

Queries

<Hive Query for Task 5>

```
select customer_id ,count( DISTINCT driver_id) from bookings
group by customer_id
order by customer_id asc;
```

```
hive> select customer_id ,count( DISTINCT driver_id) from booking
> group by customer_id
> order by customer_id asc;
Query ID = ec2-user_20221009193047_d14508b2-2779-47ac-97f9-bd457797a9b5
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1665331741273_0019)

Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 1(+1)/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
OK
10022393      1
10058402      1
10339567      1
10435129      1

Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
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
```

<Hive Query for Task 6>

```
select customer_id ,count( DISTINCT booking_id) from booking
group by customer_id
order by customer_id asc;
```

<Screenshot after executing Query>

```
hive> select customer_id ,count( DISTINCT booking_id) from booking
> group by customer_id
> order by customer_id asc;
Query ID = ec2-user_20221009193635_105825f5-4937-40b1-82c0-a6ef564b14ac
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1665331741275_0020)

Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 1(+1)/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
OK
10022393      1
10058402      1

Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
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
```

<Hive Query for Task 7>

```
select count(page_id), sum( case when is_button_click = 'Yes' then 1 else 0 end ), sum( case
when is_button_click = 'Yes' then 1 else 0 end )/count(page_id)
from
```

clickstream

where page_id='e7bc5fb2-1231-11eb-adc1-0242ac120002' and button_id='fcba68aa-1231-11eb-adc1-0242ac120002';

<Screenshot after executing Query>

```
hadoop@ip-172-31-49-133:~
hive> select count(page_id), sum( case when is_button_click = 'Yes' then 1 else 0 end ), sum( case when
> is_button_click = 'Yes' then 1 else 0 end )/count(page_id)
> from
> clickstream
> where page_id='e7bc5fb2-1231-11eb-adc1-0242ac120002' and button_id='fcba68aa-1231-11eb-adc1-0242a
c120002';
Query ID = hadoop_20221009233400_a31e8143-867f-4cbc-b6f4-7156f7d1d13b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)

-----
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: 5.16 s
-----
OK
326      167      0.5122699386503068
Time taken: 6.058 seconds, Fetched: 1 row(s)
hive>
```

<Hive Query for Task 8>

select cab_color ,count(distinct driver_id) from booking
where cab_color in ('black')
group by cab_color ;

<Screenshot after executing Query>

```
hive> select cab_color ,count(distinct driver_id ) from booking
> where cab_color in ('black')
> group by cab_color ;
Query ID = hadoop_20221009233814_f9aee9df-73e3-46e5-b3df-9b97ef6edb24
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)

-----
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: 5.76 s
-----
OK
black      72
Time taken: 6.542 seconds, Fetched: 1 row(s)
```

<Hive Query for Task 9>

```
select date_format(pickup_timestamp,'yyyy-MM-dd'),sum(tip_amount) from booking
group by date_format(pickup_timestamp,'yyyy-MM-dd');
```

<Screenshot after executing Query>

```
hive> select date_format(pickup_timestamp,'yyyy-MM-dd'),sum(tip_amount) from booking
> group by date_format(pickup_timestamp,'yyyy-MM-dd');
Query ID = hadoop_20221009234224_39028a90-05d7-4b9f-b46c-f7d1e15a2e61
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0	0
Reducer 2	container	SUCCEEDED	2	2	0	0	0	0	0

```
VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 5.76 s
OK
2020-01-01      59
2020-01-02      95
2020-01-03      11
2020-01-04     123
2020-01-05     134
2020-01-08     111
2020-01-10      77
2020-01-11      81
2020-01-12     109
2020-01-16     155
2020-01-18     240
2020-01-20     210
2020-01-21       5
2020-01-26     209
2020-01-27     231
2020-01-29     123
```

<Hive Query for Task 10>

```
select date_format(pickup_timestamp,'yyyy-MM') ,count( rating_by_customer) from
booking
where rating_by_customer < 2
group by date_format(pickup_timestamp,'yyyy-MM') ;
```

<Screenshot after executing Query>

```
hive> select date_format(pickup_timestamp,'yyyy-MM') ,count( rating_by_customer) from
> booking
> where rating_by_customer < 2
> group by date_format(pickup_timestamp,'yyyy-MM') ;
Query ID = hadoop_20221009234536_1c876f02-c18e-408d-a67c-d548baf3da50
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)
```

	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: 5.66 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: 6.333 seconds, Fetched: 10 row(s)
```

<Hive Query for Task 11>

```
select os_version ,count(distinct customer_id) from clickstream
where os_version in ('iOS')
group by os_version;
```

<Screenshot after executing Query>

```
hive> select os_version ,count(distinct customer_id) from clickstream
> where os_version in ('iOS')
> group by os_version;
Query ID = hadoop_20221009234835_2f23002f-9c6f-4eeb-8d92-8cef32487b93
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)
```

	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: 5.48 s

```
OK
iOS      1515
Time taken: 6.113 seconds, Fetched: 1 row(s)
```