Create Hive-Managed Tables

1. **Command to create database:**

```
create database if not exists cab_rides_data;
```

2. **Command to use the created database:**

```
use cab_rides_data;
```

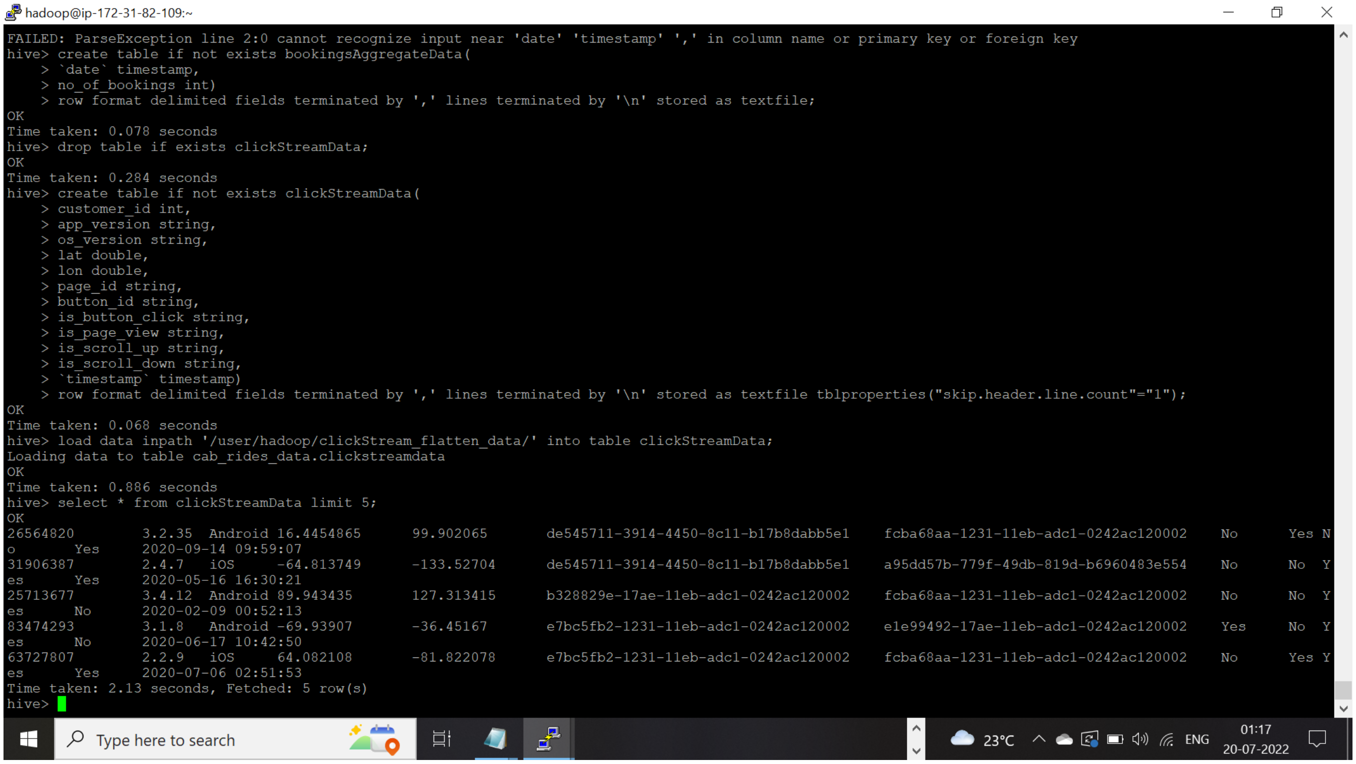
3. **Command to create clickStreamData table and load data from HDFS:**

```
create table if not exists clickStreamData(
customer_id int,
app_version string,
os_version string,
lat double,
lon double,
page_id string,
button_id string,
is_button_click string,
is_page_view string,
is_scroll_up string,
is_scroll_down string,
`timestamp` timestamp)
row format delimited fields terminated by ',' lines
terminated by '\n' stored as textfile
tblproperties("skip.header.line.count"="1");
```

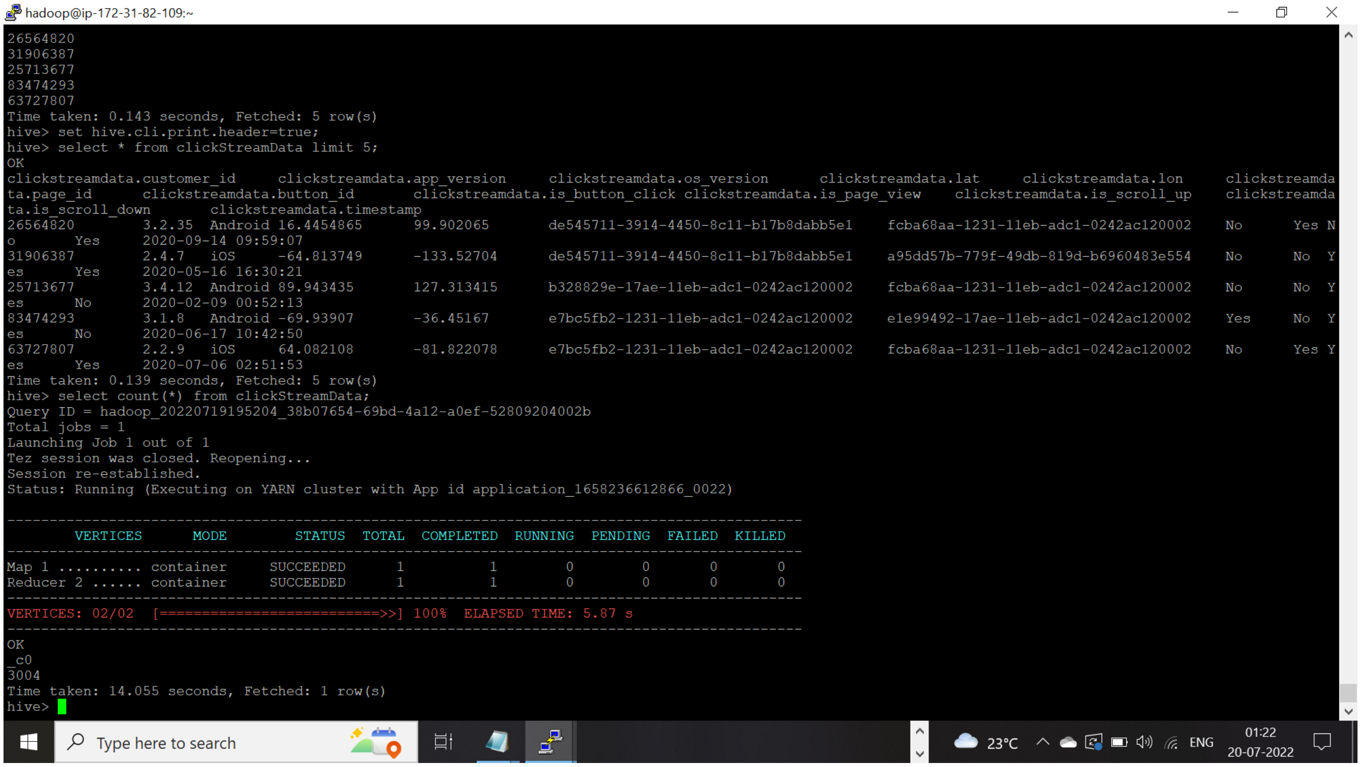
Command to load data from HDFS to clickStreamData table:

```
load data inpath '/user/hadoop/clickStream_flatten_data/' into
table clickStreamData;
```

Screenshot of the loaded data in clickStreamData table:



```
FAILED: ParseException line 2:0 cannot recognize input near 'date' 'timestamp' ',' in column name or primary key or foreign key
hive> create table if not exists bookingsAggregateData(
    > `date` timestamp,
    > no_of_bookings int)
    > row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.078 seconds
hive> drop table if exists clickStreamData;
OK
Time taken: 0.284 seconds
hive> create table if not exists clickStreamData(
    > customer_id int,
    > app_version string,
    > os_version string,
    > lat double,
    > lon double,
    > page_id string,
    > button_id string,
    > is_button_click string,
    > is_page_view string,
    > is_scroll_up string,
    > is_scroll_down string,
    > `timestamp` timestamp)
    > row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.068 seconds
hive> load data inpath '/user/hadoop/clickStream_flatten_data/' into table clickStreamData;
Loading data to table cab_rides_data.clickstreamdata
OK
Time taken: 0.886 seconds
hive> select * from clickStreamData limit 5;
OK
26564820        3.2.35  Android 16.4454865      99.902065       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes N
o       Yes     2020-09-14 09:59:07
31906387        2.4.7   iOS     -64.813749      -133.52704      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      No  Y
es      Yes     2020-05-16 16:30:21
25713677        3.4.12  Android 89.943435       127.313415      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No  Y
es      No      2020-02-09 00:52:13
83474293        3.1.8   Android -69.93907       -36.45167       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No  Y
es      No      2020-06-17 10:42:50
63727807        2.2.9   iOS     64.082108       -81.822078      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes Y
es      Yes     2020-07-06 02:51:53
Time taken: 2.13 seconds, Fetched: 5 row(s)
hive> 
```



```
26564820
31906387
25713677
83474293
63727807
Time taken: 0.143 seconds, Fetched: 5 row(s)
hive> set hive.cli.print.header=true;
hive> select * from clickStreamData limit 5;
OK
clickstreamdata.customer_id     clickstreamdata.app_version     clickstreamdata.os_version      clickstreamdata.lat     clickstreamdata.lon     clickstreamda
ta.page_id      clickstreamdata.button_id       clickstreamdata.is_button_click clickstreamdata.is_page_view    clickstreamdata.is_scroll_up    clickstreamda
ta.is_scroll_down       clickstreamdata.timestamp
26564820        3.2.35  Android 16.4454865      99.902065       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes N
o       Yes     2020-09-14 09:59:07
31906387        2.4.7   iOS     -64.813749      -133.52704      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      No  Y
es      Yes     2020-05-16 16:30:21
25713677        3.4.12  Android 89.943435       127.313415      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No  Y
es      No      2020-02-09 00:52:13
83474293        3.1.8   Android -69.93907       -36.45167       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No  Y
es      No      2020-06-17 10:42:50
63727807        2.2.9   iOS     64.082108       -81.822078      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes Y
es      Yes     2020-07-06 02:51:53
Time taken: 0.139 seconds, Fetched: 5 row(s)
hive> select count(*) from clickStreamData;
Query ID = hadoop_20220719195204_38b07654-69bd-4a12-a0ef-52809204002b
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_1658236612866_0022)

----------------------------------------------------------------------------------------------
        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: 5.87 s
----------------------------------------------------------------------------------------------
OK
_c0
3004
Time taken: 14.055 seconds, Fetched: 1 row(s)
hive> 
```

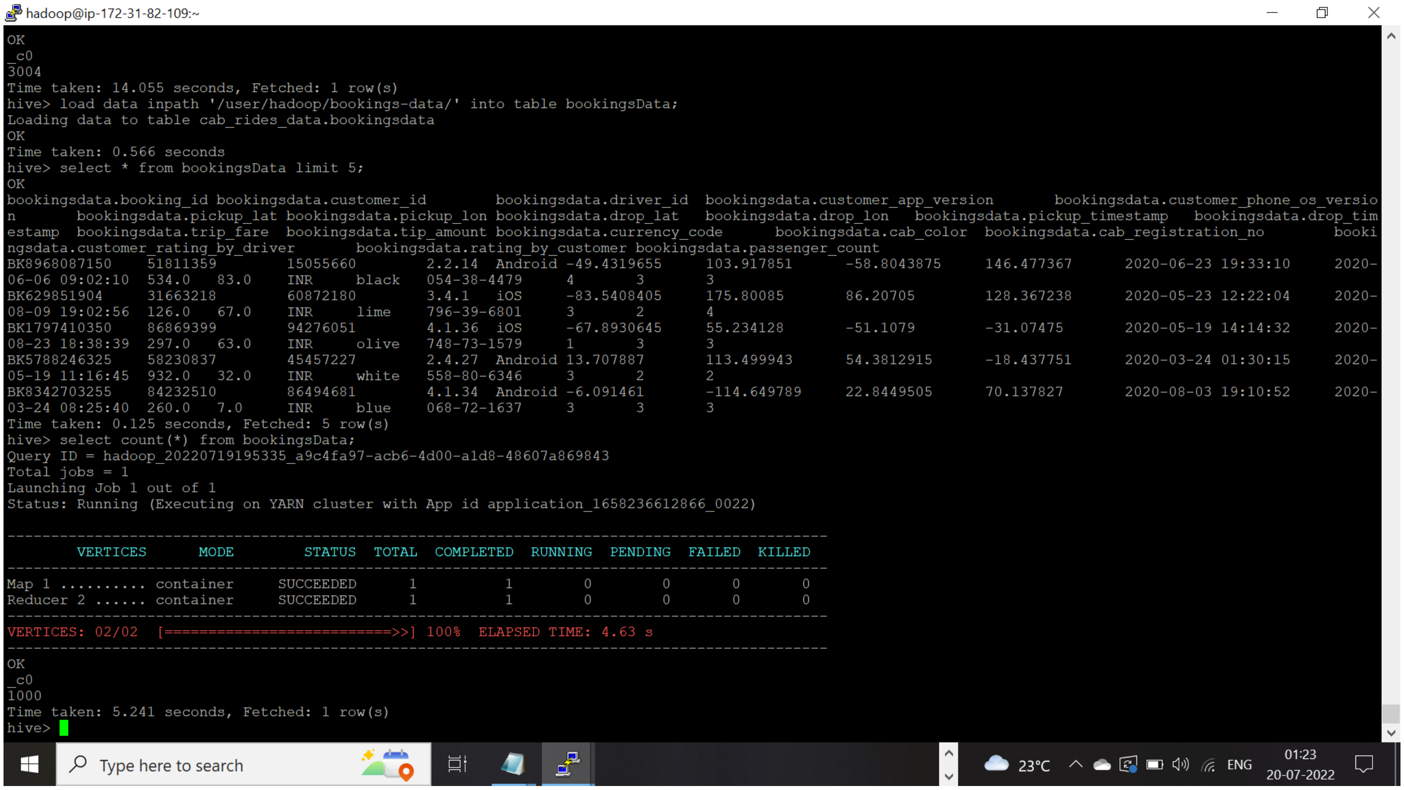
4. **Command to create bookingsData table and load data from HDFS:**

```
create table if not exists bookingsData(
booking_id string,
customer_id int,
driver_id int,
customer_app_version string,
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 double,
tip_amount double,
currency_code string,
cab_color string,
cab_registration_no string,
customer_rating_by_driver int,
rating_by_customer int,
passenger_count int)
row format delimited fields terminated by ',' lines terminated
by '\n' stored as textfile;
```

Command to load the data into bookingsData table:

```
load data inpath '/user/hadoop/bookings-
data/' into table bookingsData;
```

Screenshot of the loaded data into bookingsData table:



hadoop@ip-172-31-82-109:~

```
OK
_c0
3004
Time taken: 14.055 seconds, Fetched: 1 row(s)
hive> load data inpath '/user/hadoop/bookings-data/' into table bookingsData;
Loading data to table cab_rides_data.bookingsdata
OK
Time taken: 0.566 seconds
hive> select * from bookingsData limit 5;
OK
bookingsdata.booking_id bookingsdata.customer_id        bookingsdata.driver_id  bookingsdata.customer_app_version       bookingsdata.customer_phone_os_versio
n       bookingsdata.pickup_lat bookingsdata.pickup_lon bookingsdata.drop_lat   bookingsdata.drop_lon   bookingsdata.pickup_timestamp   bookingsdata.drop_tim
estamp  bookingsdata.trip_fare  bookingsdata.tip_amount bookingsdata.currency_code      bookingsdata.cab_color  bookingsdata.cab_registration_no        booki
ngsdata.customer_rating_by_driver       bookingsdata.rating_by_customer bookingsdata.passenger_count
BK8968087150    51811359        15055660        2.2.14  Android -49.4319655     103.917851      -58.8043875     146.477367      2020-06-23 19:33:10     2020-
06-06 09:02:10  534.0   83.0    INR     black   054-38-4479     4       3       3
BK629851904     31663218        60872180        3.4.1   iOS     -83.5408405     175.80085       86.20705        128.367238      2020-05-23 12:22:04     2020-
08-09 19:02:56  126.0   67.0    INR     lime    796-39-6801     3       2       4
BK1797410350    86869399        94276051        4.1.36  iOS     -67.8930645     55.234128       -51.1079        -31.07475       2020-05-19 14:14:32     2020-
08-23 18:38:39  297.0   63.0    INR     olive   748-73-1579     1       3       3
BK5788246325    58230837        45457227        2.4.27  Android 13.707887       113.499943      54.3812915      -18.437751      2020-03-24 01:30:15     2020-
05-19 11:16:45  932.0   32.0    INR     white   558-80-6346     3       2       2
BK8342703255    84232510        86494681        4.1.34  Android -6.091461       -114.649789     22.8449505      70.137827       2020-08-03 19:10:52     2020-
03-24 08:25:40  260.0   7.0     INR     blue    068-72-1637     3       3       3
Time taken: 0.125 seconds, Fetched: 5 row(s)
hive> select count(*) from bookingsData;
Query ID = hadoop_20220719195335_a9c4fa97-acb6-4d00-a1d8-48607a869843
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658236612866_0022)

----------------------------------------------------------------------------------------------
        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: 4.63 s
----------------------------------------------------------------------------------------------
OK
_c0
1000
Time taken: 5.241 seconds, Fetched: 1 row(s)
hive>
```

5. Command to create bookingsAggregateData and load data from HDFS:

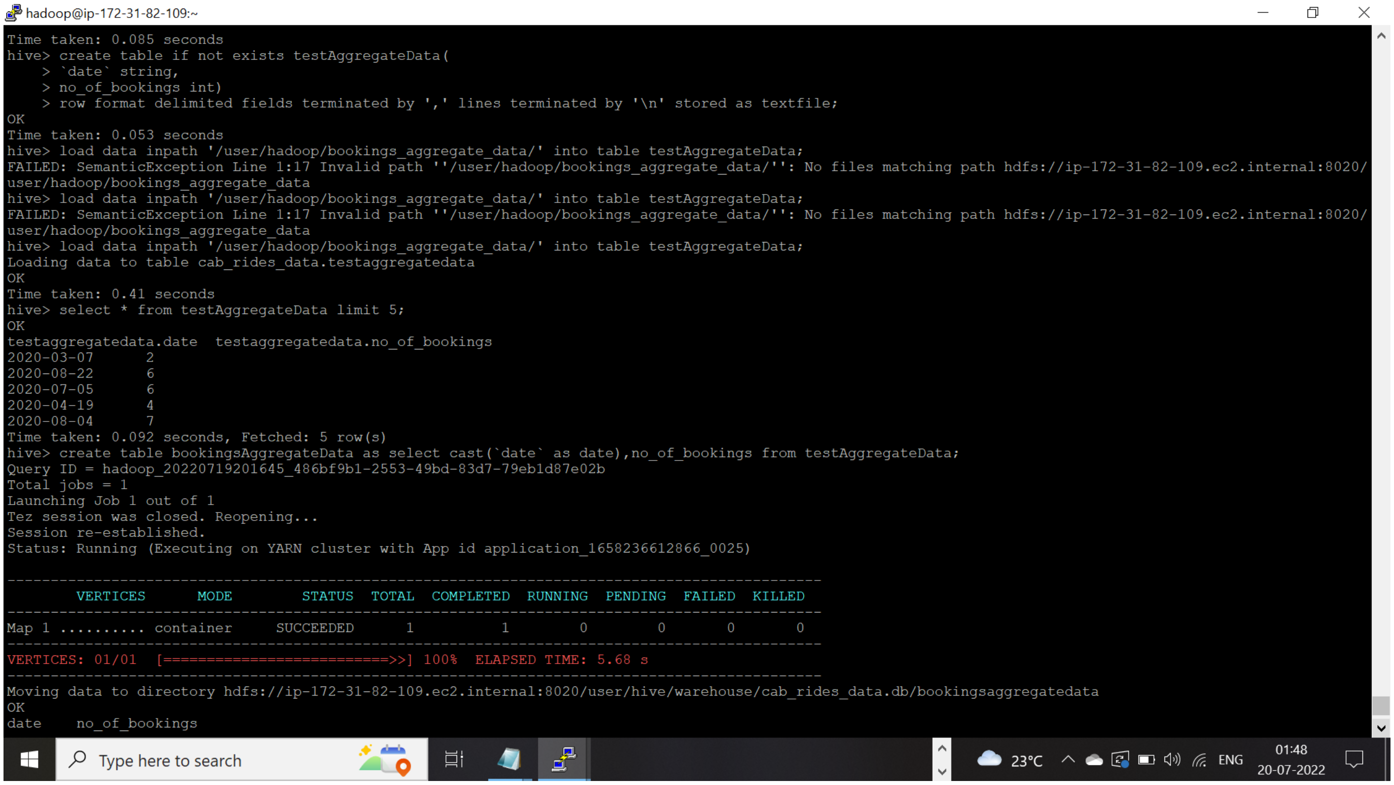
I. In bookings_aggregate_data, we have date column with date as datatype. Hence to cast the column in date type, I have created a temporary table named as testAggregateData and loaded with data from HDFS:

Command to create testAggregateData table:

```
create table if not exists testAggregateData(
`date` string,
no_of_bookings int)
row format delimited fields terminated by ','
lines terminated by '\n' stored as textfile;
```

Command to load data from HDFS to testAggregateData:

```
load data inpath '/user/hadoop/bookings_aggregate_data/'
into table testAggregateData;
```



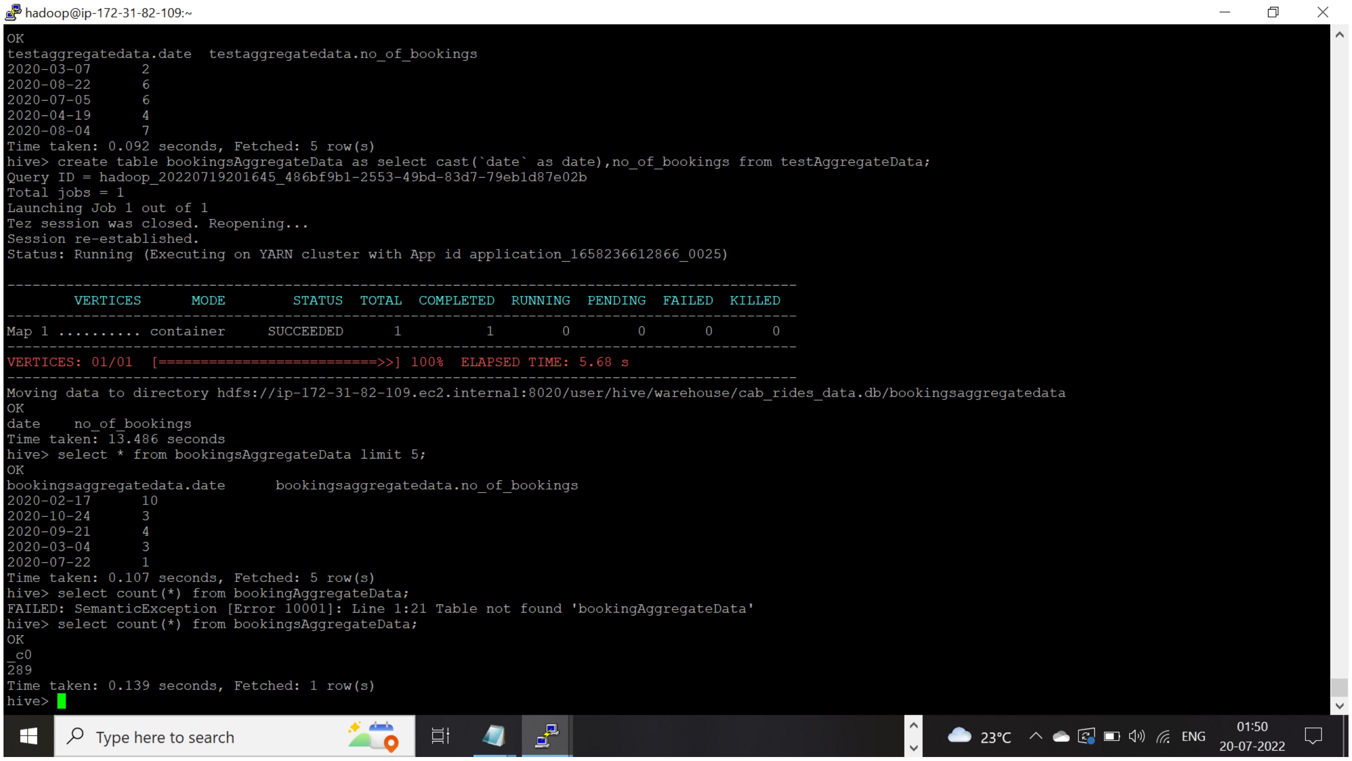
```
Time taken: 0.085 seconds
hive> create table if not exists testAggregateData(
    > `date` string,
    > no_of_bookings int)
    > row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.053 seconds
hive> load data inpath '/user/hadoop/bookings_aggregate_data/' into table testAggregateData;
FAILED: SemanticException Line 1:17 Invalid path ''/user/hadoop/bookings_aggregate_data/'': No files matching path hdfs://ip-172-31-82-109.ec2.internal:8020/
user/hadoop/bookings_aggregate_data
hive> load data inpath '/user/hadoop/bookings_aggregate_data/' into table testAggregateData;
FAILED: SemanticException Line 1:17 Invalid path ''/user/hadoop/bookings_aggregate_data/'': No files matching path hdfs://ip-172-31-82-109.ec2.internal:8020/
user/hadoop/bookings_aggregate_data
hive> load data inpath '/user/hadoop/bookings_aggregate_data/' into table testAggregateData;
Loading data to table cab_rides_data.testaggregatedata
OK
Time taken: 0.41 seconds
hive> select * from testAggregateData limit 5;
OK
testaggregatedata.date  testaggregatedata.no_of_bookings
2020-03-07      2
2020-08-22      6
2020-07-05      6
2020-04-19      4
2020-08-04      7
Time taken: 0.092 seconds, Fetched: 5 row(s)
hive> create table bookingsAggregateData as select cast(`date` as date),no_of_bookings from testAggregateData;
Query ID = hadoop_20220719201645_486bf9b1-2553-49bd-83d7-79eb1d87e02b
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_1658236612866_0025)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.68 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-82-109.ec2.internal:8020/user/hive/warehouse/cab_rides_data.db/bookingsaggregatedata
OK
date    no_of_bookings
```

II. In the next step, I have created a table named as bookingsAggregateData to cast the column date into date datatype.

Command to create bookingsAggregateData:

```
create table bookingsAggregateData as select
cast(`date` as date),no_of_bookings from
testAggregateData;
```



```
OK
testaggregatedata.date  testaggregatedata.no_of_bookings
2020-03-07      2
2020-08-22      6
2020-07-05      6
2020-04-19      4
2020-08-04      7
Time taken: 0.092 seconds, Fetched: 5 row(s)
hive> create table bookingsAggregateData as select cast(`date` as date),no_of_bookings from testAggregateData;
Query ID = hadoop_20220719201645_486bf9b1-2553-49bd-83d7-79eb1d87e02b
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_1658236612866_0025)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.68 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-82-109.ec2.internal:8020/user/hive/warehouse/cab_rides_data.db/bookingsaggregatedata
OK
date    no_of_bookings
Time taken: 13.486 seconds
hive> select * from bookingsAggregateData limit 5;
OK
bookingsaggregatedata.date      bookingsaggregatedata.no_of_bookings
2020-02-17      10
2020-10-24      3
2020-09-21      4
2020-03-04      3
2020-07-22      1
Time taken: 0.107 seconds, Fetched: 5 row(s)
hive> select count(*) from bookingAggregateData;
FAILED: SemanticException [Error 10001]: Line 1:21 Table not found 'bookingAggregateData'
hive> select count(*) from bookingsAggregateData;
OK
_c0
289
Time taken: 0.139 seconds, Fetched: 1 row(s)
hive>
```