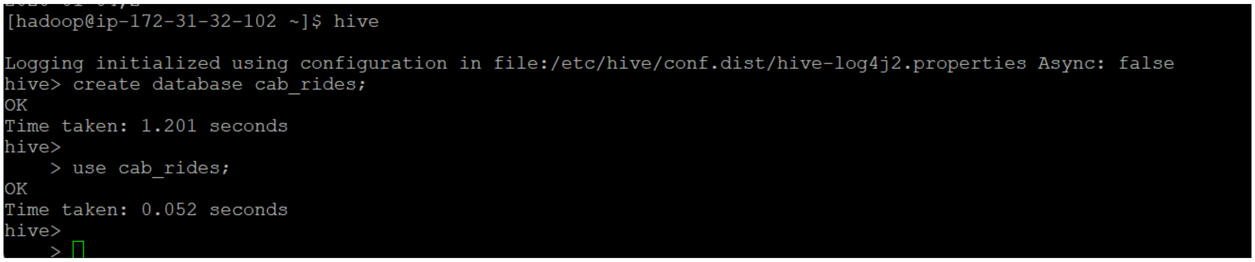
Create Hive-Managed Tables

Command to create the Hive tables

1. Connecting to Hive instance & creating database:

create database cab_rides;
use cab_rides;

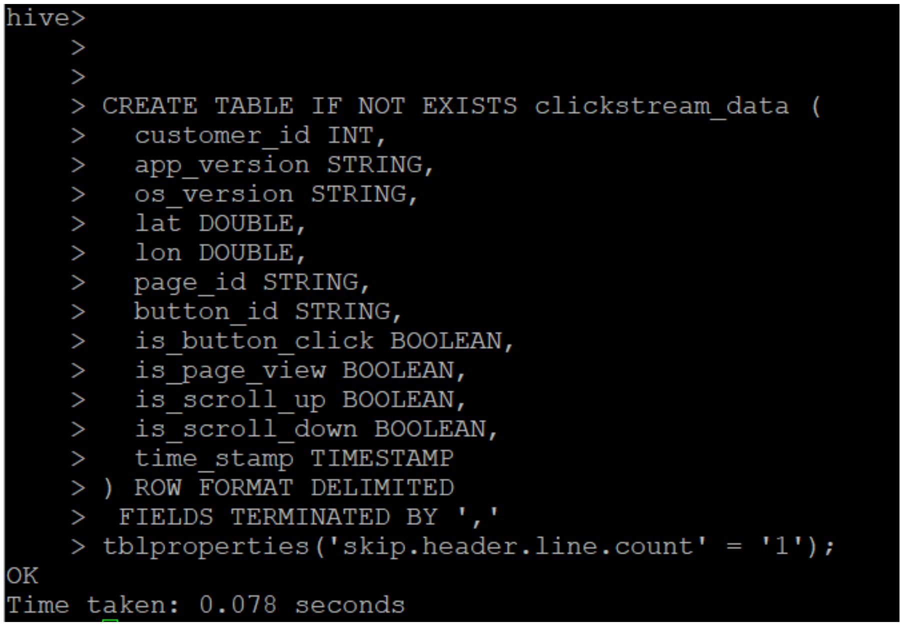


```
[hadoop@ip-172-31-32-102 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database cab_rides;
OK
Time taken: 1.201 seconds
hive>
    > use cab_rides;
OK
Time taken: 0.052 seconds
hive>
    > 
```

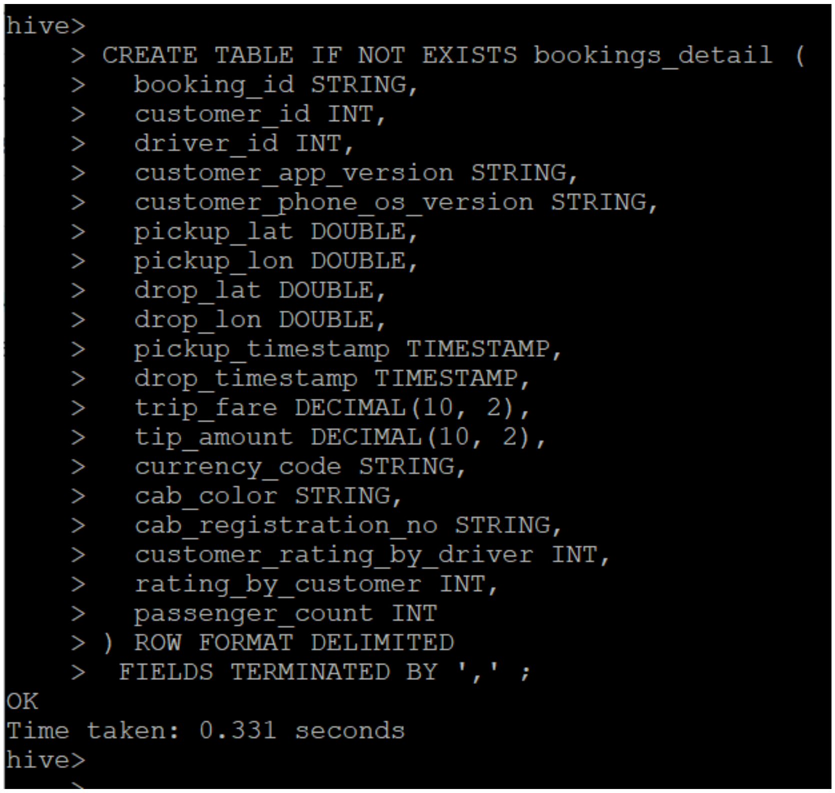
2. Creating tables in Hive

```
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
) ROW FORMAT DELIMITED
 FIELDS TERMINATED BY ','                    //To load the file with comma as delimiter
tblproperties('skip.header.line.count' = '1');         //To skip the header present in the file
```



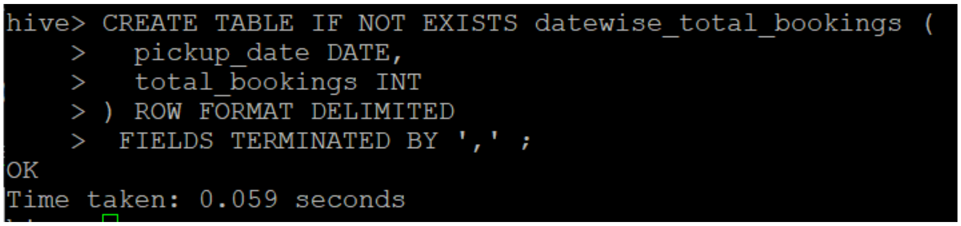
```
hive>
    >
    >
    > CREATE TABLE IF NOT EXISTS clickstream_data (
    >   customer_id INT,
    >   app_version STRING,
    >   os_version STRING,
    >   lat DOUBLE,
    >   lon DOUBLE,
    >   page_id STRING,
    >   button_id STRING,
    >   is_button_click BOOLEAN,
    >   is_page_view BOOLEAN,
    >   is_scroll_up BOOLEAN,
    >   is_scroll_down BOOLEAN,
    >   time_stamp TIMESTAMP
    > ) ROW FORMAT DELIMITED
    >  FIELDS TERMINATED BY ','
    > tblproperties('skip.header.line.count' = '1');
OK
Time taken: 0.078 seconds
```

```
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
) ROW FORMAT DELIMITED
 FIELDS TERMINATED BY ',' ;                    //To load the file with comma as delimiter
```



```
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
    > ) ROW FORMAT DELIMITED
    >  FIELDS TERMINATED BY ',' ;
OK
Time taken: 0.331 seconds
hive>
```

```
CREATE TABLE IF NOT EXISTS datewise_total_bookings (
  pickup_date DATE,
  total_bookings INT
) ROW FORMAT DELIMITED
 FIELDS TERMINATED BY ','                         //To load the file with comma as delimiter
tblproperties('skip.header.line.count' = '1');     //To skip the header present in the file
```

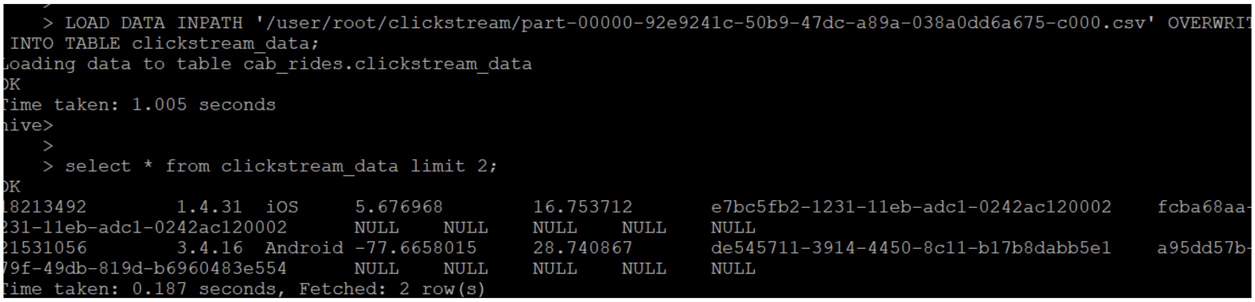


```
hive> CREATE TABLE IF NOT EXISTS datewise_total_bookings (
    >   pickup_date DATE,
    >   total_bookings INT
    > ) ROW FORMAT DELIMITED
    >  FIELDS TERMINATED BY ',' ;
OK
Time taken: 0.059 seconds
```

3. **Loading the Clickstream data into Hive tables using files in HDFS & validating the contents of the tables created**

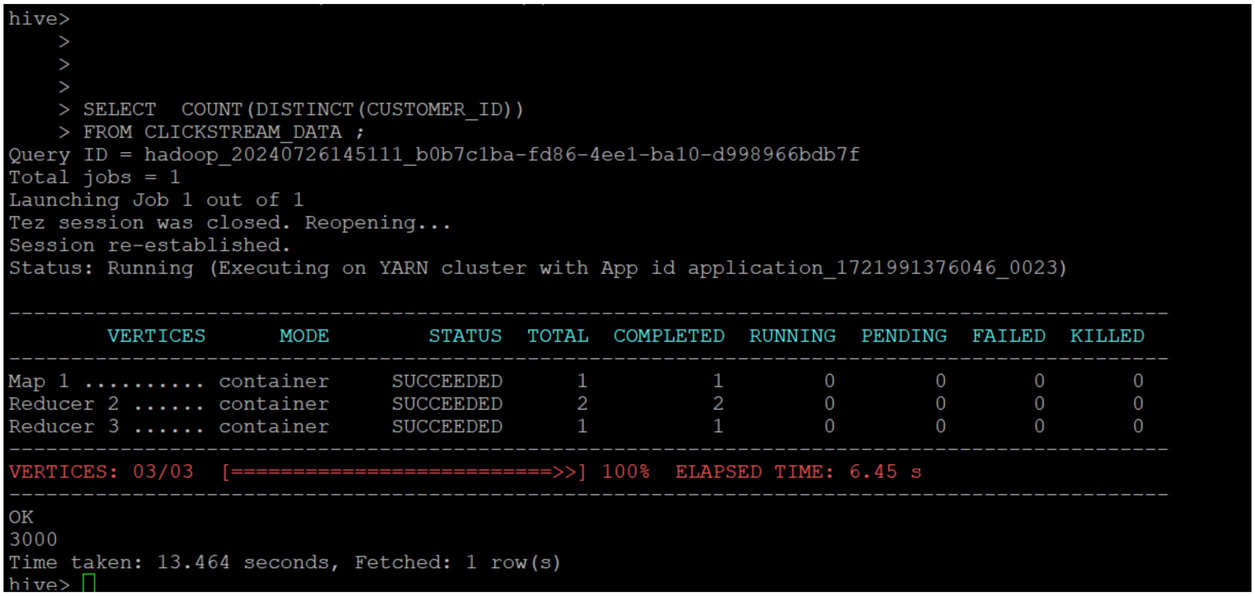
```
LOAD DATA INPATH '/user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv' OVERWRITE INTO TABLE clickstream_data;

Select * from clickstream_data limit 2;
```



```
    > LOAD DATA INPATH '/user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv' OVERWRIT
 INTO TABLE clickstream_data;
Loading data to table cab_rides.clickstream_data
OK
Time taken: 1.005 seconds
hive>
    >
    > select * from clickstream_data limit 2;
OK
18213492        1.4.31  iOS     5.676968        16.753712       e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-
231-11eb-adc1-0242ac120002      NULL    NULL    NULL    NULL    NULL
21531056        3.4.16  Android -77.6658015     28.740867       de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-
79f-49db-819d-b6960483e554      NULL    NULL    NULL    NULL    NULL
Time taken: 0.187 seconds, Fetched: 2 row(s)
```

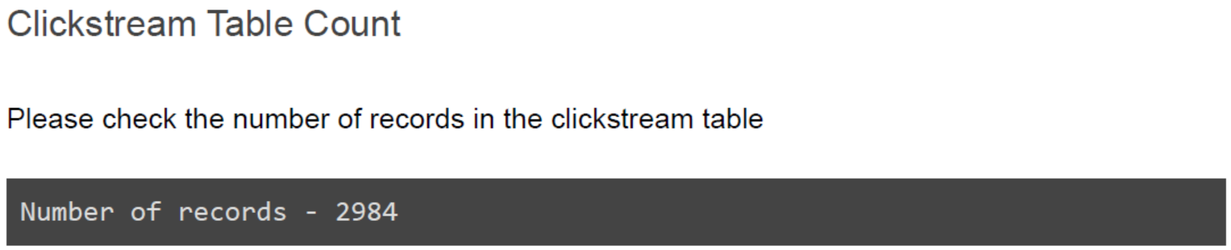
SELECT COUNT(DISTINCT(CUSTOMER_ID)) FROM CLICKSTREAM_DATA ;



```
hive>
    >
    >
    >
    > SELECT  COUNT(DISTINCT(CUSTOMER_ID))
    > FROM CLICKSTREAM_DATA ;
Query ID = hadoop_20240726145111_b0b7c1ba-fd86-4ee1-ba10-d998966bdb7f
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_1721991376046_0023)

----------------------------------------------------------------------------------------------
        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: 6.45 s
----------------------------------------------------------------------------------------------
OK
3000
Time taken: 13.464 seconds, Fetched: 1 row(s)
hive>
```

*** No of records ingested is having slight difference of 16 records with the validation documents due to live streaming data.*



Clickstream Table Count

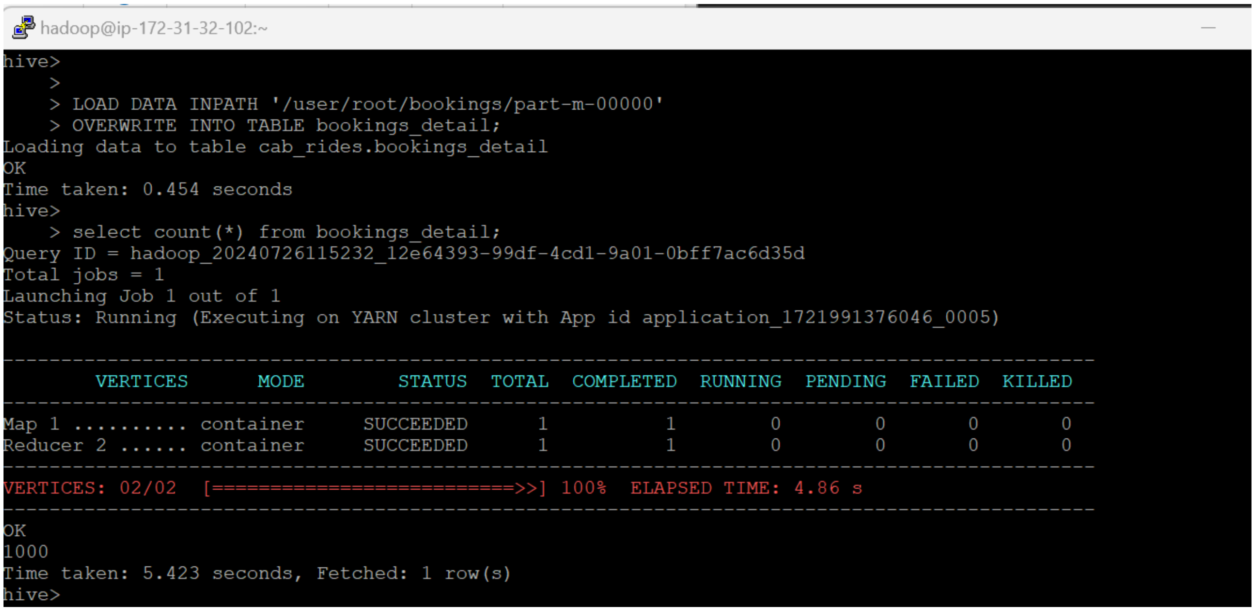
Please check the number of records in the clickstream table

```
Number of records - 2984
```

4. **Loading the Bookings data into Hive tables using files in HDFS & validating the contents of the tables created**

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

select count(*) from bookings_detail;



hadoop@ip-172-31-32-102:~

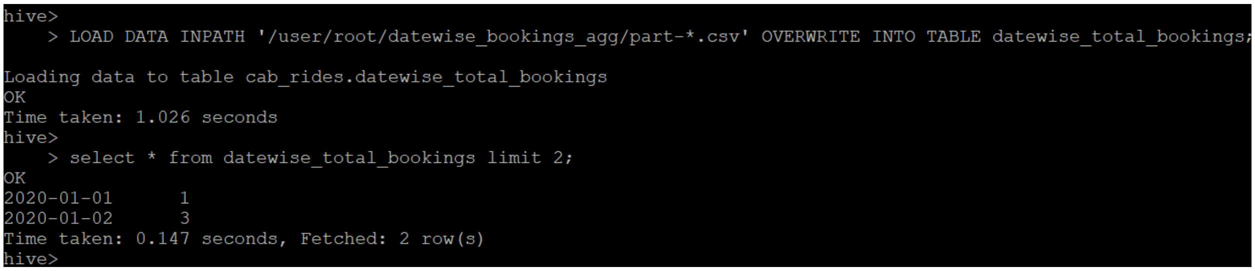
```
hive>
    >
    > LOAD DATA INPATH '/user/root/bookings/part-m-00000'
    > OVERWRITE INTO TABLE bookings_detail;
Loading data to table cab_rides.bookings_detail
OK
Time taken: 0.454 seconds
hive>
    > select count(*) from bookings_detail;
Query ID = hadoop_20240726115232_12e64393-99df-4cd1-9a01-0bff7ac6d35d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0005)

----------------------------------------------------------------------------------------------
        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: 4.86 s
----------------------------------------------------------------------------------------------
OK
1000
Time taken: 5.423 seconds, Fetched: 1 row(s)
hive>
```

*** No of records ingested is matching is matching with the validation documents*

5. Loading the Date-wise total bookings data into Hive tables using files in HDFS & validating the contents of the tables created

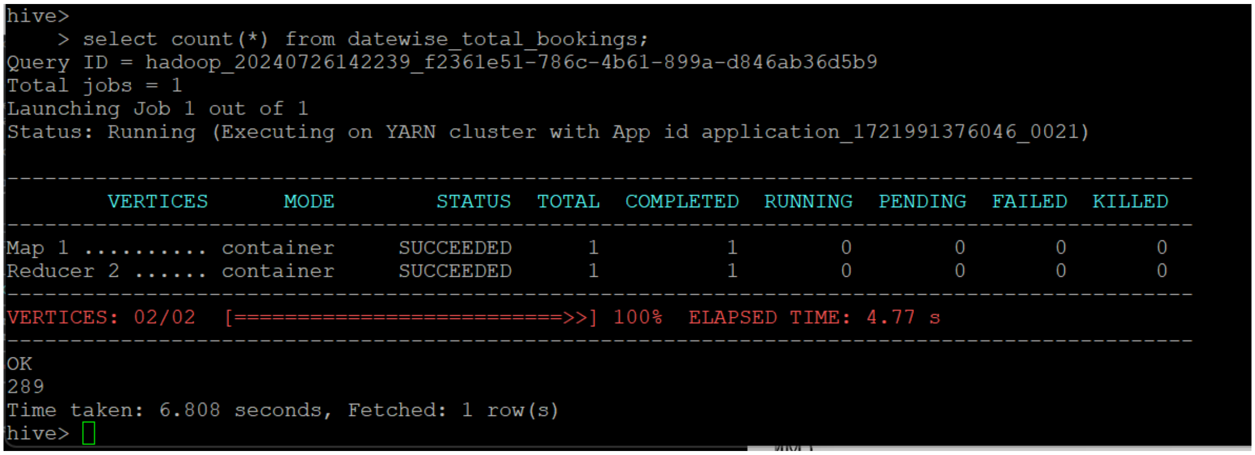
LOAD DATA INPATH '/user/root/datewise_bookings_agg/part-*.csv' OVERWRITE INTO TABLE datewise_total_bookings;



```
hive>
    > LOAD DATA INPATH '/user/root/datewise_bookings_agg/part-*.csv' OVERWRITE INTO TABLE datewise_total_bookings;

Loading data to table cab_rides.datewise_total_bookings
OK
Time taken: 1.026 seconds
hive>
    > select * from datewise_total_bookings limit 2;
OK
2020-01-01      1
2020-01-02      3
Time taken: 0.147 seconds, Fetched: 2 row(s)
hive>
```

select count(*) from datewise_total_bookings;



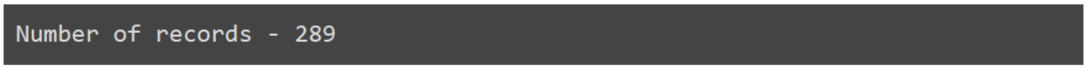
```
hive>
    > select count(*) from datewise_total_bookings;
Query ID = hadoop_20240726142239_f2361e51-786c-4b61-899a-d846ab36d5b9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

----------------------------------------------------------------------------------------------
        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: 4.77 s
----------------------------------------------------------------------------------------------
OK
289
Time taken: 6.808 seconds, Fetched: 1 row(s)
hive>
```

*** No of records ingested is matching is matching with the validation documents*

Bookings Aggregates Table Count

Please check the number of records in the bookings aggregates table



```
Number of records - 289
```