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 COUNT(DISTINCT(DRIVER_ID));
```

Output:

```
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1739507174616_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: 5.78 s
--------------------------------------------------------------------------------
OK
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
```

Note: Result is matching with the validation document output.

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

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);
```

Output:

```
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739507174616_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: 5.51 s
--------------------------------------------------------------------------------
OK
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
```

Note: Result is matching with the validation document output.

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

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;
```

Output:

```
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_20250214053009_79372e62-67b5-49f9-a2f2-9dd51ab2f669
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739507174616_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 22.83 s
--------------------------------------------------------------------------------
OK
total_page_visits       total_button_pressed    convertion_ratio
1205385 1187543 0.9852
Time taken: 23.053 seconds, Fetched: 1 row(s)
hive>
```

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.

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

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';
```

Output:

```
hive> SELECT COUNT(BOOKING_ID) AS TOTAL_TRIPS_BY_BLACK_CABS
    > FROM BOOKINGS_DETAIL
    > WHERE CAB_COLOR = 'black';
Query ID = hadoop_20250214053817_7c3d388a-5d4f-459f-a171-0307c90c56e8
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1739507174616_0004)

--------------------------------------------------------------------------------
        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: 5.89 s
--------------------------------------------------------------------------------
OK
total_trips_by_black_cabs
72
```

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;
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

Output:

```
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_20250214054116_be545579-e803-45d8-83a3-9447bd4d54e5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739507174616_0004)

--------------------------------------------------------------------------------
        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: 5.05 s
--------------------------------------------------------------------------------
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
```

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;
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

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_20250214054637_9bbe7e98-7e7f-4de5-810a-c4d1970af451
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739507174616_0004)

--------------------------------------------------------------------------------
        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: 6.53 s
--------------------------------------------------------------------------------
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: 6.852 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';
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

Output:

```
hive> SELECT COUNT(DISTINCT(CUSTOMER_ID)) AS TOTAL_IOS_USERS
    > FROM CLICKSTREAM_DATA
    > WHERE OS_VERSION = 'iOS';
Query ID = hadoop_20250214054847_6c35b1ec-0e78-4472-bb64-b35e389750c6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739507174616_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        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: 22.37 s
--------------------------------------------------------------------------------
OK
total_ios_users
1515
```

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

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved