Logic For First Submission

<Properly explain the code, list the steps to run the code provided by you and attach screenshots of code execution>

Note: Be as descriptive as possible.

Fetch_click_stream_data.py

This file is used to read the kafka stream and store the read data into hdfs file system. First the spark session is created. Then, lines data frame is created using spark.readStream method. Post which, the stream is type casted into stream and stored intosad /user/ec2-user/path location in hdfs in JSON format.

```
from pyspark.sql import SparkSession

# Create a spark session using SparkSession builder
spark = SparkSession  \
        .builder  \
        .appName("Load Kafka ClickStream")  \
        .getOrCreate()

# Setting log level to ERROR
spark.sparkContext.setLogLevel('ERROR')


# Subscribe to the topic
# Bootstrap-server - 18.211.252.152
# Port - 9092
# Topic - de-capstone3

lines = spark  \
        .readStream  \
        .format("kafka")  \
        .option("kafka.bootstrap.servers","18.211.252.152:9092")  \
        .option("subscribe","de-capstone3")  \
        .option("startingOffsets","earliest") \
        .load()

# Reveal the schema of the dataFrame
lines.printSchema()

kafkaDF = lines.selectExpr("cast(key as string)","cast(value as string)")
```

```
query = kafkaDF. \
        writeStream \
        .outputMode("append") \
        .format("json") \
        .option("truncate","false") \
        .option("path", "/user/ec2-user/path") \
        .option("checkpointLocation","/user/ec2-user/checkpoint") \
        .start()

query.awaitTermination()
```

Run above file

Above file is then ran with command: `spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2 spark_kafka_to_local.py`

spark-submit is run in a spark version 3.1.2
–packages is used to add the dependency in spark which isn't present in the default distribution.
Correct version of this dependency can be found here:
https://mvnrepository.com/artifact/org.apache.spark/spark-streaming_2.13/3.2.0

Verify the data stored in hadoop file system

Command: `hadoop fs -ls /user/ec2-user/path`

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls /user/ec2-user/path
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
drwxr-xr-x   - hadoop ec2-user          0 2022-01-23 08:44 /user/ec2-user/path/_spark_metadata
-rw-r--r--   1 hadoop ec2-user    1255605 2022-01-23 08:44 /user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json
```

Hadoop fs -ls shows that the fetched file is not present in the hdfs.

Screenshot of ls and data printed

Total lines read from kafka is 3000

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat /user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json | wc -l
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
3000
[hadoop@ip-172-31-31-78 ~]$
```

Spark_local_flatten.py

Following file flattens the data stored in /user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json location

We use spark.read method to read the file from hdfs location, then get_json_object is used to read particular field from input. The new flatten data frame is printed for logging and stored in file in 'com.databricks.spark.csv' format.

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *

# Create a spark session using SparkSession builder
spark = SparkSession.builder \
    .master("local") \
    .appName("Kafka To HDFS") \
    .getOrCreate()

# spark read in json format
# file location:  /user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json
jsonDF = spark \
        .read \
```

```
        .json('/user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json')

df = jsonDF.select( \
        get_json_object(jsonDF["value"], "$.customer_id").alias("customer_id"), \
        get_json_object(jsonDF["value"], "$.app_version").alias("app_version"), \
        get_json_object(jsonDF["value"], "$.OS_version").alias("OS_version"), \
        get_json_object(jsonDF["value"], "$.lat").alias("lat"), \
        get_json_object(jsonDF["value"], "$.lon").alias("lon"), \
        get_json_object(jsonDF["value"], "$.page_id").alias("page_id"), \
        get_json_object(jsonDF["value"], "$.button_id").alias("button_id"), \
        get_json_object(jsonDF["value"], "$.is_button_click").alias("is_button_click"), \
        get_json_object(jsonDF["value"], "$.is_page_view").alias("is_page_view"), \
        get_json_object(jsonDF["value"], "$.is_scroll_up").alias("is_scroll_up"), \
        get_json_object(jsonDF["value"], "$.is_scroll_down").alias("is_scroll_down"), \
        get_json_object(jsonDF["value"], "$.timestamp\n").alias("timestamp"), \
)

# Log some data
print(df.schema)
df.show(5)

# storing the CSV file
# location: 'user/ec2-user/ClickStreamData'
df \
        .coalesce(1) \
        .write \
        .format('com.databricks.spark.csv') \
        .mode('overwrite') \
        .save('user/ec2-user/ClickStreamData', header = 'true')
```

The file is ran using following command:

a. `hadoop fs -cat user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv`

Output of command:

- hadoop fs -ls user/ec2-user/ClickStreamData
  - Output file can be seen
- hadoop fs -cat user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv
  - CSV ClickStream data can be seen

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/ClickStreamData
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-01-23 12:56 user/ec2-user/ClickStreamData/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup     460733 2022-01-23 12:56 user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
customer_id,app_version,OS_version,lat,lon,page_id,button_id,is_button_click,is_page_view,is_scroll_up,is_scroll_down,timestamp
26564820,3.2.35,Android,16.4454865,99.902065,de545711-3914-4450-8c11-b17b8dabb5e1,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,No,Yes,"2020-09-14 09:59:07"
31906387,2.4.7,iOS,-64.813749,-133.527040,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,No,No,Yes,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,Yes,No,"2020-02-09 00:52:13"
83474293,3.1.8,Android,-69.939070,-36.451670,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,No,Yes,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,Yes,Yes,"2020-07-06 02:51:53"
73737907,4.3.19,Android,-18.850508,-116.358375,b328829e-17ae-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,No,Yes,"2020-04-26 06:18:16"
36927433,3.2.26,iOS,-84.6857245,-146.507678,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,Yes,Yes,No,Yes,"2020-02-06 10:21:18"
```

Step 2: Read the data from Sqoop

Squoop is used to transfer data from SQL to HDFS file system in CSV format.

Command explanation

–connect - JDBC string of the SQL database

–table - table name which is supposed to be moved

–username username to whom the access is provided of the DB

–password password of the user mentioned above

–target-dir - place where sqoop will store the file

Command:

```
sqoop import --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --table bookings --username student --password STUDENT123 --target-dir user/ec2-user/bookings_data -m 1
```

Command 1: hadoop fs -ls user/ec2-user/bookings_data

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/bookings_data
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLogge
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/St
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-01-23 13:46 user/ec2-user/bookings_data/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup     165678 2022-01-23 13:46 user/ec2-user/bookings_data/part-m-00000
```

Command 2: hadoop fs -cat user/ec2-user/bookings_data/part-m-00000

Command 3: hadoop fs -cat user/ec2-user/bookings_data/part-m-00000 | wc -l

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/bookings_data/part-m-00000 | wc -l
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/s
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
1000
```

Step3: Create date wise aggregation of bookings data

Datewise_bookings_aggregates_spark.py
Following script is used to create date wise aggregation. First, CSV data is read using spark.read command. Since the data stored by sqoop didnt have header, column names were changed fromi _c0, _c1 to respective names. Then groupBy('date') is done in the data frame and both original data with column names and aggregated numbers data is stored into different hdfs folders

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *

spark = SparkSession \
        .builder \
        .appName("Aggregate Date Wise Bookings Data") \
        .getOrCreate()

bookingsDF = spark \
        .read \
        .csv("user/ec2-user/bookings_data/part-m-00000", inferSchema=True)

bookingsDFWithHeader = bookingsDF.withColumnRenamed("_c0","booking_id") \
            .withColumnRenamed("_c1","customer_id") \
```

```
        .withColumnRenamed("_c2","driver_id") \
        .withColumnRenamed("_c3","customer_app_version")  \
        .withColumnRenamed("_c4","customer_phone_os_version") \
        .withColumnRenamed("_c5","pickup_lat") \
        .withColumnRenamed("_c6","pickup_lon") \
        .withColumnRenamed("_c7","drop_lat") \
        .withColumnRenamed("_c8","drop_lon") \
        .withColumnRenamed("_c9","pickup_timestamp")  \
        .withColumnRenamed("_c10","drop_timestamp")  \
        .withColumnRenamed("_c11","trip_fare") \
        .withColumnRenamed("_c12","tip_amount")  \
        .withColumnRenamed("_c13","currency_code") \
        .withColumnRenamed("_c14","cab_color")  \
        .withColumnRenamed("_c15","cab_registration_no") \
        .withColumnRenamed("_c16","customer_rating_by_driver")  \
        .withColumnRenamed("_c17","rating_by_customer")  \
        .withColumnRenamed("_c18","passenger_count")

bookingsDFWithHeader =  bookingsDFWithHeader.withColumn("date",
date_format('pickup_timestamp', "yyyy-MM-dd"))
bookingsDFWithHeader.show(10)

# Creating Date wise aggregated data flow
dateAggregatedDF = bookingsDFWithHeader \
        .select('date') \
        .groupBy('date') \
        .count()

# Log the new data flow and find count. Should be 289
dateAggregatedDF.show()
dateAggregatedDF.count()


# write the dataframe to HDFS
bookingsDFWithHeader \
        .coalesce(1) \
        .write \
        .format('com.databricks.spark.csv') \
        .mode('overwrite') \
        .save('user/ec2-user/bookings_data_with_header', header = 'true')



# write the aggregated dataframe to HDFS
```

```
dateAggregatedDF \
        .coalesce(1) \
        .write \
        .format('com.databricks.spark.csv') \
        .mode('overwrite') \
        .save('user/ec2-user/date_aggregated_bookings', header='true')
```

Run spark-submit command to run the above file

- Command: `spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2 datewise_bookings_aggregates_spark.py`

Files in HDFS:

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBi
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/Stati
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 4 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 12:56 user/ec2-user/ClickStreamData
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 13:46 user/ec2-user/bookings_data
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/bookings_data_with_header
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/date_aggregated_bookings
[hadoop@ip-172-31-31-78 ~]$
```

LS bookings data with header

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/bookings_data_with_header
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/bookings_data_with_header/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup     176961 2022-01-23 15:01 user/ec2-user/bookings_data_with_header/part-00000-df625fbb-34aa-4ea0-b4db-7bdd3aa5468a-c000.csv
[hadoop@ip-172-31-31-78 ~]$
```

cat bookings data with header

Number of lines in bookings data with header

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/bookings_data_with_header/part-00000-df625fbb-34aa-4ea0-b4db-7bdd3aa5468a-c000.csv | wc -l
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
1001
```

Ls cat aggregated data

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/date_aggregated_bookings
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/date_aggregated_bookings/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       3769 2022-01-23 15:01 user/ec2-user/date_aggregated_bookings/part-00000-d9ea8de7-41c1-4f0d-84cc-964db6285270-c000.csv
[hadoop@ip-172-31-31-78 ~]$
```

cat date aggregated data

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/date_aggregated_bookings/part-00000-d9ea8de7-41c1-4f0d-84cc-964db6285270-c000.csv
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
date,count
2020-06-20,1
2020-04-20,3
2020-05-14,3
2020-04-22,2
2020-03-16,2
2020-09-16,2
2020-05-16,5
2020-01-18,4
2020-10-04,5
2020-03-05,5
2020-04-25,4
2020-05-26,2
2020-01-10,2
2020-06-07,3
2020-10-05,4
2020-08-04,7
2020-02-02,6
2020-05-18,2
2020-08-23,3
2020-08-17,4
2020-03-11,2
2020-10-08,4
2020-03-28,1
2020-08-22,6
```

Number of lines in date aggregated data

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/date_aggregated_bookings/part-00000-d9ea8de7-41c1-4f0d-84cc-964db6285270-c000.csv | wc -l
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
290
```

Step 4: Create HIVE tables and add data

Use “hive;” to use hive.

-- Create database

```
CREATE DATABASE IF NOT EXISTS yoc_riders ;
```

-- use the new db

```
use yoc_riders ;
```

-- create click stream table

```
CREATE TABLE IF NOT EXISTS click_stream_data(customer_id INT,
app_version STRING, os_version STRING, lat DOUBLE, lon DOUBLE, page_id
STRING, button_id STRING, is_button_click VARCHAR(3), is_page_view
VARCHAR(3), is_scroll_up VARCHAR(3), is_scroll_down VARCHAR(3),
`timestamp` TIMESTAMP) ROW FORMAT DELIMITED FIELDS TERMINATED BY ","
LINES TERMINATED BY '\n';
```

-- create bookings data table

```
CREATE TABLE IF NOT EXISTS bookings_data( 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 INT, tip_amount INT, currency_code
STRING, cab_color STRING, cab_registration_no STRING,
customer_rating_by_driver INT, rating_by_customer INT, passenger_count
INT, `date` TIMESTAMP) ROW FORMAT DELIMITED FIELDS TERMINATED BY ","
LINES TERMINATED BY '\n';
```

-- create date aggregated booking data

```
 CREATE TABLE IF NOT EXISTS datewise_booking_data(`date` STRING,
`count` INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LINES
TERMINATED BY '\n';
```

Verification if tables are created

> show tables;

```
hive> show tables;
OK
bookings_data
click_stream_data
datewise_booking_data
Time taken: 0.135 seconds, Fetched: 3 row(s)
```

> describe bookings_data;

```
hive> describe bookings_data;
OK
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                   int
tip_amount                  int
currency_code               string
cab_color                   string
cab_registration_no         string
customer_rating_by_driver           int
rating_by_customer          int
passenger_count             int
date                        timestamp
Time taken: 0.047 seconds, Fetched: 20 row(s)
```

> describe clickstrem data

```
hive> describe click_stream_data;
OK
customer_id             int
app_version             string
os_version              string
lat                     double
lon                     double
page_id                 string
button_id               string
is_button_click         varchar(3)
is_page_view            varchar(3)
is_scroll_up            varchar(3)
is_scroll_down          varchar(3)
timestamp               timestamp
Time taken: 0.046 seconds, Fetched: 12 row(s)
```

> describe datewise_booking_data

```
hive> describe datewise_booking_data;
OK
date                    string
count                   int
Time taken: 0.043 seconds, Fetched: 2 row(s)
```

Load data in the above created tables

We can use LOAD DATA INPATH command to load the data which requires the location as parameter.

-- Load Click Stream data

```
LOAD DATA INPATH 'user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv' INTO TABLE click_stream_data ;
```

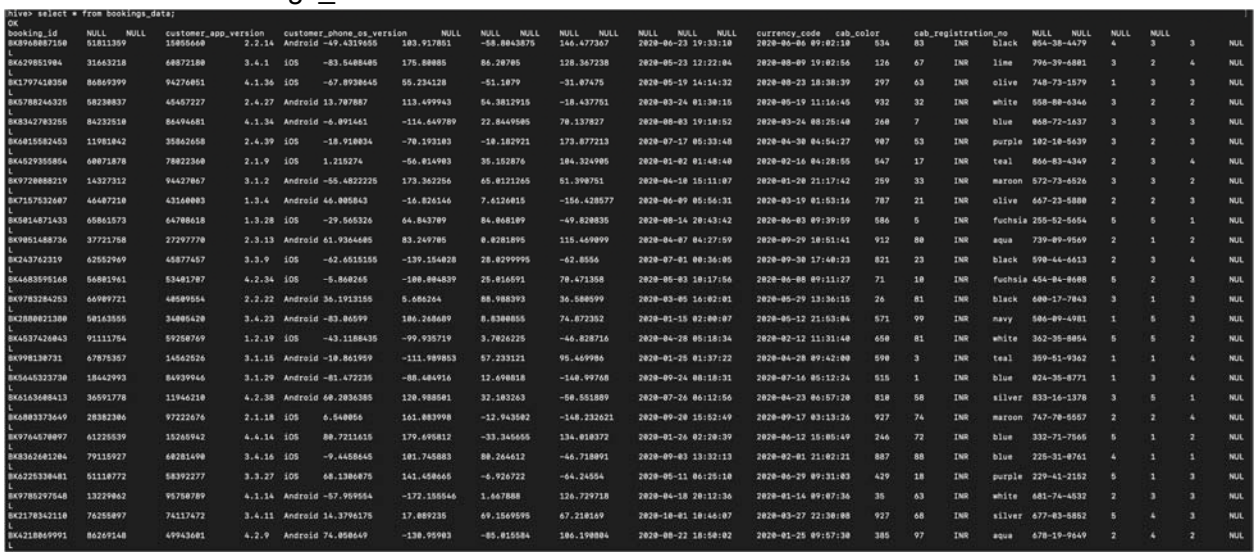
-- Load bookings data

```
LOAD DATA INPATH 'user/ec2-user/bookings_data_with_header/part-00000-
df625fbb-34aa-4ea0-b4db-7bdd3aa5468a-c000.csv' into TABLE
bookings_data ;
```

-- Load datewise aggregated data

```
LOAD DATA INPATH 'user/ec2-user/date_aggregated_bookings/part-00000-
d9ea8de7-41c1-4f0d-84cc-964db6285270-c000.csv' into TABLE
datewise_booking_data;
```

> select * from bookings_data



```
hive> select * from bookings_data;
OK
booking_id      NULL    NULL    customer_app_version    customer_phone_os_version       NULL    NULL    NULL    NULL    NULL    NULL    currency_code   cab_color       cab_registration_no     NULL    NULL    NULL    NULL
BK8968887150    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     83      INR     black   054-38-4479     4       3       3       NUL
L
BK629851984     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     67      INR     lime    796-39-6801     3       2       4       NUL
L
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     63      INR     olive   748-73-1579     1       3       3       NUL
L
BK5788244326    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     32      INR     white   558-80-6346     3       2       2       NUL
L
BK8342703255    84232510        86494681        4.1.34  Android -6.091461       -114.649789     22.8449585      70.137827       2020-08-03 19:10:52     2020-03-24 08:25:40     260     7       INR     blue    068-72-1637     3       3       3       NUL
L
BK6015582453    11981042        35862658        2.4.39  iOS     -18.910034      -70.193103      -10.182921      173.877213      2020-07-17 05:33:48     2020-04-30 04:54:27     907     53      INR     purple  102-10-5639     3       2       3       NUL
L
BK4529355854    60071878        78022360        2.1.9   iOS     1.215274        -56.014903      35.152876       104.324905      2020-01-02 01:48:40     2020-02-16 04:28:55     647     17      INR     teal    866-83-4349     2       3       4       NUL
L
BK9720088219    14327312        94427067        3.1.2   Android -55.4822225     173.362256      65.8121265      51.398751       2020-04-10 15:11:07     2020-01-20 21:17:42     259     33      INR     maroon  572-73-6526     3       3       2       NUL
L
BK7157532607    46407210        43160003        1.3.4   Android 46.005843       -16.826146      7.6126015       -156.428577     2020-06-09 05:56:31     2020-03-19 01:53:16     787     21      INR     olive   667-23-5880     2       2       3       NUL
L
BK5014871433    65861573        64708618        1.3.28  iOS     -29.565326      64.843709       84.868109       -49.820835      2020-08-14 20:43:42     2020-06-03 09:39:59     586     5       INR     fuchsia 255-52-5654     5       5       1       NUL
L
BK9051488736    37721758        27297770        2.3.13  Android 61.9364685      83.249705       0.0281895       115.469099      2020-04-07 04:27:59     2020-09-29 10:51:41     912     80      INR     aqua    739-09-9569     2       1       2       NUL
L
BK243762319     62552969        45877457        3.3.9   iOS     -62.6515155     -139.154028     28.0299995      -62.8556        2020-07-01 00:36:05     2020-09-30 17:40:23     821     23      INR     black   590-44-6613     2       3       4       NUL
L
BK4683595168    56801961        53401707        4.2.34  iOS     -5.860265       -100.004839     25.016591       70.471358       2020-05-03 10:17:56     2020-06-08 09:11:27     71      10      INR     fuchsia 454-84-8608     5       2       3       NUL
L
BK9783284253    66909721        48589554        2.2.22  Android 36.1913155      5.686264        88.988393       36.580599       2020-03-05 16:02:01     2020-05-29 13:36:15     26      81      INR     black   600-17-7043     3       1       3       NUL
L
BK2880021380    50163555        34005420        3.4.23  Android -83.86599        186.268689     8.8300855       74.872352       2020-01-15 02:00:07     2020-05-12 21:53:04     571     99      INR     navy    506-09-4981     1       5       3       NUL
L
BK4537426843    91111754        59250769        1.2.19  iOS     -43.1188435     -99.935719      3.7026225       -46.828716      2020-04-28 05:18:34     2020-02-12 11:31:40     650     81      INR     white   362-35-8054     5       5       2       NUL
L
BK998130731     67875357        14562526        3.1.15  Android -10.861959       -111.989853    57.233121       95.469986       2020-01-25 01:37:22     2020-04-28 09:42:00     590     3       INR     teal    359-51-9362     1       1       4       NUL
L
BK5445323730    18442993        84939946        3.1.29  Android -81.472235       -88.484916     12.690818       -140.99768      2020-09-24 08:18:31     2020-07-16 05:12:24     515     1       INR     blue    024-35-8771     1       3       4       NUL
L
BK6163608413    36591778        11946210        4.2.38  Android 60.2036385       120.988501     32.103263       -50.551889      2020-07-26 06:12:56     2020-04-23 06:57:20     810     58      INR     silver  833-16-1378     3       5       1       NUL
L
BK4803373649    28382306        97222676        2.1.18  iOS     6.540056        161.083998      -12.943502      -148.232621     2020-09-20 15:52:49     2020-09-17 03:13:26     927     74      INR     maroon  747-70-5557     2       2       4       NUL
L
BK9764670097    61225539        15265942        4.4.14  iOS     80.7211615      179.695812      -33.345655      134.010372      2020-01-26 02:20:39     2020-06-12 15:05:49     246     72      INR     blue    332-71-7565     5       1       2       NUL
L
BK8362601204    79115927        60281490        3.4.16  iOS     -9.4458645      101.745883      80.264612       -46.718091      2020-09-03 13:32:13     2020-02-01 21:02:21     887     88      INR     blue    225-31-0761     4       1       1       NUL
L
BK6225330481    51110772        58392277        3.3.27  iOS     68.1306875      141.458665      -6.926722       -64.24554       2020-05-11 06:25:10     2020-06-29 09:31:03     429     18      INR     purple  229-41-2152     5       1       3       NUL
L
BK9785297548    13229862        95750789        4.1.14  Android -67.959584       -172.155546    1.667888        126.729718      2020-04-18 20:12:36     2020-01-14 09:07:36     35      63      INR     white   681-74-4532     2       3       3       NUL
L
BK2178342110    76255097        74117472        3.4.11  Android 14.3796175       17.889235      69.1569595      67.210149       2020-10-01 10:46:07     2020-03-27 22:30:08     927     68      INR     silver  677-83-5852     5       4       3       NUL
L
BK4218069991    86269148        49943681        4.2.9   Android 74.050649        -130.95983     -85.015584      186.198884      2020-08-22 18:50:02     2020-01-25 09:57:30     385     97      INR     aqua    678-19-9649     2       4       2       NUL
L
```

> select * from click_stream_data;

```
hive> select * from click_stream_data;
OK
NULL    app_version     OS_version      NULL    NULL    page_id button_id       is_     is_     is_     is_     NULL
26564820        3.2.35  Android 16.4454865      99.902065       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes     No      Yes     NULL
31906387        2.4.7   iOS     -64.813749      -133.52704      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      No      Yes     Yes     NULL
25713677        3.4.12  Android 89.943435       127.313415      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No      Yes     No      NULL
83474293        3.1.8   Android -69.93907       -36.45167       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No      Yes     No      NULL
63727807        2.2.9   iOS     64.082108       -81.822078      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes     Yes     Yes     NULL
73737907        4.3.19  Android -18.850508      -116.358375     b328829e-17ae-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    No      Yes     No      Yes     NULL
36927433        3.2.26  iOS     -84.6857245     -146.507678     de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     No      Yes     NULL
12691783        3.3.11  Android 54.3852925      -37.411814      de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     Yes     No      No      NULL
22635021        4.4.36  iOS     -31.8055        150.65565       e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      No      No      No      NULL
23593546        1.2.16  Android 8.8918475       -83.929878      de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No      Yes     No      NULL
16929816        1.2.25  Android -34.8576385     -136.639835     e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No      No      Yes     NULL
41929865        1.1.14  iOS     -34.3613155     -137.467833     b328829e-17ae-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      Yes     No      No      NULL
68940320        1.1.17  Android 81.738239       -104.207763     de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     No      Yes     No      NULL
95405928        1.2.10  Android 44.185627       78.125488       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    No      No      No      No      NULL
56235860        1.4.16  Android -61.178407      8.534126        de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     Yes     Yes     NULL
81758845        4.4.30  Android 78.984984       167.315596      e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     No      No      NULL
46822067        1.2.26  Android 45.8993985      135.895118      de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     Yes     Yes     Yes     NULL
37623067        3.2.27  iOS     -74.1131505     -170.715421     b328829e-17ae-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      Yes     No      No      NULL
14571693        3.4.21  iOS     -28.8329055     161.132369      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes     No      No      NULL
62605529        2.1.36  Android 44.196239       -15.354515      e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     No      No      NULL
83657253        2.1.2   iOS     38.7036085      -59.081929      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes     No      Yes     NULL
47086145        3.3.16  Android 62.2950045      47.37952        de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    No      No      No      Yes     NULL
40025212        4.4.12  iOS     -87.4878555     -139.785288     b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No      Yes     No      NULL
19283684        2.4.23  iOS     47.447583       31.679116       e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    Yes     No      No      No      NULL
21563520        1.3.31  iOS     -24.616214      -66.741365      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     No      Yes     No      NULL
35705060        2.2.13  iOS     -11.2298655     8.010661        de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      No      Yes     No      NULL
12252116        2.1.13  iOS     -7.7110045      -49.999098      e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     Yes     No      Yes     NULL
88636479        3.2.22  iOS     -46.236733      -21.683563      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes     No      Yes     NULL
41664607        4.3.9   Android -0.2391655      -77.952798      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     No      Yes     Yes     NULL
72578040        3.3.25  Android 49.2911795      -152.594458     e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    No      Yes     No      Yes     NULL
17535587        4.2.13  iOS     -43.4202565     24.324168       e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      No      Yes     No      NULL
14734830        4.2.29  Android -25.862119      65.408339       e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes     Yes     No      NULL
68194999        2.3.4   iOS     51.4876125      73.580584       e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      No      No      Yes     NULL
96812430        3.1.29  iOS     28.057742       -81.253091      e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     Yes     Yes     No      NULL
47631980        2.1.23  iOS     48.842616       -66.047399      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     No      No      Yes     NULL
64011321        2.2.30  iOS     -11.1664985     171.128929      b328829e-17ae-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     Yes     No      NULL
53911958        2.2.39  iOS     5.9806555       179.002821      de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    No      No      No      Yes     NULL
98568662        2.1.25  Android 5.6975445       37.819333       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      No      Yes     No      NULL
83589769        3.2.1   iOS     37.072072       -49.406494      e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    No      No      No      Yes     NULL
49512087        4.4.9   Android 18.5024405      88.260556       de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      Yes     No      No      NULL
22704267        3.3.32  iOS     -2.525445       -69.681696      de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     Yes     No      No      NULL
21265572        2.4.29  Android -28.9880665     -85.010425      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     No      Yes     Yes     NULL
86078309        3.3.21  iOS     14.5942885      -164.817167     e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No      Yes     No      NULL
68422739        1.3.15  Android 31.731629       11.394737       e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      Yes     No      No      NULL
85346419        3.3.36  Android 68.4022575      152.412777      b328829e-17ae-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      Yes     Yes     No      NULL
66710552        3.3.28  iOS     -69.3930215     -77.610069      e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    Yes     No      Yes     No      NULL
11316378        1.4.11  Android 28.470736       150.0967        de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     No      Yes     NULL
83156650        1.4.14  iOS     48.209126       95.714431       b328829e-17ae-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      Yes     No      No      NULL
29702837        1.1.12  Android 1.1570665       -152.366834     b328829e-17ae-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     No      No      NULL
35776107        1.3.3   iOS     29.3031035      -156.45416      b328829e-17ae-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No      No      Yes     NULL
44822406        2.1.12  iOS     -39.698084      19.976282       de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No      Yes     No      NULL
72311620        1.2.25  iOS     -15.094244      -58.166225      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     Yes     Yes     NULL
99713267        1.2.13  Android -63.011675      -168.394573     de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      Yes     No      No      NULL
66968218        3.2.3   iOS     -71.376849      -141.492428     e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     Yes     Yes     Yes     NULL
22783844        1.1.20  Android -25.972366      17.315069       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    No      No      No      No      NULL
92178756        3.3.29  Android -85.98095       -12.773762      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     Yes     Yes     NULL
93617994        1.1.24  iOS     -85.7651895     -116.523285     de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     Yes     Yes     No      NULL
93840678        1.4.19  iOS     10.996018       -154.33489      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     Yes     Yes     Yes     NULL
65044856        3.4.30  Android -86.735896      -168.343227     e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     Yes     Yes     Yes     NULL
20005705        1.4.11  Android 44.3767255      -15.33208       de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     Yes     Yes     Yes     NULL
86904724        4.3.15  Android -3.0758015      -97.024893      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    Yes     Yes     No      No      NULL
91583778        3.4.5   Android -18.2869335     84.429953       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    Yes     No      No      No      NULL
62561362        3.3.10  iOS     28.044871       -16.207224      e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    No      Yes     Yes     No      NULL
84536540        4.4.12  iOS     36.1815295      -126.499135     b328829e-17ae-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No      No      Yes     NULL
```

> select * from datewise_booking_data;

```
hive> select * from datewise_booking_data;
OK
date    NULL
2020-06-20      1
2020-04-20      3
2020-05-14      3
2020-04-22      2
2020-03-16      2
2020-09-16      2
2020-05-16      5
2020-01-18      4
2020-10-04      5
2020-03-05      5
2020-04-25      4
2020-05-26      2
2020-01-10      2
2020-06-07      3
2020-10-05      4
2020-08-04      7
2020-02-02      6
2020-05-18      2
2020-08-23      3
2020-08-17      4
2020-03-11      2
2020-10-08      4
2020-03-28      1
2020-08-22      6
2020-02-11      1
2020-10-07      1
2020-03-01      1
```