Logic For Final Submission

Queries with output and explanation:

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

Query:

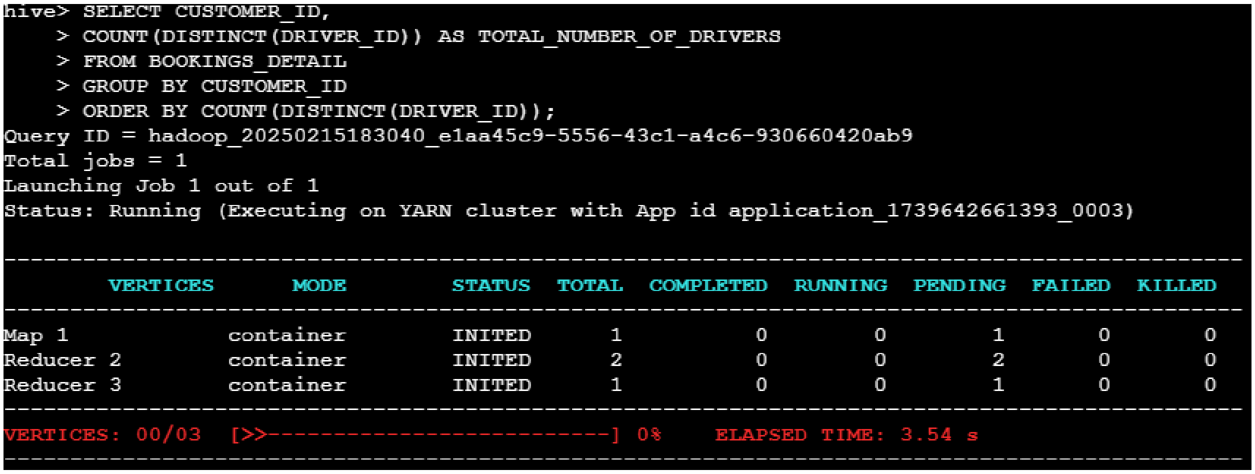
SELECT CUSTOMER_ID,
COUNT(DISTINCT(DRIVER_ID)) AS TOTAL_NUMBER_OF_DRIVERS
FROM BOOKINGS_DETAIL
GROUP BY CUSTOMER_ID
ORDER BY COUNT(DISTINCT(DRIVER_ID));

Explanation:

This query counts the unique drivers (driver_id) by each customer_id. This will give the total number of different drivers for each customer. The query first groups the data by customer_id and then counts number of different driver_ids for each customer_id. The result is ordered by count of distinct driver_id. The output will be sorted in ascending order of count of distinct driver_ids.

This query will give insights on customers who have booked rides with multiple drivers.

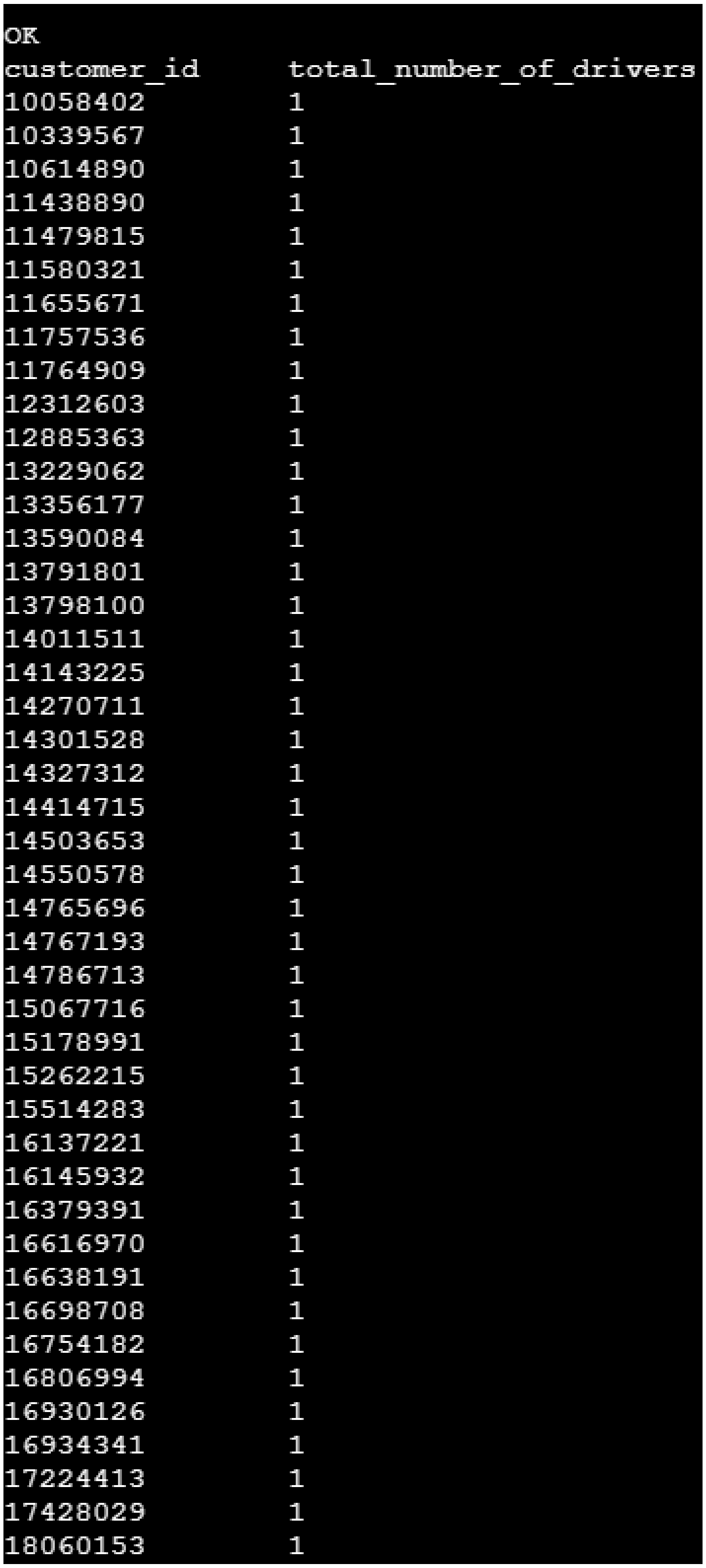
Screenshot of Query execution:



```
hive> SELECT CUSTOMER_ID,
    > COUNT(DISTINCT(DRIVER_ID)) AS TOTAL_NUMBER_OF_DRIVERS
    > FROM BOOKINGS_DETAIL
    > GROUP BY CUSTOMER_ID
    > ORDER BY COUNT(DISTINCT(DRIVER_ID));
Query ID = hadoop_20250215183040_e1aa45c9-5556-43c1-a4c6-930660420ab9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739642661393_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container       INITED      1          0        0        1       0       0
Reducer 2        container       INITED      2          0        0        2       0       0
Reducer 3        container       INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 3.54 s
----------------------------------------------------------------------------------------------
```

Screenshot of output:



```
OK
customer_id     total_number_of_drivers
10058402        1
10339567        1
10614890        1
11438890        1
11479815        1
11580321        1
11655671        1
11757536        1
11764909        1
12312603        1
12885363        1
13229062        1
13356177        1
13590084        1
13791801        1
13798100        1
14011511        1
14143225        1
14270711        1
14301528        1
14327312        1
14414715        1
14503653        1
14550578        1
14765696        1
14767193        1
14786713        1
15067716        1
15178991        1
15262215        1
15514283        1
16137221        1
16145932        1
16379391        1
16616970        1
16638191        1
16698708        1
16754182        1
16806994        1
16930126        1
16934341        1
17224413        1
17428029        1
18060153        1
```

Note: Result is matching with the validation document output

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

Query:

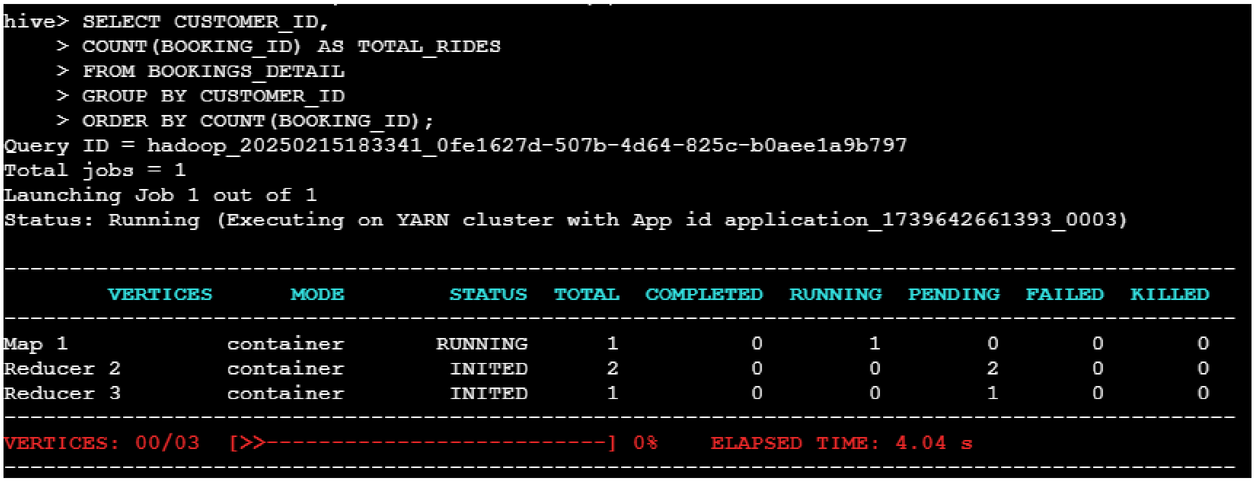
SELECT CUSTOMER_ID,
COUNT(BOOKING_ID) AS TOTAL_RIDES
FROM BOOKINGS_DETAIL
GROUP BY CUSTOMER_ID
ORDER BY COUNT(BOOKING_ID);

Explanation:

This query counts the number of rides (booking_id) by each customer. This will give the total number of rides booked by each customer. The query first groups the data by customer_id and then counts number of booking_id for each customer_id. The result is ordered by count of booking_id. The output will be sorted in ascending order of count of booking_ids.

This query will give insights on customers who have booked rides more often. This could help find out most valuable customers for the company and company can roll out offer for such frequent riders.

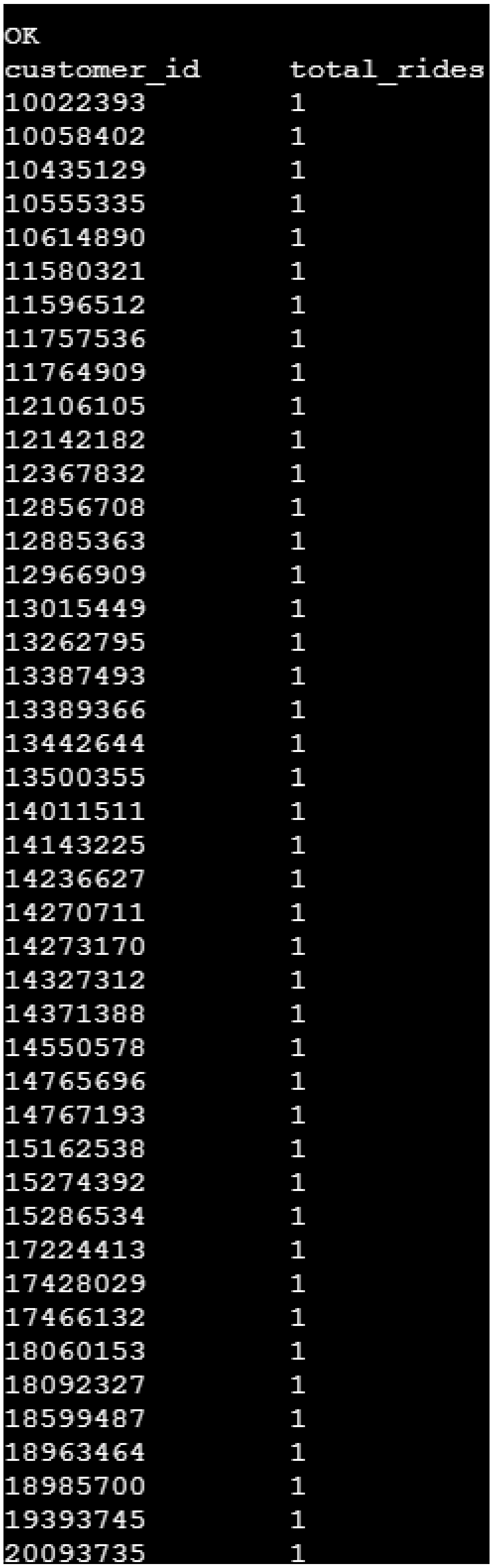
Screenshot of Query execution:



```
hive> SELECT CUSTOMER_ID,
    > COUNT(BOOKING_ID) AS TOTAL_RIDES
    > FROM BOOKINGS_DETAIL
    > GROUP BY CUSTOMER_ID
    > ORDER BY COUNT(BOOKING_ID);
Query ID = hadoop_20250215183341_0fe1627d-507b-4d64-825c-b0aee1a9b797
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739642661393_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container       RUNNING      1          0        1        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 4.04 s
----------------------------------------------------------------------------------------------
```

Screenshot of output:



```
OK
customer_id     total_rides
10022393        1
10058402        1
10435129        1
10555335        1
10614890        1
11580321        1
11596512        1
11757536        1
11764909        1
12106105        1
12142182        1
12367832        1
12856708        1
12885363        1
12966909        1
13015449        1
13262795        1
13387493        1
13389366        1
13442644        1
13500355        1
14011511        1
14143225        1
14236627        1
14270711        1
14273170        1
14327312        1
14371388        1
14550578        1
14765696        1
14767193        1
15162538        1
15274392        1
15286534        1
17224413        1
17428029        1
17466132        1
18060153        1
18092327        1
18599487        1
18963464        1
18985700        1
19393745        1
20093735        1
```

Note: Result is matching with the validation document output.

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:

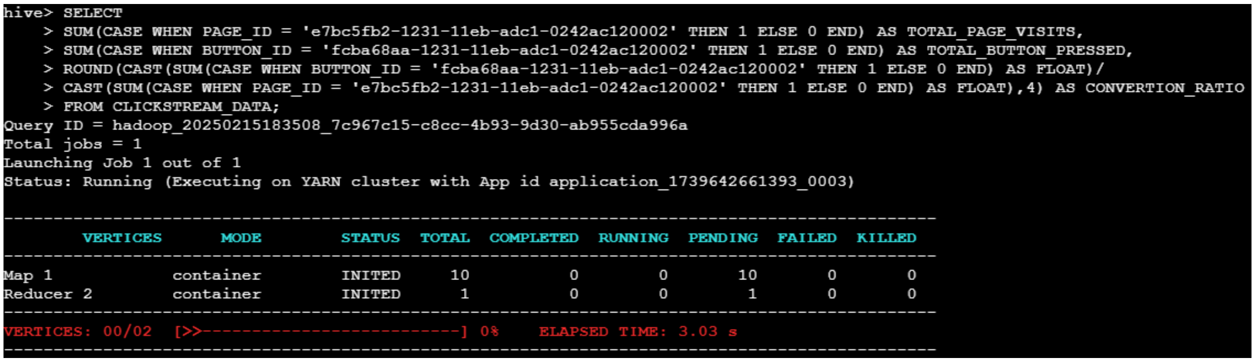
```
SELECT
SUM(CASE WHEN PAGE_ID = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' THEN 1 ELSE 0
END) AS TOTAL_PAGE_VISITS,
SUM(CASE WHEN BUTTON_ID = 'fcba68aa-1231-11eb-adc1-0242ac120002' THEN 1 ELSE
0 END) AS TOTAL_BUTTON_PRESSED,
ROUND(CAST(SUM(CASE WHEN BUTTON_ID = 'fcba68aa-1231-11eb-adc1-0242ac120002'
THEN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

Explanation:

This query will count the total number of page visits made by customers on the booking page and also count the total number of times Book Now button was pressed by customer. Then we calculate the conversion ratio of page visit to button press which is total number of Book Now button press divided by total visit made by customers on the booking page.

This will give insight into customer’s behaviour like when the visit booking page how many times they book a ride. Conversion ratio is metric which indicates the probability of booking if any customer visit booking page.

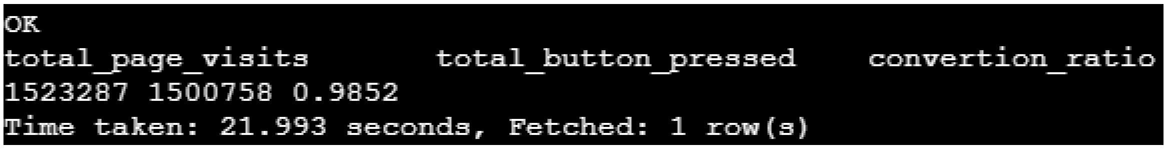
Screenshot of Query execution:



```
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 CONVERTION_RATIO
    > FROM CLICKSTREAM_DATA;
Query ID = hadoop_20250215183508_7c967c15-c8cc-4b93-9d30-ab955cda996a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739642661393_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container       INITED     10          0        0       10       0       0
Reducer 2        container       INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/02  [>>--------------------------] 0%    ELAPSED TIME: 3.03 s
----------------------------------------------------------------------------------------------
```

Screenshot of output:



```
OK
total_page_visits       total_button_pressed    convertion_ratio
1523287 1500758 0.9852
Time taken: 21.993 seconds, Fetched: 1 row(s)
```

Note: The conversion rate is 0.9852. compared to 0.9688 in validation document. This is because higher number of records in clickstream data table.

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

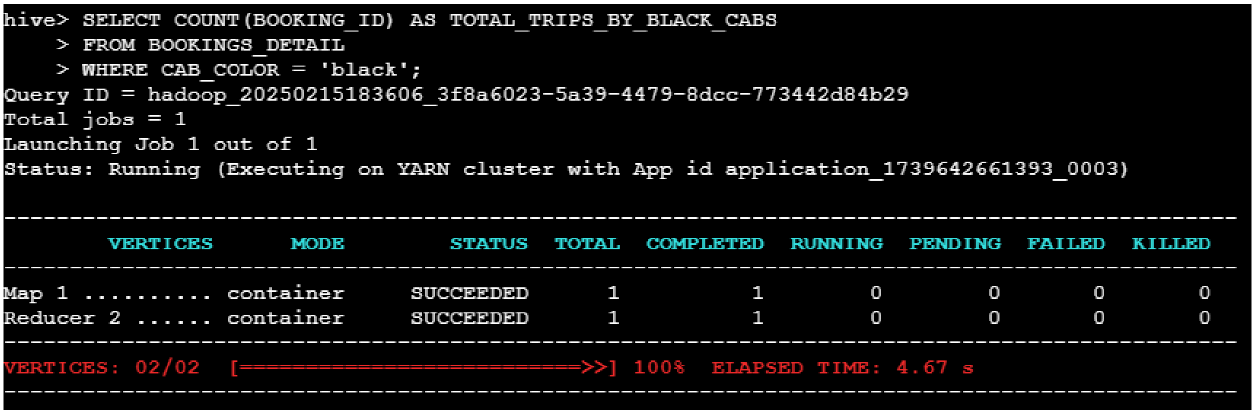
Query:

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

Explanation:

This query will give total number trips done by black cabs. We used Where condition to filter CAB_COLOR=’black’ and then count total number of booking_ids. This will identify colour preference of customer booking rides.

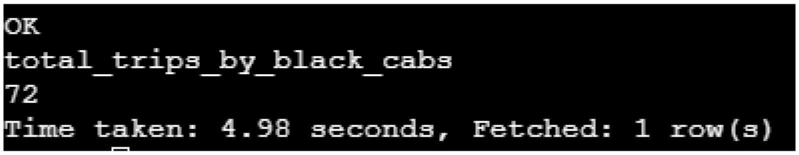
Screenshot of Query execution:



```
hive> SELECT COUNT(BOOKING_ID) AS TOTAL_TRIPS_BY_BLACK_CABS
    > FROM BOOKINGS_DETAIL
    > WHERE CAB_COLOR = 'black';
Query ID = hadoop_20250215183606_3f8a6023-5a39-4479-8dcc-773442d84b29
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739642661393_0003)

----------------------------------------------------------------------------------------------
        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: 4.67 s
----------------------------------------------------------------------------------------------
```

Screenshot of output:



```
OK
total_trips_by_black_cabs
72
Time taken: 4.98 seconds, Fetched: 1 row(s)
```

Note: Count of all trips done on black cabs: 72 Matches with validation document count.

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

Query:

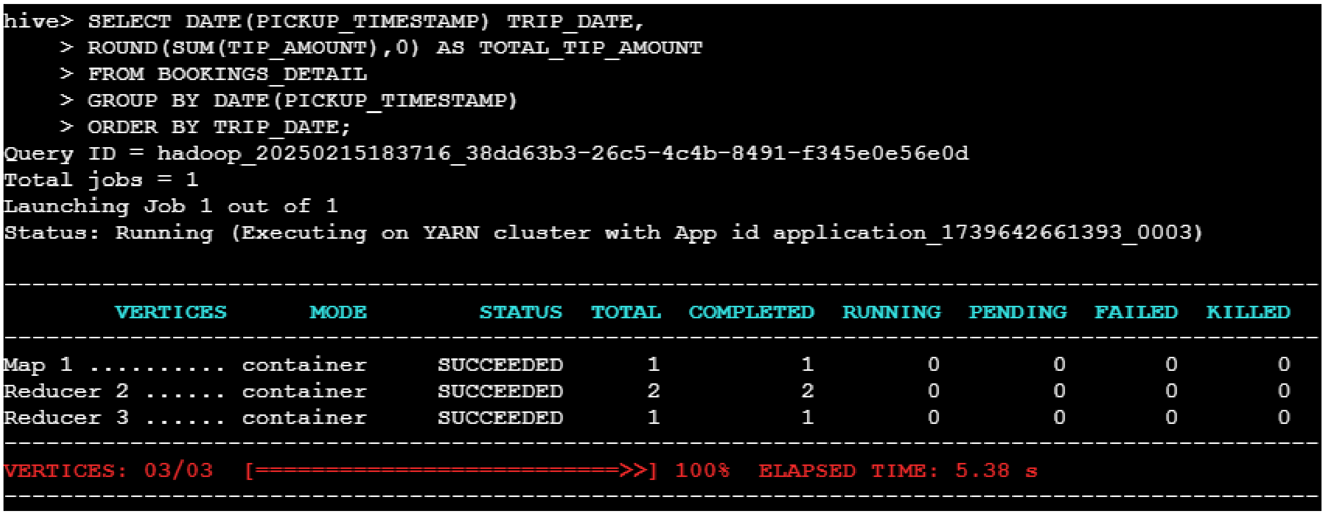
SELECT DATE(PICKUP_TIMESTAMP) TRIP_DATE,
ROUND(SUM(TIP_AMOUNT),0) AS TOTAL_TIP_AMOUNT
FROM BOOKINGS_DETAIL
GROUP BY DATE(PICKUP_TIMESTAMP)
ORDER BY TRIP_DATE;

Explanation:

This query will give total amount of tips given by customers to all drivers date wise. First, we extract the date from timestamp column to get date and group by the date column. Then we sum up the total tips on each day. The output is ordered by the date in ascending order.

This will help us to identify if customers give more tips on any specific occasion or specific dates. Based on this analysis, the company can provide offers to customers on such occasions.

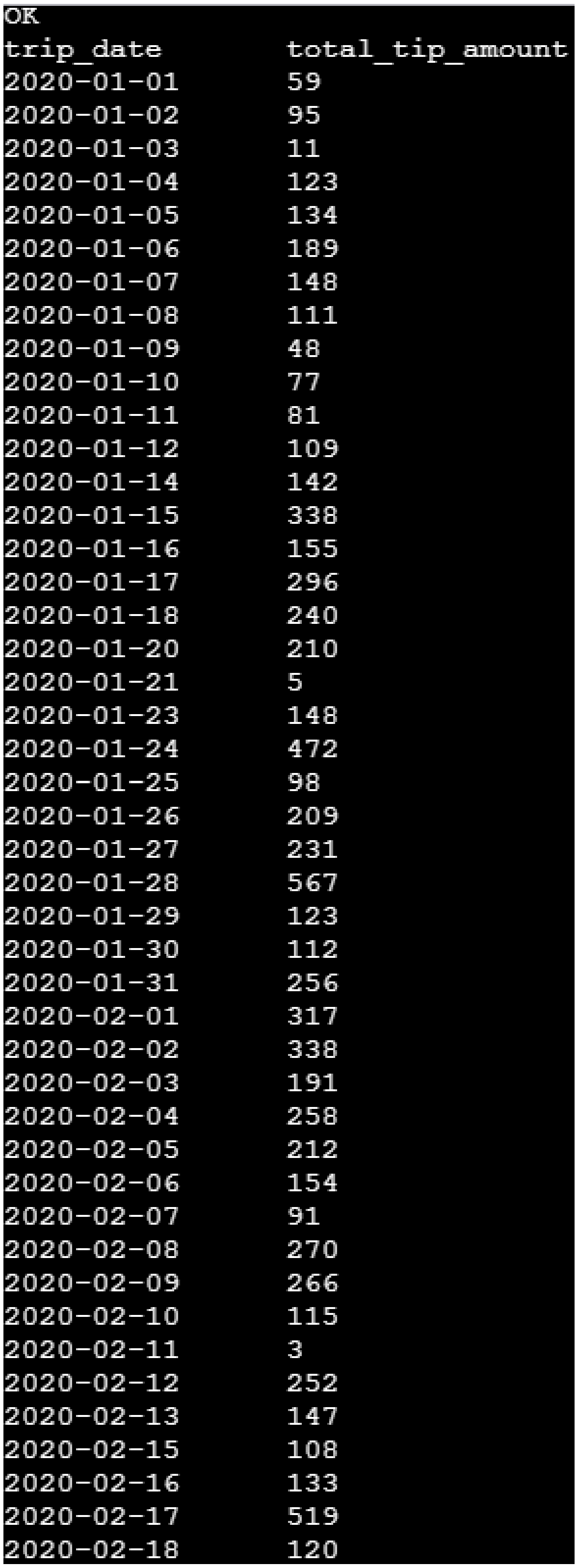
Screenshot of Query execution:



```
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_20250215183716_38dd63b3-26c5-4c4b-8491-f345e0e56e0d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739642661393_0003)

----------------------------------------------------------------------------------------------
        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: 5.38 s
----------------------------------------------------------------------------------------------
```

Screenshot of output:



```
OK
trip_date         total_tip_amount
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
2020-01-24        472
2020-01-25        98
2020-01-26        209
2020-01-27        231
2020-01-28        567
2020-01-29        123
2020-01-30        112
2020-01-31        256
2020-02-01        317
2020-02-02        338
2020-02-03        191
2020-02-04        258
2020-02-05        212
2020-02-06        154
2020-02-07        91
2020-02-08        270
2020-02-09        266
2020-02-10        115
2020-02-11        3
2020-02-12        252
2020-02-13        147
2020-02-15        108
2020-02-16        133
2020-02-17        519
2020-02-18        120
```

Note: The output is exactly matching with validation document output.

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

Query:

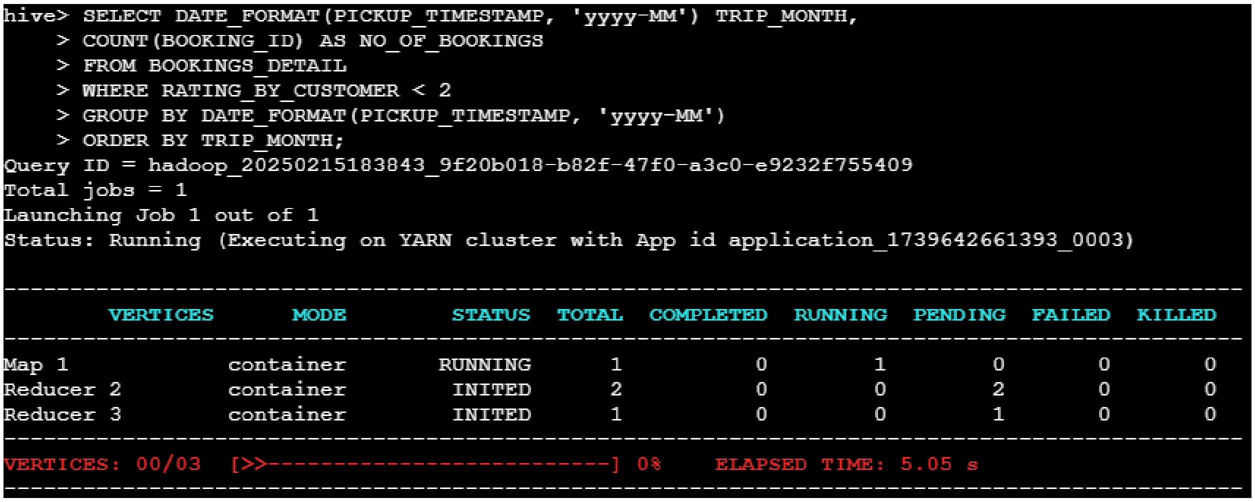
SELECT DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM') TRIP_MONTH,
COUNT(BOOKING_ID) AS NO_OF_BOOKINGS
FROM BOOKINGS_DETAIL
WHERE RATING_BY_CUSTOMER < 2
GROUP BY DATE_FORMAT(PICKUP_TIMESTAMP, 'yyyy-MM')
ORDER BY TRIP_MONTH;

Explanation:

DATE_FORMAT function formats datetimestamp value in the specified format like yyyy-MM which results like 2023-06. Then we used WHERE clause which is used to filter bookings where rating given by customers is less than 2 which indicates customer dissatisfaction. ORDER BY clause with Trip month alias is used to show output in ascending order of pickup month.

This analysis could help to understand number of trips by month where customers were not happy. Also, could give insight or a hidden pattern in dissatisfactory rides in a specific month or period which could be different factors like low rating because of AC was not on, late pickup and drop traffic etc. Based on this analysis, instructions can be given to driver to make customers happy and take care of things which could lead to low customer rating.

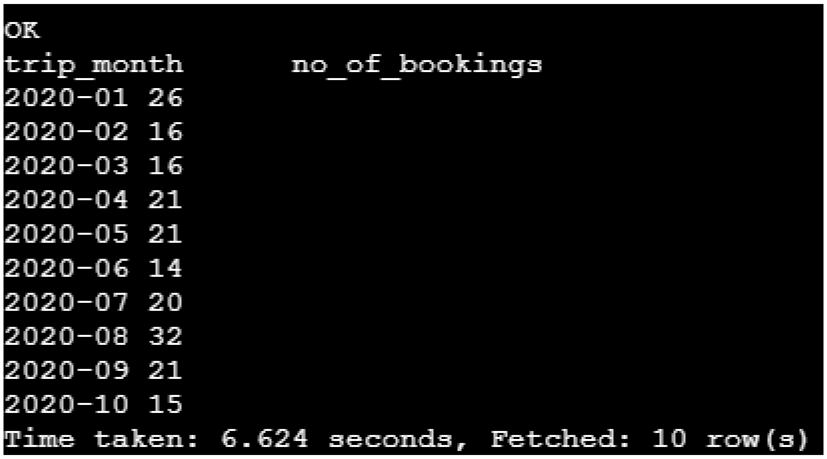
Screenshot of Query execution:



```
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_20250215183843_9f20b018-b82f-47f0-a3c0-e9232f755409
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739642661393_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container       RUNNING      1          0        1        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 5.05 s
----------------------------------------------------------------------------------------------
```

Screenshot of output:



```
OK
trip_month      no_of_bookings
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: 6.624 seconds, Fetched: 10 row(s)
```

Note: The output is exactly matching with validation document output.

Task 11: Calculate the count of total iOS users.

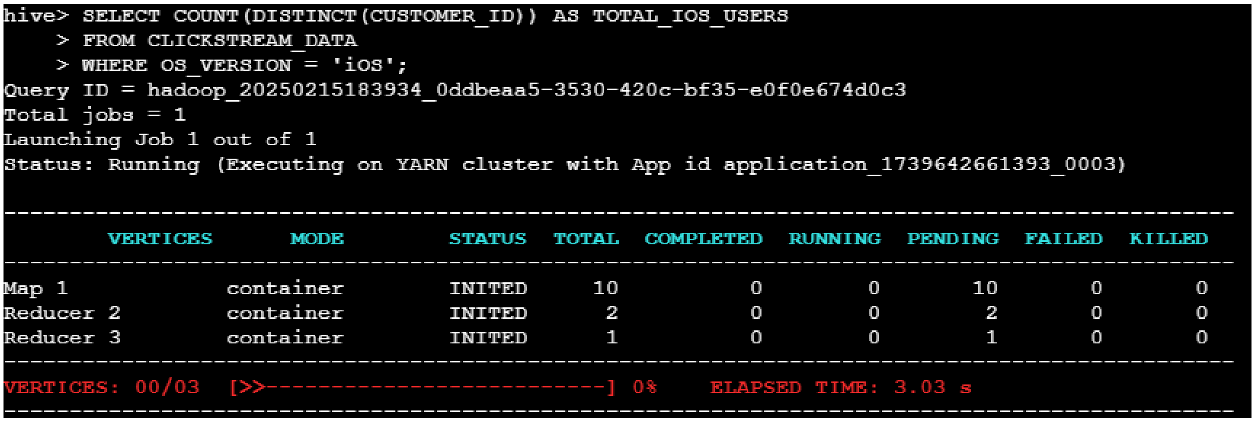
Query:

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

Explanation:
This query gives the count of distinct customers who are using iOS devices. Where clause is used to filter out customers who are using iOS devices.

This analysis will give insights into how many or percentage of customers using a specific type of devices.

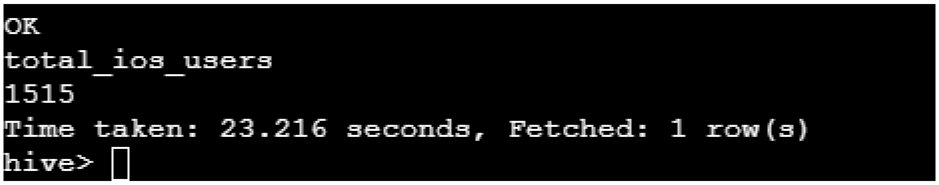
Screenshot of Query execution:



```
hive> SELECT COUNT(DISTINCT(CUSTOMER_ID)) AS TOTAL_IOS_USERS
    > FROM CLICKSTREAM_DATA
    > WHERE OS_VERSION = 'iOS';
Query ID = hadoop_20250215183934_0ddbeaa5-3530-420c-bf35-e0f0e674d0c3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739642661393_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container       INITED     10          0        0       10       0       0
Reducer 2        container       INITED      2          0        0        2       0       0
Reducer 3        container       INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 3.03 s
----------------------------------------------------------------------------------------------
```

Screenshot of output:



```
OK
total_ios_users
1515
Time taken: 23.216 seconds, Fetched: 1 row(s)
hive> 
```

Note: The output is 1515 compared top 1503 due to higher number of records in clickstream_data table.