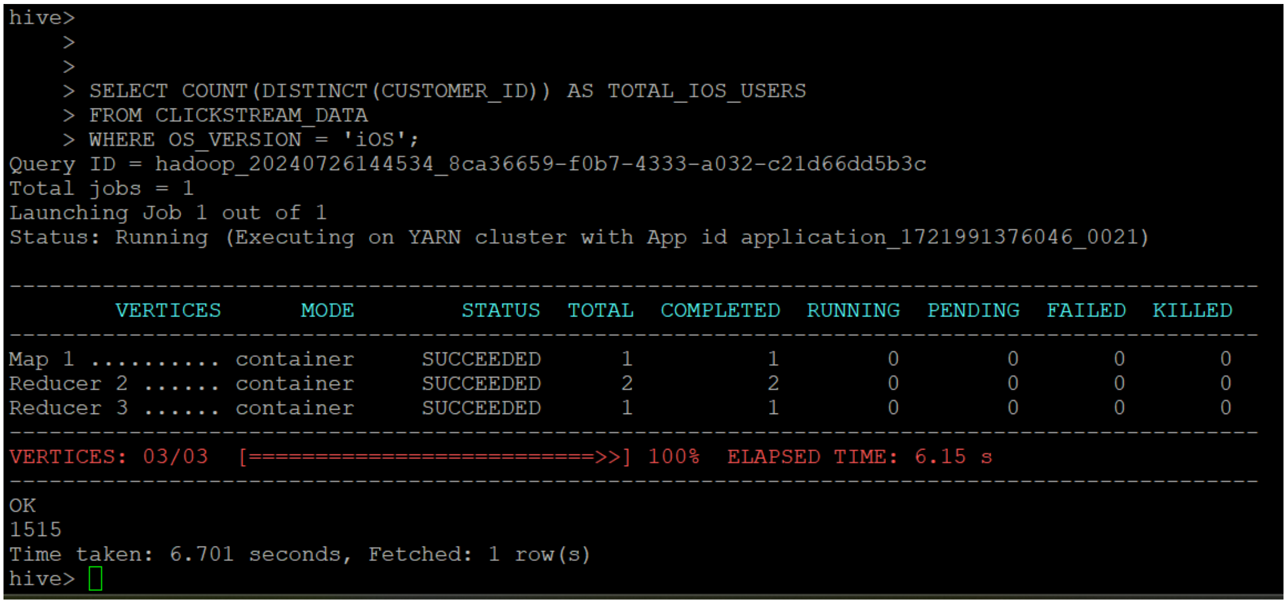
	trip_month	no_of_bookings
1	2020-01	26
2	2020-02	16
3	2020-03	16
4	2020-04	21
5	2020-05	21
6	2020-06	14
7	2020-07	20
8	2020-08	32
9	2020-09	21
10	2020-10	15

Task 11: Calculate the count of total iOS users.

Query:-

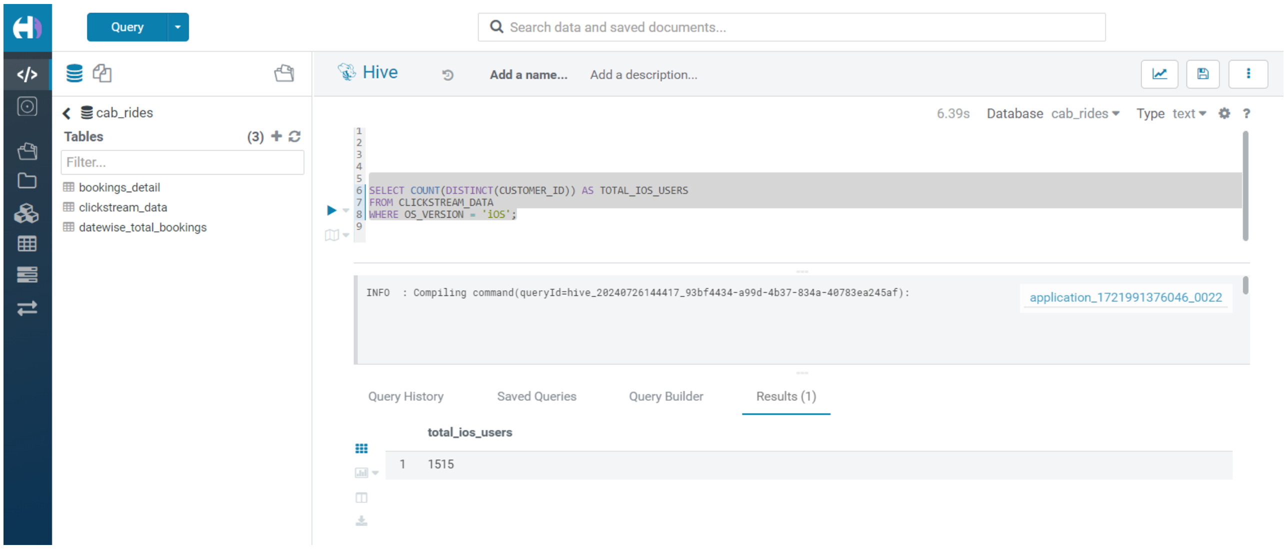
SELECT COUNT(DISTINCT(CUSTOMER_ID)) AS TOTAL_IOS_USERS
FROM CLICKSTREAM_DATA
WHERE OS_VERSION = 'iOS';

Output:-



```
hive>
    >
    >
    > SELECT COUNT(DISTINCT(CUSTOMER_ID)) AS TOTAL_IOS_USERS
    > FROM CLICKSTREAM_DATA
    > WHERE OS_VERSION = 'iOS';
Query ID = hadoop_20240726144534_8ca36659-f0b7-4333-a032-c21d66dd5b3c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.15 s
----------------------------------------------------------------------------------------------
OK
1515
Time taken: 6.701 seconds, Fetched: 1 row(s)
hive>
```

** There is slight difference 12 records as per the validation document due to the additional 16 records present in the loaded clickstream data in HDFS.



cab_rides

Tables (3)

Filter...

bookings_detail
clickstream_data
datewise_total_bookings

6.39s Database cab_rides Type text

```
SELECT COUNT(DISTINCT(CUSTOMER_ID)) AS TOTAL_IOS_USERS
FROM CLICKSTREAM_DATA
WHERE OS_VERSION = 'iOS';
```

```
INFO  : Compiling command(queryId=hive_20240726144417_93bf4434-a99d-4b37-834a-40783ea245af):
```

application_1721991376046_0022

	total_ios_users
1	1515