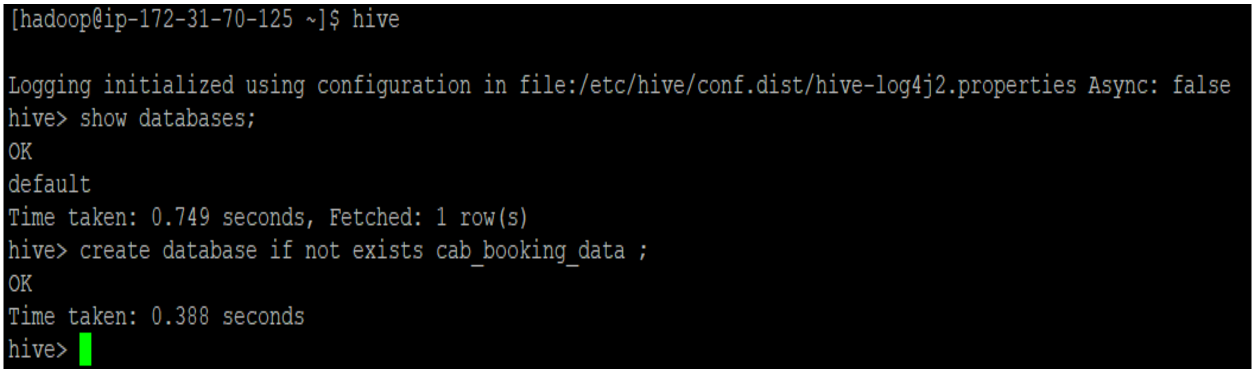
Create Hive-Managed Tables

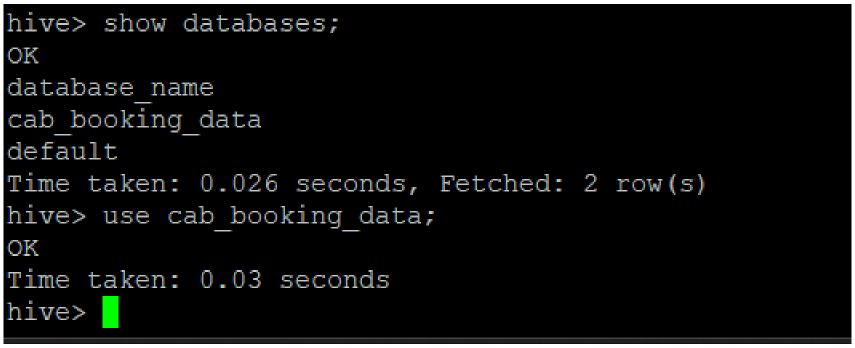
<Command to create the Hive tables>

1. First create a database
create database if not exists cab_booking_data ;
use cab_booking_data ;



```
[hadoop@ip-172-31-70-125 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show databases;
OK
default
Time taken: 0.749 seconds, Fetched: 1 row(s)
hive> create database if not exists cab_booking_data ;
OK
Time taken: 0.388 seconds
hive>
```

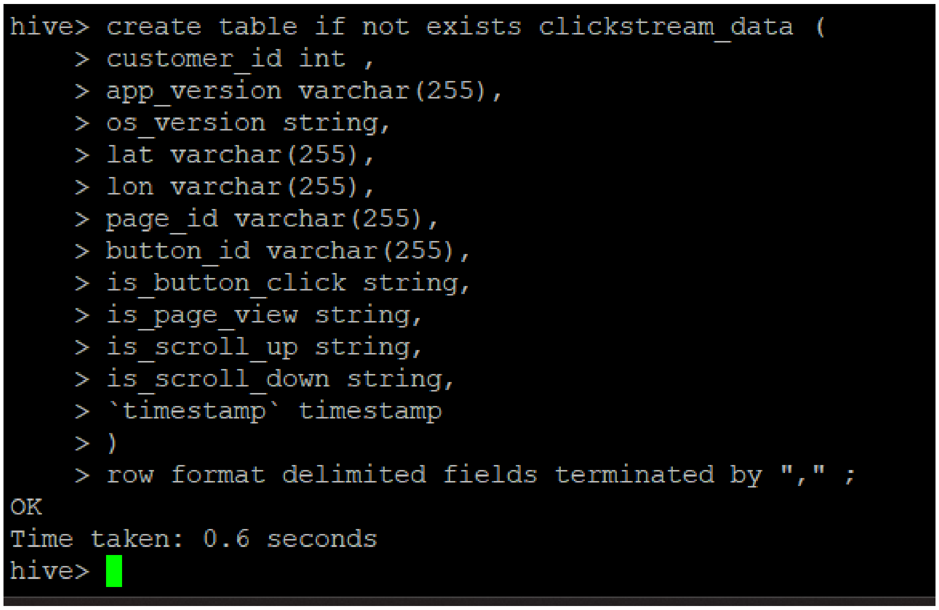


```
hive> show databases;
OK
database_name
cab_booking_data
default
Time taken: 0.026 seconds, Fetched: 2 row(s)
hive> use cab_booking_data;
OK
Time taken: 0.03 seconds
hive>
```

2. Creating a Hive-managed table for clickstream data

```
create table if not exists clickstream_data (
customer_id int ,
app_version varchar(255),
os_version string,
lat varchar(255),
lon varchar(255),
page_id varchar(255),
button_id varchar(255),
is_button_click string,
is_page_view string,
is_scroll_up string,
is_scroll_down string,
`timestamp` timestamp
```

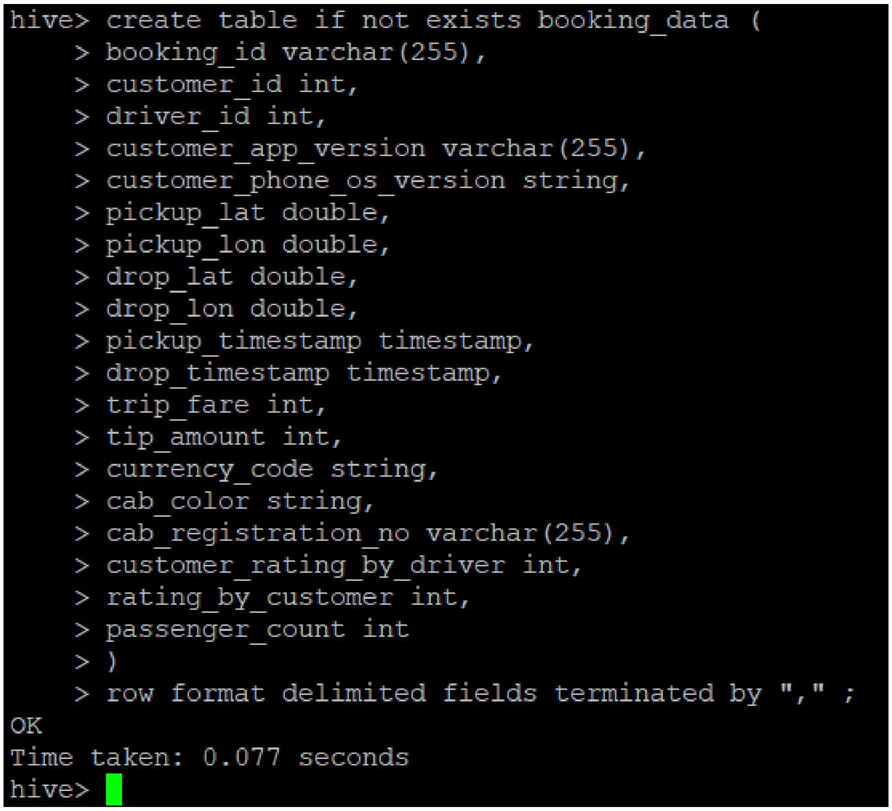
)
row format delimited fields terminated by "," ;



```
hive> create table if not exists clickstream_data (
    > customer_id int ,
    > app_version varchar(255),
    > os_version string,
    > lat varchar(255),
    > lon varchar(255),
    > page_id varchar(255),
    > button_id varchar(255),
    > is_button_click string,
    > is_page_view string,
    > is_scroll_up string,
    > is_scroll_down string,
    > `timestamp` timestamp
    > )
    > row format delimited fields terminated by "," ;
OK
Time taken: 0.6 seconds
hive>
```

3. Creating a Hive-managed table for bookings data

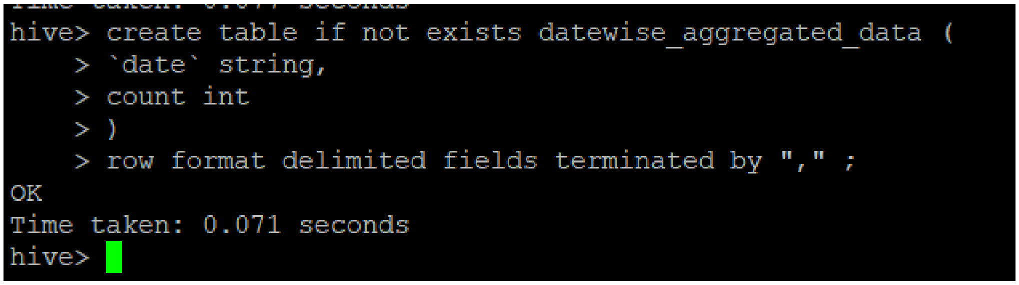
create table if not exists booking_data (
booking_id varchar(255),
customer_id int,
driver_id int,
customer_app_version varchar(255),
customer_phone_os_version string,
pickup_lat double,
pickup_lon double,
drop_lat double,
drop_lon double,
pickup_timestamp timestamp,
drop_timestamp timestamp,
trip_fare int,
tip_amount int,
currency_code string,
cab_color string,
cab_registration_no varchar(255),
customer_rating_by_driver int,
rating_by_customer int,
passenger_count int
)
row format delimited fields terminated by "," ;



```
hive> create table if not exists booking_data (
    > booking_id varchar(255),
    > customer_id int,
    > driver_id int,
    > customer_app_version varchar(255),
    > customer_phone_os_version string,
    > pickup_lat double,
    > pickup_lon double,
    > drop_lat double,
    > drop_lon double,
    > pickup_timestamp timestamp,
    > drop_timestamp timestamp,
    > trip_fare int,
    > tip_amount int,
    > currency_code string,
    > cab_color string,
    > cab_registration_no varchar(255),
    > customer_rating_by_driver int,
    > rating_by_customer int,
    > passenger_count int
    > )
    > row format delimited fields terminated by "," ;
OK
Time taken: 0.077 seconds
hive>
```

4. Creating a Hive-managed table for aggregated data in Task 3

create table if not exists datewise_aggregated_data (
`date` string,
count int
)
row format delimited fields terminated by "," ;



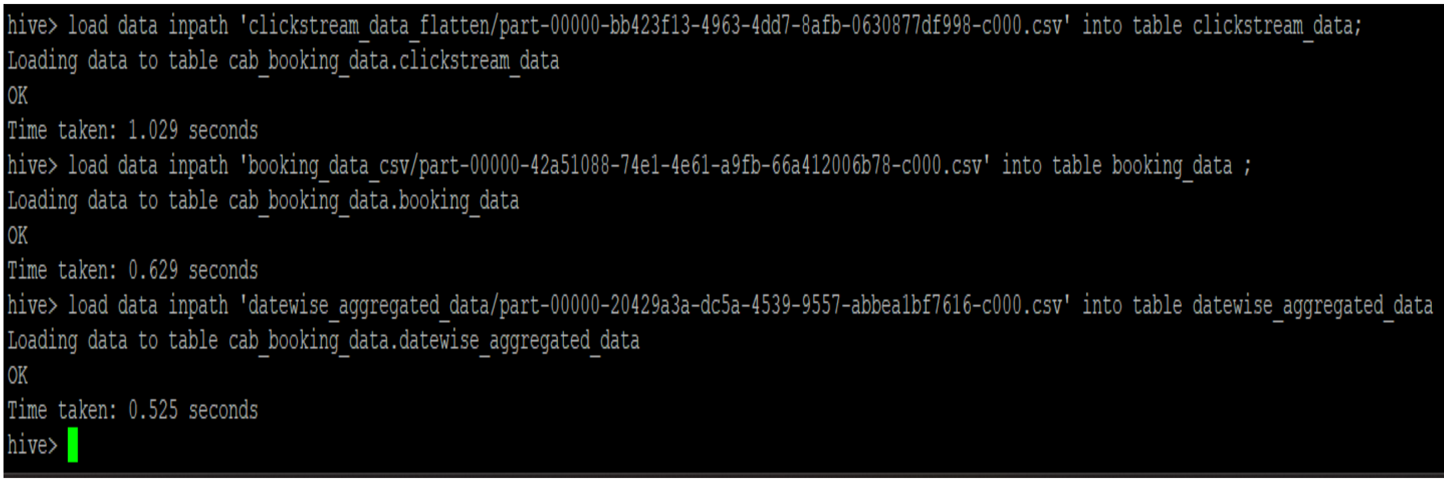
```
hive> create table if not exists datewise_aggregated_data (
    > `date` string,
    > count int
    > )
    > row format delimited fields terminated by "," ;
OK
Time taken: 0.071 seconds
hive>
```

<Command to load the data into Hive tables>

1. load data inpath 'clickstream_data_flatten/part-00000-bb423f13-4963-4dd7-8afb-0630877df998-c000.csv' into table clickstream_data ;

2. load data inpath 'booking_data_csv/part-00000-42a51088-74e1-4e61-a9fb-66a412006b78-c000.csv’ into table booking_data ;

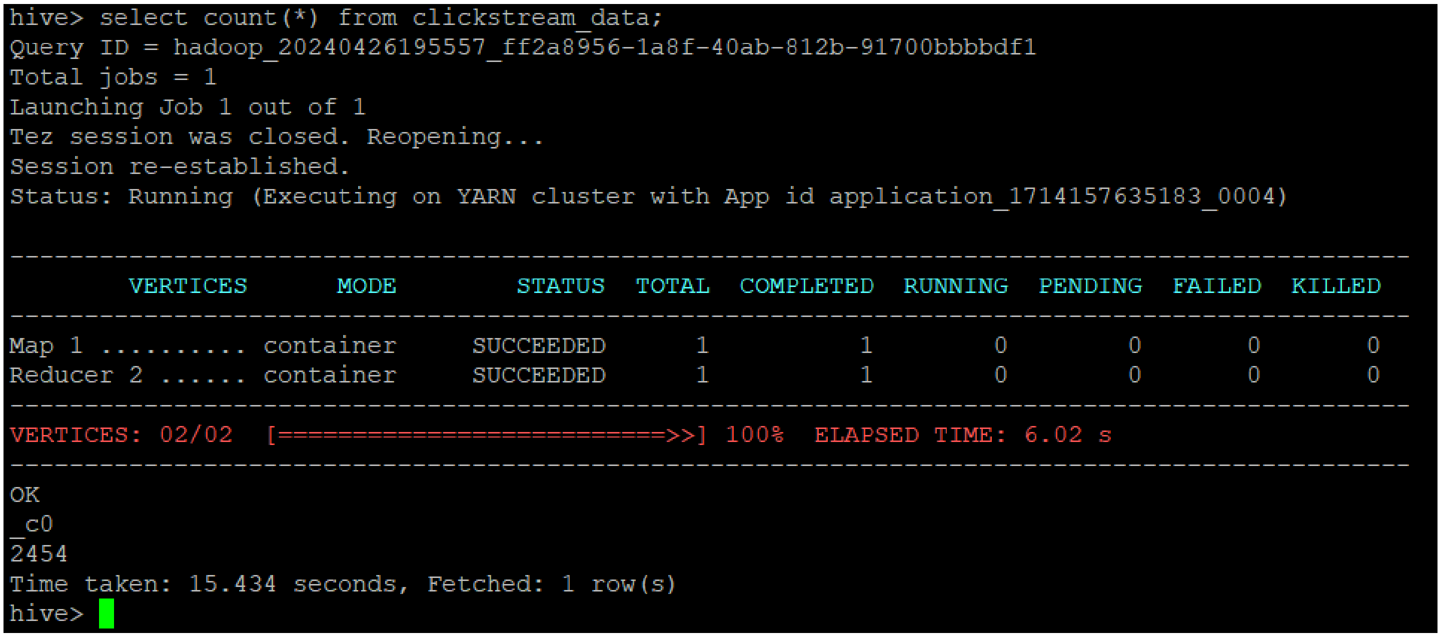
3. load data inpath 'datewise_aggregated_data/part-00000-20429a3a-dc5a-4539-9557-abbea1bf7616-c000.csv’ into table datewise_aggregated_data ;



```
hive> load data inpath 'clickstream_data_flatten/part-00000-bb423f13-4963-4dd7-8afb-0630877df998-c000.csv' into table clickstream_data;
Loading data to table cab_booking_data.clickstream_data
OK
Time taken: 1.029 seconds
hive> load data inpath 'booking_data_csv/part-00000-42a51088-74e1-4e61-a9fb-66a412006b78-c000.csv' into table booking_data ;
Loading data to table cab_booking_data.booking_data
OK
Time taken: 0.629 seconds
hive> load data inpath 'datewise_aggregated_data/part-00000-20429a3a-dc5a-4539-9557-abbea1bf7616-c000.csv' into table datewise_aggregated_data
Loading data to table cab_booking_data.datewise_aggregated_data
OK
Time taken: 0.525 seconds
hive>
```

4. Verify the data in hive tables

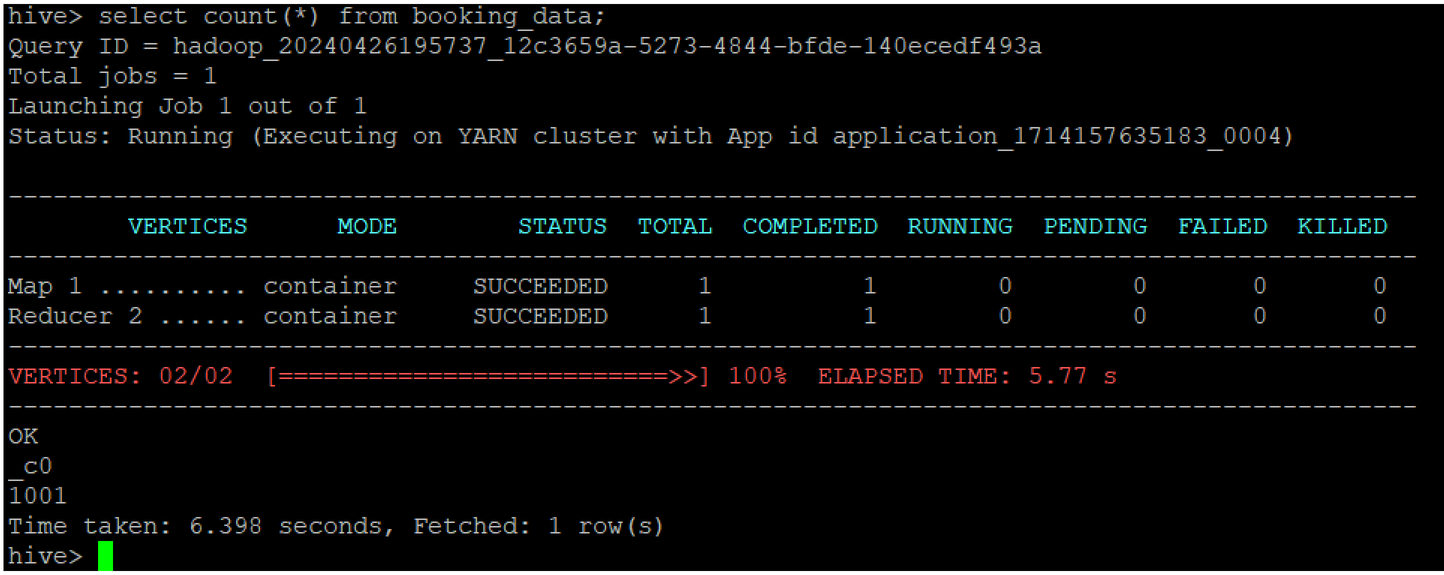
select count(*) from clickstream_data;



```
hive> select count(*) from clickstream_data;
Query ID = hadoop_20240426195557_ff2a8956-1a8f-40ab-812b-91700bbbbdf1
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_1714157635183_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: 6.02 s
----------------------------------------------------------------------------------------------
OK
_c0
2454
Time taken: 15.434 seconds, Fetched: 1 row(s)
hive>
```

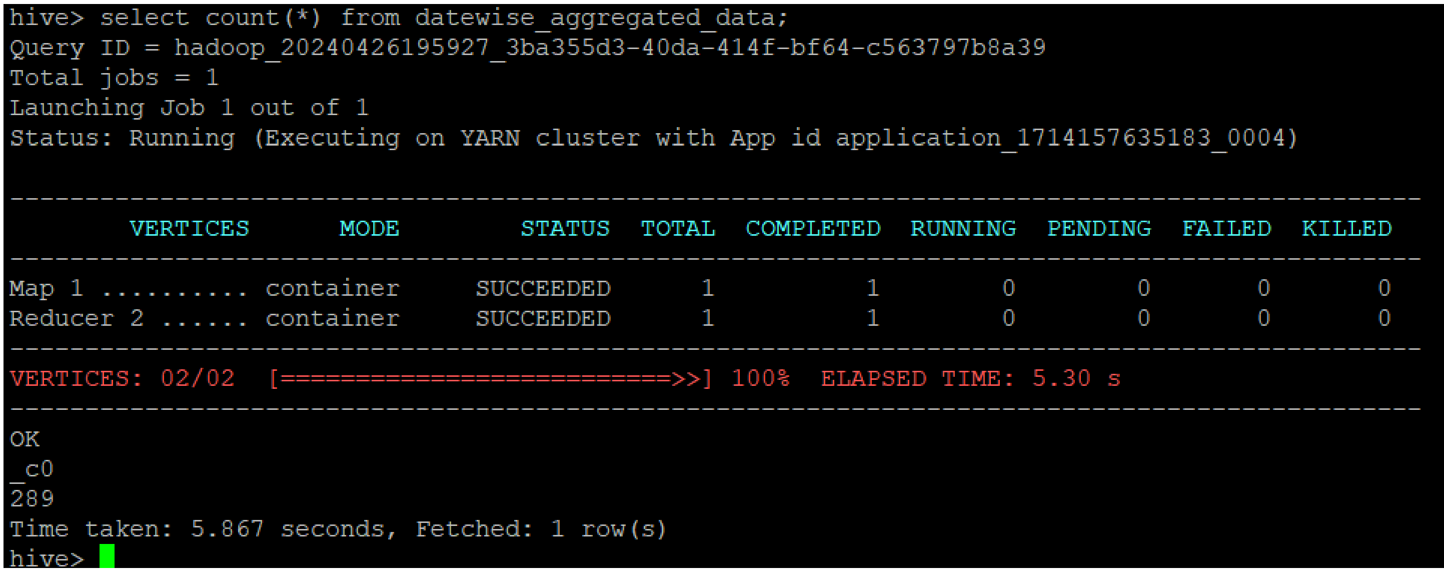
select count(*) from booking_data;



```
hive> select count(*) from booking_data;
Query ID = hadoop_20240426195737_12c3659a-5273-4844-bfde-140ecedf493a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1714157635183_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: 5.77 s
----------------------------------------------------------------------------------------------
OK
_c0
1001
Time taken: 6.398 seconds, Fetched: 1 row(s)
hive>
```

select count(*) from datewise_aggregated_data;



```
hive> select count(*) from datewise_aggregated_data;
Query ID = hadoop_20240426195927_3ba355d3-40da-414f-bf64-c563797b8a39
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1714157635183_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: 5.30 s
----------------------------------------------------------------------------------------------
OK
_c0
289
Time taken: 5.867 seconds, Fetched: 1 row(s)
hive>
```