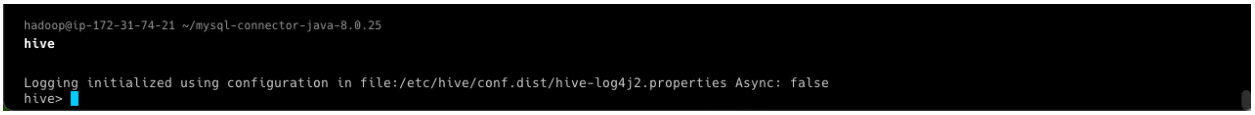
Create Hive-Managed Tables

create database and tables

1. Connect to hive instance

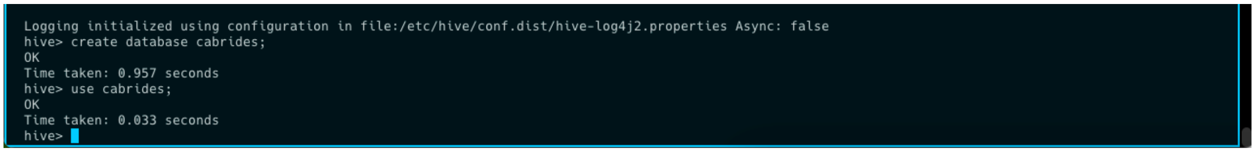
```
hive
```



```
hadoop@ip-172-31-74-21 ~/mysql-connector-java-8.0.25
hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

2. Create database and aggregated tables in it

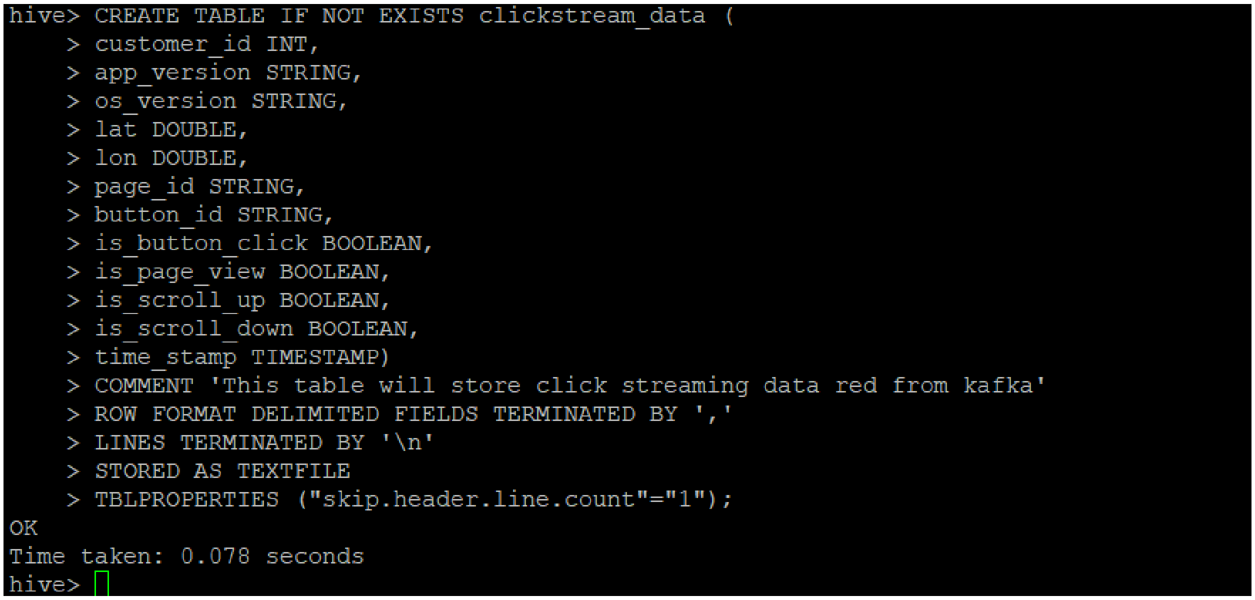
```
create database cabrides;
use cabrides;
```



```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database cabrides;
OK
Time taken: 0.957 seconds
hive> use cabrides;
OK
Time taken: 0.033 seconds
hive>
```

3. Create clickstream_data table

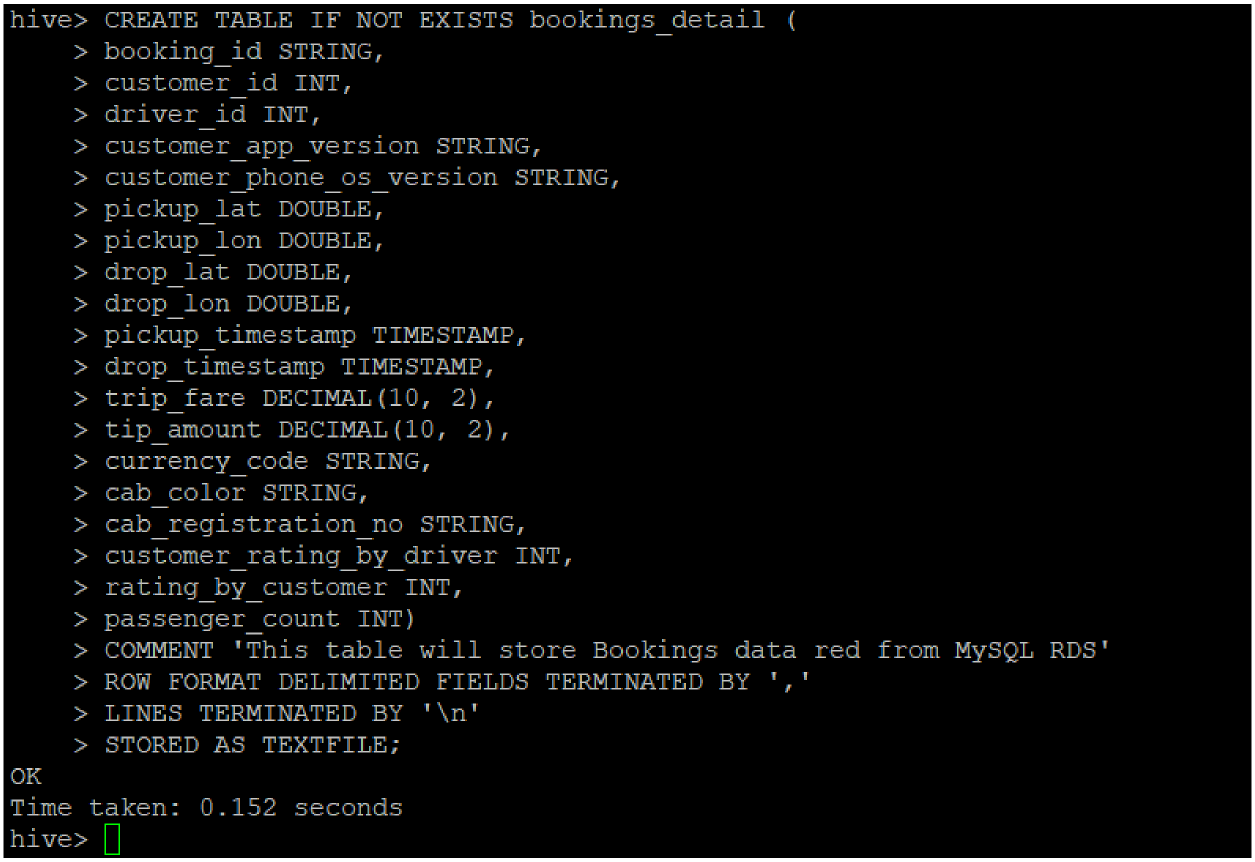
```
CREATE TABLE IF NOT EXISTS clickstream_data (
customer_id INT,
app_version STRING,
os_version STRING,
lat DOUBLE,
lon DOUBLE,
page_id STRING,
button_id STRING,
is_button_click BOOLEAN,
is_page_view BOOLEAN,
is_scroll_up BOOLEAN,
is_scroll_down BOOLEAN,
time_stamp TIMESTAMP)
COMMENT 'This table will store click streaming data red from kafka'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```



```
hive> CREATE TABLE IF NOT EXISTS clickstream_data (
    > customer_id INT,
    > app_version STRING,
    > os_version STRING,
    > lat DOUBLE,
    > lon DOUBLE,
    > page_id STRING,
    > button_id STRING,
    > is_button_click BOOLEAN,
    > is_page_view BOOLEAN,
    > is_scroll_up BOOLEAN,
    > is_scroll_down BOOLEAN,
    > time_stamp TIMESTAMP)
    > COMMENT 'This table will store click streaming data red from kafka'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.078 seconds
hive> 
```

4. Create bookings_detail table

```
CREATE TABLE IF NOT EXISTS bookings_detail (
    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 DECIMAL(10, 2),
    tip_amount DECIMAL(10, 2),
    currency_code STRING,
    cab_color STRING,
    cab_registration_no STRING,
    customer_rating_by_driver INT,
    rating_by_customer INT,
    passenger_count INT
)
COMMENT 'This table will store Bookings'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```

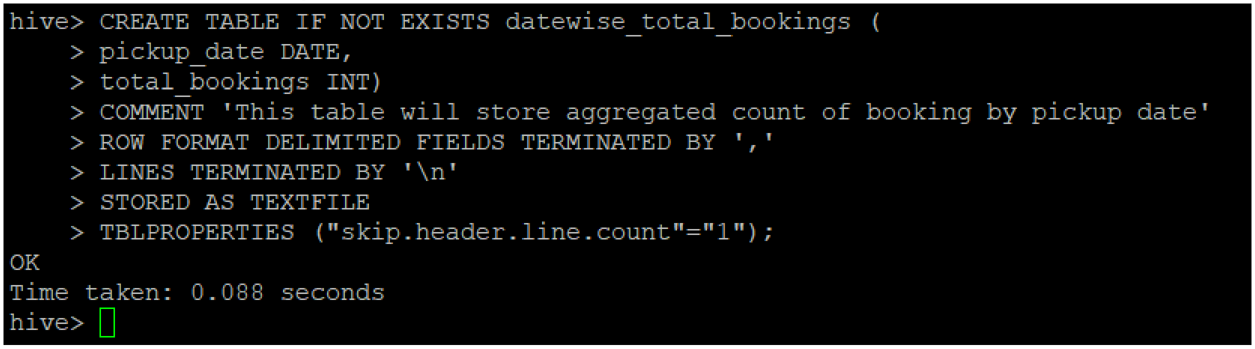


```
hive> CREATE TABLE IF NOT EXISTS bookings_detail (
    > booking_id STRING,
    > customer_id INT,
    > driver_id INT,
    > customer_app_version STRING,
    > customer_phone_os_version STRING,
    > pickup_lat DOUBLE,
    > pickup_lon DOUBLE,
    > drop_lat DOUBLE,
    > drop_lon DOUBLE,
    > pickup_timestamp TIMESTAMP,
    > drop_timestamp TIMESTAMP,
    > trip_fare DECIMAL(10, 2),
    > tip_amount DECIMAL(10, 2),
    > currency_code STRING,
    > cab_color STRING,
    > cab_registration_no STRING,
    > customer_rating_by_driver INT,
    > rating_by_customer INT,
    > passenger_count INT)
    > COMMENT 'This table will store Bookings data red from MySQL RDS'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE;
OK
Time taken: 0.152 seconds
hive>
```

5. create datewise_total_bookings table

```
CREATE TABLE IF NOT EXISTS datewise_total_bookings (

pickup_date DATE,

total_bookings INT)

COMMENT 'This table will store aggregated count of booking by pickup
Date'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED AS TEXTFILE
```

```
TBLPROPERTIES ("skip.header.line.count"="1");
```

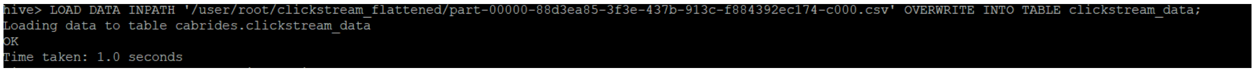


```
hive> CREATE TABLE IF NOT EXISTS datewise_total_bookings (
    > pickup_date DATE,
    > total_bookings INT)
    > COMMENT 'This table will store aggregated count of booking by pickup date'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.088 seconds
hive>
```

Load data in hive tables

1. Load click stream data into clickstream_data table;

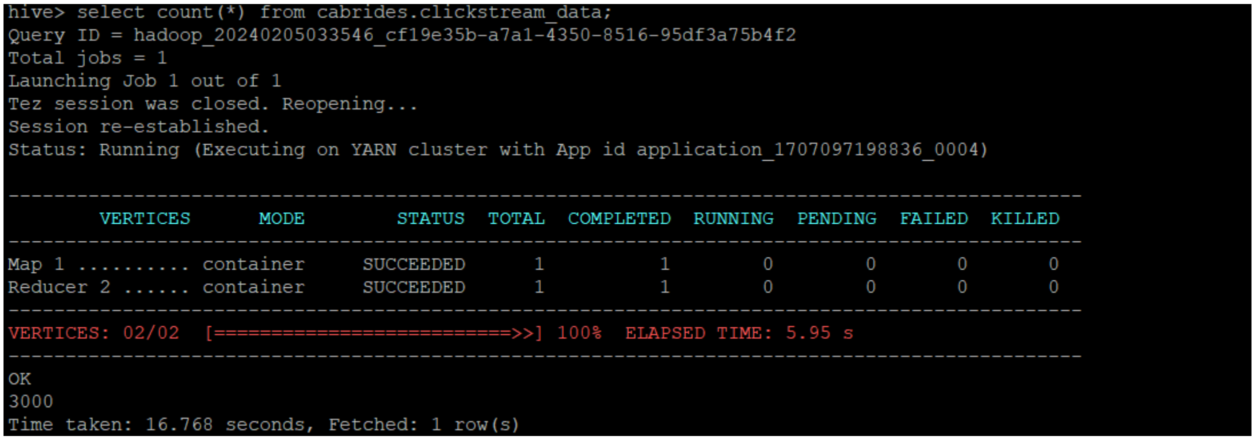
```
LOAD DATA INPATH '/user/root/clickstream_flattened/part-00000-88d3ea85-3f3e-437b-913c-f884392ec174-c000.csv' OVERWRITE INTO TABLE clickstream_data
```



```
hive> LOAD DATA INPATH '/user/root/clickstream_flattened/part-00000-88d3ea85-3f3e-437b-913c-f884392ec174-c000.csv' OVERWRITE INTO TABLE clickstream_data;
Loading data to table cabrides.clickstream_data
OK
Time taken: 1.0 seconds
```

2. verify records count in clickstream_data table

```
select count(*) from clickstream_data;
```

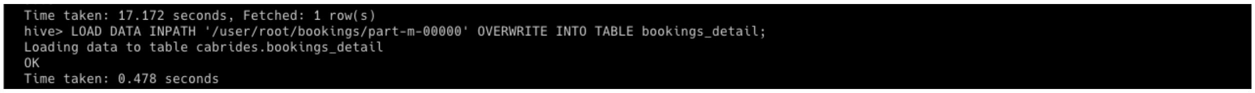


```
hive> select count(*) from cabrides.clickstream_data;
Query ID = hadoop_20240205033546_cf19e35b-a7a1-4350-8516-95df3a75b4f2
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1707097198836_0004)

----------------------------------------------------------------------------------------------
        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: 5.95 s
----------------------------------------------------------------------------------------------
OK
3000
Time taken: 16.768 seconds, Fetched: 1 row(s)
```

Around 3000 records are present, 3 records removed because of customer id is empty in spark_local_flatten.py code.

3. Load booking file into bookings_detail table

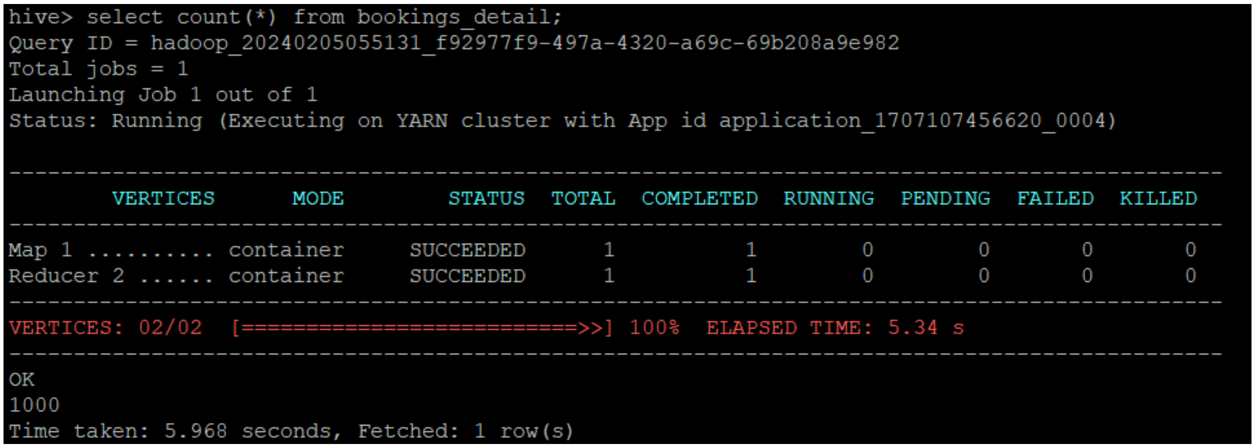
```
LOAD DATA INPATH '/user/root/bookings/part-m-00000' OVERWRITE INTO TABLE
bookings_detail;
```



```
Time taken: 17.172 seconds, Fetched: 1 row(s)
hive> LOAD DATA INPATH '/user/root/bookings/part-m-00000' OVERWRITE INTO TABLE bookings_detail;
Loading data to table cabrides.bookings_detail
OK
Time taken: 0.478 seconds
```

4. verify records in bookings_detail table

```
select count(*) from bookings_detail;
```

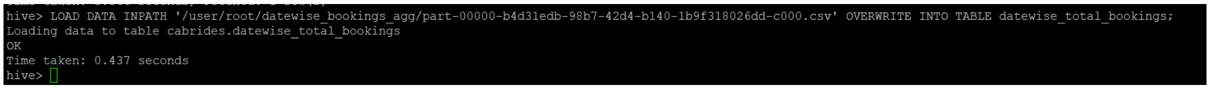


```
hive> select count(*) from bookings_detail;
Query ID = hadoop_20240205055131_f92977f9-497a-4320-a69c-69b208a9e982
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707107456620_0004)

----------------------------------------------------------------------------------------------
        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: 5.34 s
----------------------------------------------------------------------------------------------
OK
1000
Time taken: 5.968 seconds, Fetched: 1 row(s)
```

Around 1000 records are present

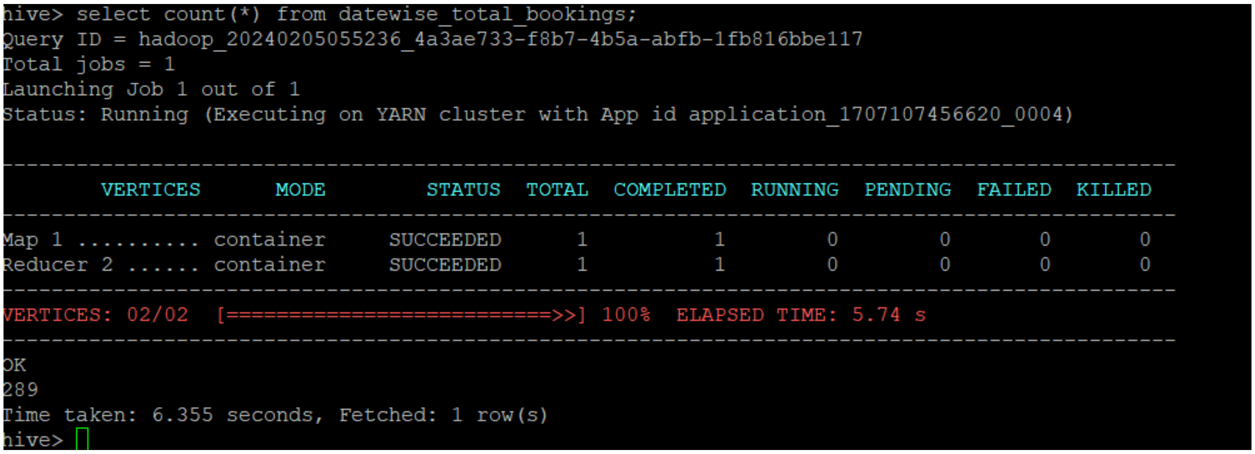
5. Loading data into datewise_totoal_bookings and verifying records count

```
LOAD DATA INPATH '/user/root/datewise_bookings_agg/part-00000-d43d1e9b-
987b-42d4-b140-1b9f318026dd-c000.csv' OVERWRITE INTO TABLE
datewise_total_bookings;
```



```
hive> LOAD DATA INPATH '/user/root/datewise_bookings_agg/part-00000-b4d31edb-98b7-42d4-b140-1b9f318026dd-c000.csv' OVERWRITE INTO TABLE datewise_total_bookings;
Loading data to table cabrides.datewise_total_bookings
OK
Time taken: 0.437 seconds
hive> 
```

```
select count(*) from datewise_total_bookings;
```



```
hive> select count(*) from datewise_total_bookings;
Query ID = hadoop_20240205055236_4a3ae733-f8b7-4b5a-abfb-1fb816bbe117
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1707107456620_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: 5.74 s
----------------------------------------------------------------------------------------------
OK
289
Time taken: 6.355 seconds, Fetched: 1 row(s)
hive>
```

Around 289 records are present