

Queries

** While loading the data please add the following code

Load data inpath 'location of part file' into table 'table name'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','

<Hive Query for Task 5>

Select customer_id,count(distinct driver_id)
 From bookings
 Group by customer_id

- <Screenshot after executing Query>

```
Time taken: 1.367 seconds, Fetched: 10 row(s)
hive> Select customer_id,count(distinct driver_id) as driver_count from bookings group by customer_id;
Query ID = hadoop_20240202063441_2c268979-831e-4b23-b016-bef9e6f7f11b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0010)

-----
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
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 5.16 s
-----
OK
10022393      1
10435129      1
10555335      1
10592274      1
10678994      1
11264797      1
11353346      1
11418437      1
11454977      1
11518953      1
11596512      1
11608791      1
11860278      1
11981042      1
12106105      1
12142182      1
12334699      1
12367832      1
12856708      1
12913608      1
12914577      1
12966909      1
13015449      1
13262795      1
13387493      1
13389366      1
13442644      1
```

<Hive Query for Task 6>

Select customer_id, count(distinct booking_id) as booking_count
 From bookings
 Group by customer_id

- <Screenshot after executing Query>

```
Time taken: 1.36 seconds, Fetched: 10 row(s)
hive> Select count(distinct driver_id) as driver_count from bookings group by customer_id;
Query ID = hadoop_20240202063441_2c268979-831e-4b23-b016-bef9e6f7f11b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0010)

-----
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
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 5.16 s

OK
10022393      1
10435129      1
10555335      1
10592274      1
10678994      1
11264797      1
11353346      1
11418437      1
11454977      1
11518953      1
11596512      1
11608791      1
11860278      1
11981042      1
12106105      1
12142102      1
12334699      1
12367832      1
```

<Hive Query for Task 7>

Select count(case when is_button_click="Yes" then button_id end)/count(case when is_page_view = "Yes" then page_id end) as conversion_ratio
from clickstream_data
where customer_id is not null;

• <Screenshot after executing Query>

```
hive> Select count(case when is_button_click="Yes" then button_id end)/count(case when is_page_view = "Yes" then page_id end) as conversion_ratio from clickstream_data where customer_id is not null;
Query ID = hadoop_20240202070519_75d22ad7-3070-495e-ae06-caac0ca334d4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0012)

-----
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: 5.28 s

OK
0.9887566137566137
Time taken: 5.61 seconds, Fetched: 1 row(s)
hive>
```

Select count(case when is_page_view = "Yes" then page_id end) as page_views ,
count(case when is_button_click="Yes" then button_id end) as button_click
from clickstream_data
where customer_id is not null;

```
hive> Select count(case when is_page_view = "Yes" then page_id end) as page_views,count(case when is_button_click="Yes" then button_id end) as button_clicks from clickstream_data where customer_id is not null;
Query ID = hadoop_20240202070119_02d2fef8-52a9-4dd1-9d95-9582fbb00836
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0012)

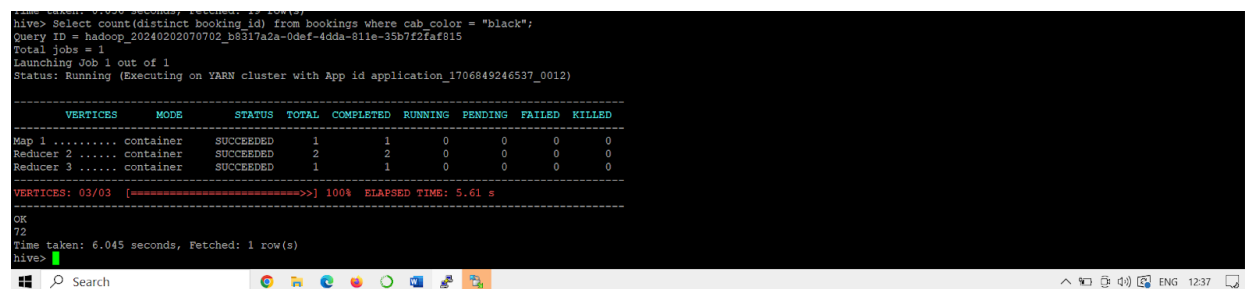
-----
VERTICES      MODE        STATUS      TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1        1          0        0        0        0
Reducer 2 ..... container  SUCCEEDED    1        1          0        0        0        0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 5.92 s

OK
1512      1495
Time taken: 6.213 seconds, Fetched: 1 row(s)
hive>
```

<Hive Query for Task 8>

```
Select count(distinct booking_id)
from bookings
where cab_Color='black'
```

- <Screenshot after executing Query>



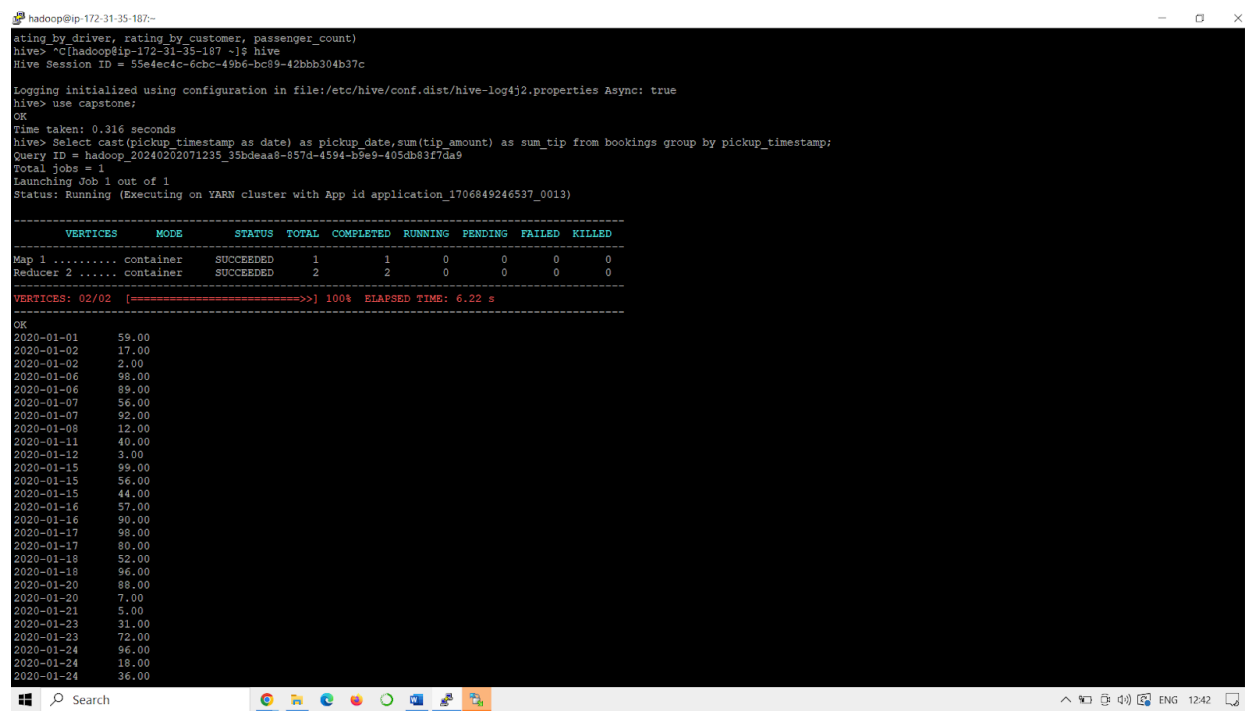
```
hive> select count(distinct booking_id) from bookings where cab_color = "black";
Query ID = hadoop_20240202070702_b8317a2a-0def-4dda-811e-35b7f2fa815
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0012)

-----
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: 5.61 s
OK
72
Time taken: 6.045 seconds, Fetched: 1 row(s)
hive>
```

<Hive Query for Task 9>

```
Select cast(pickup_timestamp as date) as pickup_date,sum(tip_amount) as sum_tip
from bookings
group by pickup_timestamp;
```

- <Screenshot after executing Query>



```
hadoop@ip-172-31-35-187:~$
ating_by driver, rating by customer, passenger_count)
hive> ^C[hadoop@ip-172-31-35-187 ~]$ hive
Hive Session ID = 55e4ec4c-6cbc-49b6-bc89-42bbb304b37c

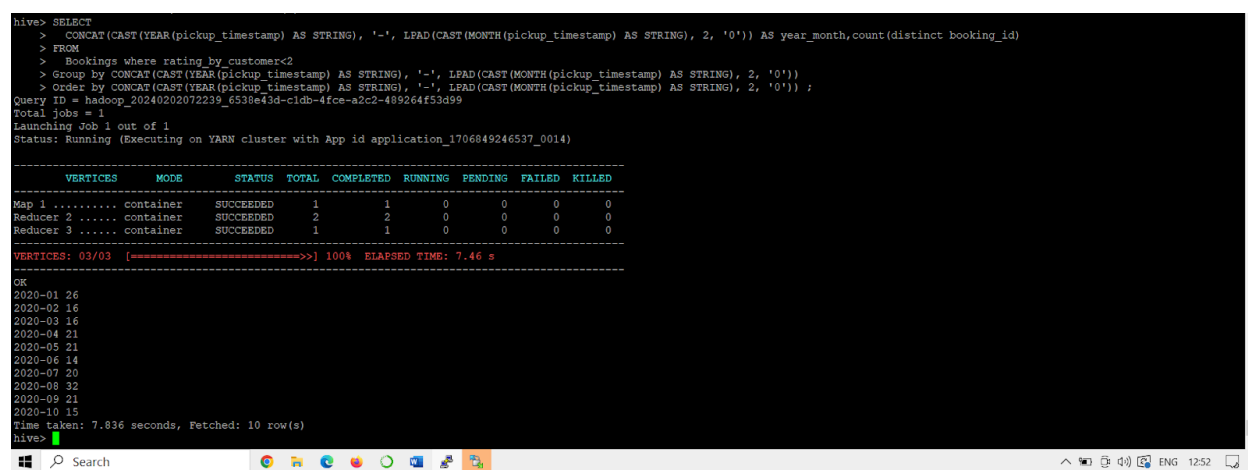
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> use capstone;
OK
Time taken: 0.316 seconds
hive> Select cast(pickup_timestamp as date) as pickup_date,sum(tip_amount) as sum_tip from bookings group by pickup_timestamp;
Query ID = hadoop_20240202071235_35bdeaa8-857d-4594-b9e9-405db83f7da9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0013)

-----
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
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 6.22 s
OK
2020-01-01      59.00
2020-01-02      17.00
2020-01-02       2.00
2020-01-06     98.00
2020-01-06     89.00
2020-01-07     56.00
2020-01-07     92.00
2020-01-08     12.00
2020-01-11     40.00
2020-01-12       3.00
2020-01-15     99.00
2020-01-15     56.00
2020-01-15     44.00
2020-01-16     57.00
2020-01-16     90.00
2020-01-17     98.00
2020-01-17     80.00
2020-01-18     52.00
2020-01-18     96.00
2020-01-20     88.00
2020-01-20       7.00
2020-01-21       5.00
2020-01-23     31.00
2020-01-23     72.00
2020-01-24     96.00
2020-01-24     18.00
2020-01-24     36.00
```

<Hive Query for Task 10>

```
SELECT
CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-', LPAD(CAST(MONTH(pickup_timestamp) AS
STRING), 2, '0')) AS year_month,count(distinct booking_id)
FROM
Bookings where rating_by_customer<2
Group by CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-',
LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0'))
Order by CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-',
LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0')) ;
```

- <Screenshot after executing Query>



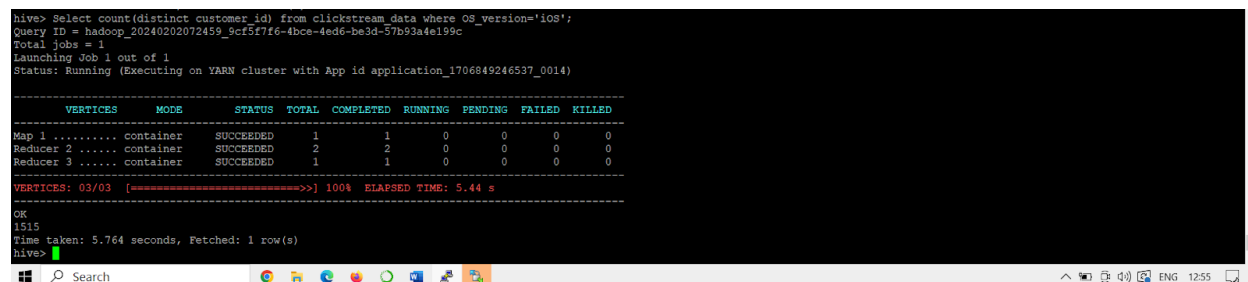
```
hive> SELECT
>   CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-', LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0')) AS year_month,count(distinct booking_id)
> FROM
>   Bookings where rating_by_customer<2
> Group by CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-', LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0'))
> Order by CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-', LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0')) ;
Query ID = hadoop_20240202072239_6530e43d-cldb-4fce-a2c2-489264f53d99
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0014)

-----
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: 7.46 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: 7.836 seconds, Fetched: 10 row(s)
hive>
```

<Hive Query for Task 11>

```
Select count(distinct customer_id)
from clickstream_data
where OS_version='iOS';
```

- <Screenshot after executing Query>



```
hive> Select count(distinct customer_id) from clickstream_data where OS_version='iOS';
Query ID = hadoop_20240202072459_9cf57f76-4bce-4ed6-be3d-57b93ade199c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0014)

-----
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: 5.44 s
-----
OK
1515
Time taken: 5.764 seconds, Fetched: 1 row(s)
hive>
```