

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 into [/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/efs/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

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat /user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/share/aws/emrfs/lib/sl4fj-log4j12-1.7.25.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-x--  hadoop ec2-user 2556489 Dec 14 07:22 .
drwxr-x--  hadoop ec2-user 2556489 Dec 14 07:22 ..
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat /user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/share/aws/emrfs/lib/sl4fj-log4j12-1.7.25.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]
{"value": {"customer_id": "31984387", "app_version": "2.2.9", "os_version": "Android", "lat": "16.445865", "lon": "+99.992865", "page_id": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\fcba0ddaa-1231-11e0-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-11 09:09:37(\n)"}, {"value": {"customer_id": "31984387", "app_version": "2.4.1", "os_version": "Android", "lat": "-64.813749", "lon": "+133.527848", "page_id": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-11 10:38:13(\n)"}, {"value": {"customer_id": "20753677", "app_version": "3.4.23", "os_version": "Android", "lat": "+69.043439", "lon": "+127.033451", "page_id": "\0328829-17ae-11eb-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-11 07:53:11(\n)"}, {"value": {"customer_id": "63727807", "app_version": "2.2.9", "os_version": "Android", "lat": "17.131105", "lon": "+106.805813", "page_id": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\ele99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-11 17:28:33(\n)"}, {"value": {"customer_id": "63727807", "app_version": "2.9.2", "os_version": "Android", "lat": "+81.622687", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\fcba0ddaa-1231-11e0-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-07-08 02:51:53(\n)"}, {"value": {"customer_id": "63727807", "app_version": "2.2.9", "os_version": "Android", "lat": "+81.622687", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\fcba0ddaa-1231-11e0-adc1-0242ac120002", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-04-26 06:18:10(\n)"}, {"value": {"customer_id": "35022433", "app_version": "3.2.26", "os_version": "iOS", "lat": "+34.485724", "lon": "-144.587679", "page_id": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-11 07:34:27(\n)"}, {"value": {"customer_id": "12692783", "app_version": "3.3.11", "os_version": "Android", "lat": "+54.885929", "lon": "-137.411914", "page_id": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\ele99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-08 04:23:56(\n)"}, {"value": {"customer_id": "55239869", "app_version": "1.2.16", "os_version": "Android", "lat": "-18.015786", "lon": "-105.486039", "page_id": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-06-02 00:33:58(\n)"}, {"value": {"customer_id": "23598541", "app_version": "1.2.19", "os_version": "Android", "lat": "+63.929878", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\ele99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-06-01 23:01:15(\n)"}, {"value": {"customer_id": "61692984", "app_version": "1.2.25", "os_version": "Android", "lat": "-34.857635", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-13 16:30:17(\n)"}, {"value": {"customer_id": "12290021", "app_version": "1.2.27", "os_version": "Android", "lat": "-34.467833", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-13 16:41:13(\n)"}, {"value": {"customer_id": "68944329", "app_version": "1.1.17", "os_version": "Android", "lat": "+184.207763", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\fcba0ddaa-1231-11e0-adc1-0242ac", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-14 03:39:05(\n)"}, {"value": {"customer_id": "55239869", "app_version": "1.2.18", "os_version": "Android", "lat": "-16.174497", "lon": "-55.085684", "page_id": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-07-05 10:53:17(\n)"}, {"value": {"customer_id": "78759845", "app_version": "3.4.43", "os_version": "Android", "lat": "+27.375598", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-07-01 15:28:34(\n)"}, {"value": {"customer_id": "386541", "app_version": "1.0.220", "os_version": "Android", "lat": "+45.999998", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\fcba0ddaa-1231-11e0-adc1-0242ac", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-06-17 19:52:06(\n)"}, {"value": {"customer_id": "37622807", "app_version": "1.0.5", "os_version": "Android", "lat": "-74.131598", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-01-28 00:05:01(\n)"}, {"value": {"customer_id": "62740529", "app_version": "2.1.26", "os_version": "Android", "lat": "+44.2939", "lon": "\ude5d51-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-28 11:20:48(\n)"}, {"value": {"customer_id": "83657203", "app_version": "2.1.16", "os_version": "Android", "lat": "+38.793409", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-10 09:08:19(\n)"}, {"value": {"customer_id": "70825143", "app_version": "1.3.23", "os_version": "Android", "lat": "+62.229084", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\ele99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-15 20:53:51(\n)"}, {"value": {"customer_id": "83657203", "app_version": "2.0.16", "os_version": "Android", "lat": "+38.793409", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-10 09:08:20(\n)"}, {"value": {"customer_id": "13567322", "app_version": "1.0.14", "os_version": "Android", "lat": "-26.814188", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-08 14:47:35(\n)"}, {"value": {"customer_id": "62740529", "app_version": "2.1.26", "os_version": "Android", "lat": "+44.2939", "lon": "\ude5d51-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-08-28 11:20:49(\n)"}, {"value": {"customer_id": "83657203", "app_version": "2.1.16", "os_version": "Android", "lat": "+38.793409", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-10 09:08:21(\n)"}, {"value": {"customer_id": "13567322", "app_version": "1.0.14", "os_version": "Android", "lat": "-26.814188", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-08 14:47:36(\n)"}, {"value": {"customer_id": "13567322", "app_version": "1.0.14", "os_version": "Android", "lat": "-26.814188", "lon": "\ude5d71-3914-4458-8c11-017b0dd0051", "button_id": "\v95dd57b-779f-49d0-819d-b47648d43", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp": "+2020-05-08 14:47:37(\n)"}]
```

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/sl4fj-log4j12-1.7.25.jar!/org/slf4j.impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emrfs/lib/sl4fj-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-4559-4e99-8185-3cccd472023f2-c000.csv
[[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3cccd472023f2-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,16,4454845,99,982065,de545711-3914-4450-8c11-b17b8dabb5e1,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,No,Yes,"2020-09-14 09:59:07"
31996387,2.4.7,iOS,-64,813749,-133,527940,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b690483e554,No,No,Yes,Yes,"2020-05-16 16:39:21"
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,-69,939876,-36,451670,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,No,Yes,No,"2020-06-17 18:42:50"
63728087,2.2.9,iOS,64,082198,-81,822078,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,Yes,Yes,"2020-07-06 02:51:53"
73737987,4.3.19,Android,-18,858508,-116,358375,b328829e-17ae-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,Yes,"2020-04-26 06:18:18"
36977433,3.2.26,iOS,-84,6R57245,-146,507478,dm545711-3914-4450-Rc11-17hdahh5n1.a95ffn5fh-779f-49dh-R19d-h690483e554,Yes,Yes,No,Yes,"2020-07-06 10:21:18"
```

Step 2: Read the data from Squeryl

Squeryl 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 squeryl will store the file

Command:

```
squeryl 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/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 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

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/bookings_data/part-m-00000
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]
BK948887158, 51811359, 15855668, 2.2.14, Android, -49, 4319455, 103, 917851, -58, 8043875, 146, 477367, 2028-06-23 19:33:10, 0, 2020-06-06 09:20:10, 0, 534, 83, INR, black, 854-38-4479, 4, 3, 3
PK29851984, 31663218, 68872188, 3, 4, 1, iOS, -83, 5408485, 175, 80085, 86, 28705, 128, 367238, 2028-05-23 12:22:04, 0, 2020-08-09 19:02:56, 0, 126, 67, INR, lime, 796-39-6801, 3, 2, 4
BK1797418359, 86869394, 94276951, 4, 1, 36, iOS, -67, 8938645, 55, 234128, -51, 1979, -31, 07475, 2028-05-19 14:14:32, 0, 2028-08-23 18:30:39, 0, 297, 63, INR, olive, 748-73-1579, 1, 3, 3
BK5789744235, 58230837, 46487227, 2, 4, 27, Android, 13, 707887, 112, 99943, 54, 3812915, -18, 437751, 2028-03-24 01:30:15, 0, 2028-06-19 11:16:45, 0, 932, 32, INR, white, 558-80-4346, 3, 2, 2
BK83423519, 8649481, 4, 1, 34, Android, -6, 891461, 114, 649789, 22, 8449505, 70, 137827, 2028-08-03 19:10:52, 0, 2028-03-24 08:25:40, 0, 266, 7, INR, blue, 868-72-1637, 3, 3, 3
BK6011582453, 11981047, 35862658, 55, 234128, -51, 1979, -31, 07475, 2028-05-19 14:14:32, 0, 2028-08-23 18:30:39, 0, 297, 63, INR, purple, 162-10-5639, 3, 2, 3
BK4529355884, 68071878, 78822368, 2, 1, 9, iOS, 1, 215274, 56, 814983, 35, 152876, 184, 324985, 2028-01-02 01:48:48, 0, 2028-02-16 04:28:55, 0, 547, 17, INR, teal, 866-83-4349, 2, 3, 4
BK9720888219, 14327312, 94427067, 3, 1, 2, Android, 13, 707887, 112, 99943, 54, 3812915, -18, 437751, 2028-04-16 15:11:07, 0, 2028-01-20 21:17:42, 0, 259, 33, INR, maroon, 572-73-6526, 3, 3, 2
BK157832697, 43160003, 1, 3, 4, Android, 46, 005843, -16, 826146, 7, 6126015, -156, 428577, 2028-06-09 05:56:31, 0, 2028-03-19 01:53:16, 0, 787, 21, INR, olive, 667-23-5880, 2, 2, 3
BK560148736, 37721758, 2797770, 2, 3, 13, Android, 61, 9364686, 83, 249795, 0, 0281895, 115, 469099, 2028-04-07 04:27:59, 0, 2028-09-29 10:51:41, 0, 912, 88, INR, aqua, 739-09-9569, 2, 1, 2
BK243762319, 62552969, 45877487, 3, 3, 9, iOS, -62, 6515155, -139, 154286, 28, 0299995, -62, 8556, 2828-07-01 08:36:05, 0, 2028-09-38 17:48:23, 0, 821, 23, INR, black, 599-44-6613, 2, 3, 4
BK4683595168, 56801961, 53481707, 4, 2, 34, iOS, -5, 860265, 108, 0848439, 25, 016591, 70, 471358, 2028-05-03 18:17:56, 0, 2028-06-08 09:11:27, 0, 71, 10, INR, fuchsia, 454-04-0608, 5, 2, 3
BK9783284253, 66989721, 40509554, 2, 2, 22, Android, 34, 1913155, 5, 686264, 88, 988393, 36, 588659, 2028-03-05 16:02:01, 0, 2028-05-29 13:36:15, 0, 26, 81, INR, black, 680-17-7043, 3, 1, 3
BK2880021388, 58163555, 34005428, 3, 4, 23, Android, -83, 06599, 186, 268689, 8, 8308855, 74, 872352, 2028-01-15 02:00:07, 0, 2028-05-12 21:53:04, 0, 571, 99, INR, navy, 586-09-4981, 1, 5, 3
BK4537426043, 91111751, 59258769, 1, 3, 19, iOS, -29, 1188435, -99, 935719, 3, 7926225, -46, 828716, 2028-04-28 05:18:34, 0, 2028-02-12 11:31:40, 0, 658, 81, INR, white, 362-35-8054, 5, 5, 2
BK9798130731, 67875357, 14562526, 3, 1, 15, Android, -10, 861959, -111, 989853, 57, 233121, 95, 469986, 2028-01-25 01:37:22, 0, 2028-04-28 09:42:08, 0, 593, 3, INR, teal, 359-51-9362, 1, 1, 4
BK5645232738, 18429493, 84939946, 3, 1, 29, Android, -81, 472235, -88, 404916, 12, 6980818, -148, 99768, 2028-09-24 08:18:31, 0, 2028-07-16 05:12:24, 0, 515, 1, INR, blue, 824-35-8771, 1, 3, 4
BK6163608413, 36591778, 11946218, 4, 2, 38, Android, 60, 2836385, 128, 988501, 32, 103263, -50, 551889, 2028-07-26 06:12:56, 0, 2028-04-23 06:57:20, 0, 818, 58, INR, silver, 833-16-1378, 3, 1
BK4883373649, 28382386, 97222676, 2, 1, 18, iOS, 6, 540856, 161, 083994, -148, 232621, 2028-09-28 15:48:23, 0, 2028-09-17 03:13:26, 0, 927, 74, INR, maroon, 747-70-5557, 2, 2, 4
BK97645780897, 61225539, 15265942, 4, 4, 14, iOS, 80, 7211615, 179, 695812, -33, 345655, 134, 018372, 2028-01-26 02:20:39, 0, 2028-06-12 15:05:49, 0, 246, 72, INR, blue, 332-71-7565, 5, 1, 2
BK8362601284, 79115927, 682812498, 3, 4, 16, iOS, -9, 4458645, 181, 745883, 88, 264612, -46, 718891, 2028-09-03 13:32:13, 0, 2028-02-01 21:02:21, 0, 887, 88, INR, blue, 225-31-0761, 4, 1, 1
BK4225338481, 51110772, 58392277, 3, 3, 27, iOS, 68, 1386975, 141, 458665, -6, 24554, 2028-05-11 06:25:10, 0, 2028-06-29 09:43:03, 0, 429, 18, INR, purple, 229-41-2152, 5, 1, 3
BK9785297548, 13229862, 95780789, 4, 1, 14, Android, -57, 959554, -172, 155546, 1, 667888, 126, 729718, 2028-04-18 20:12:36, 0, 2028-01-14 09:07:36, 0, 35, 63, INR, white, 681-74-4532, 2, 3, 3
BK2178342118, 76255897, 74117472, 3, 4, 11, Android, 14, 3796175, 17, 089235, 69, 0159595, 67, 218169, 2028-10-01 10:46:07, 0, 2028-03-27 22:38:08, 0, 927, 68, INR, silver, 677-03-5852, 5, 4, 3
BK4218869901, 86269118, 49943681, 4, 2, 9, Android, 74, 058649, -138, 95093, -85, 015584, 184, 198884, 2028-08-22 18:50:02, 0, 2028-01-25 09:57:30, 0, 385, 97, INR, aqua, 478-19-9649, 2, 4, 2
```

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/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]
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 from _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/Static
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/efs/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-df625fb-34aa-4ea0-b4db-7bdd3aa5468a-c000
[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

Date		Customer		Booking		Product		Order		Payment		Delivery		Review		
OK		NULL	NULL	CUSTOMER_APP_VERSION	customer_phone_no_version	NULL	NULL	NULL	NULL	CAB_CODE	CAB_COLOR	CAB_REGISTRATION_NO	NULL	NULL	NULL	
BOOKING_ID	51881359	15885648	2.2.14	Android -49.4319469	183.917851	-58.8843875	146.477307	2828-06-23 19:31:18	2028-06-26 09:21:10	534	83	INR	black	054-38-4479	4	
L	UK29851984	31663218	3.4.1	iOS -83.5468495	176.88985	84.29795	128.367238	2828-05-23 12:21:04	2028-08-09 19:02:56	126	67	INR	lime	796-39-6881	3	
I	UK1797420350	86864939	4.1.36	iOS -67.8936445	55.234128	-51.1879	-31.07475	2028-05-19 14:11:32	2028-08-23 18:38:39	297	63	INR	olive	748-73-1579	1	
L	UK5788246328	582398837	2.4.27	Android 13.7078867	113.499943	54.3812915	-18.437701	2028-03-24 01:30:15	2028-05-19 11:16:45	932	32	INR	white	558-88-6346	3	
I	UK8342783256	84232518	4.1.34	Android -6.0911441	-114.649789	22.8449585	70.137827	2028-08-03 19:18:05	2028-03-24 08:25:48	268	7	INR	blue	068-72-1637	3	
I	UK415582453	11981842	2.4.39	iOS -18.910834	-70.191803	-18.182921	175.877213	2028-07-17 06:13:48	2028-04-20 04:15:27	987	53	INR	purple	182-18-5639	3	
L	UK429355854	64971278	2.1.9	iOS 1.215274	-56.814983	35.152876	184.324995	2028-01-02 01:48:40	2028-02-16 04:28:55	847	17	INR	teal	866-83-4349	2	
I	UK972888219	14327312	3.1.2	Android -55.4822228	173.362266	65.8121265	51.399751	2028-04-18 15:11:07	2028-01-20 21:17:42	289	33	INR	maroon	572-73-6526	3	
I	UK157522687	46487218	4.3160003	Android 46.005843	-16.826146	7.6126015	-156.428077	2028-06-09 05:56:31	2028-03-19 01:53:16	787	21	INR	olive	667-23-5880	2	
L	UK5914871433	65861573	1.3.28	iOS -29.665526	64.843789	84.068189	-49.828035	2028-08-14 20:43:42	2028-06-03 09:39:59	586	5	INR	fuchsia	255-52-5654	5	
I	UK9851488736	37722178	2.3.13	Android 41.9364485	83.249785	0.8281995	115.469999	2028-04-07 04:27:59	2028-09-29 10:51:41	912	88	INR	aqua	739-89-9569	2	
L	UK243762319	62552969	45877457	3.3.9	iOS -62.6515185	-139.154828	28.8299995	-62.8556	2028-07-01 00:36:05	2028-09-30 17:48:23	821	23	INR	black	598-44-6613	2
I	UK4483599168	56881961	3.4.34	iOS -5.868265	-108.884839	28.816991	78.473358	2028-06-03 10:17:56	2028-06-08 09:11:27	71	10	INR	fuchsia	454-84-8688	5	
I	UK9783284253	60989721	2.2.22	Android 36.1913155	5.686264	88.988393	30.588099	2028-03-05 16:01:01	2028-05-29 13:30:15	26	81	INR	black	600-17-7843	3	
L	UK2880821308	54163555	34895428	3.4.23	Android -83.865999	186.268689	0.8380055	74.872302	2028-01-18 02:08:07	2028-05-12 21:53:04	571	99	INR	navy	586-89-4981	1
I	UK4537426843	91111754	2.1.19	iOS -43.1188436	-99.935719	3.7026225	-46.828716	2028-04-26 05:18:34	2028-02-12 11:31:48	658	81	INR	white	362-35-8864	5	
L	UK998138731	67878307	14562026	3.1.18	Android -18.861959	-111.989853	57.235121	90.469986	2028-01-25 01:37:22	2028-04-28 09:42:00	698	3	INR	teal	359-51-9362	1
I	UK56453232738	18442993	84939466	3.1.29	Android -81.472235	-88.484916	12.698018	-140.979758	2028-09-24 06:11:31	2028-07-16 05:12:24	915	1	INR	blue	824-35-8771	1
L	UK1634486413	36591778	11946219	4.2.38	Android 40.2936385	120.988601	37.293163	-50.551889	2028-07-26 00:11:56	2028-04-23 06:57:20	818	58	INR	silver	833-14-5370	3
I	UK4883376449	28382386	77222076	2.1.18	iOS 6.548056	161.883998	-12.943582	-148.232021	2028-09-28 15:15:24	2028-09-17 03:13:26	927	74	INR	maroon	747-78-5557	2
I	UK9744578897	61225639	15269942	4.4.14	iOS 88.7211415	179.695812	-33.346565	134.018372	2028-01-26 02:29:09	2028-06-12 15:08:49	246	72	INR	blue	332-73-7565	5
L	UK8362681284	79115927	68281498	3.4.16	iOS -9.4458645	101.745883	88.264612	-46.718091	2028-09-03 13:32:13	2028-02-01 21:02:21	887	88	INR	blue	228-31-0761	4
I	UK6225339481	51118772	58392277	3.3.27	iOS 68.1386875	141.498665	-9.926722	-64.24564	2028-06-11 06:12:18	2028-06-29 09:31:03	429	18	INR	purple	229-41-2152	5
L	UK9785297548	13229862	95768789	4.1.14	Android -87.999504	-372.155644	1.667888	126.279718	2028-04-18 20:12:36	2028-01-14 09:07:36	36	63	INR	white	681-74-4832	2
I	UK2170342118	76258897	74117472	3.4.11	Android 14.3796175	17.889725	69.1569595	67.210169	2028-18-01 10:46:07	2028-03-27 22:30:08	927	68	INR	silver	677-83-5882	5
I	UK4218849951	86269148	49943681	4.2.9	Android 74.0586449	-138.95983	-85.815584	186.198084	2028-08-22 18:58:02	2028-01-25 09:57:38	385	97	INR	aqua	678-19-9649	2

> select * from click_stream_data;

	app_version	OS_version	NULL	NULL	page_id	button_id	is_i	NULL						
26564826	3.2.35	Android 16.4454865	99.902065		de545711-3914-4458-8c11-b17b8dabb5e1									No Yes No Yes NULL
31986387	2.4.7	iOS	-64.813749	-133.52704	de545711-3914-4458-8c11-b17b8dabb5e1									No No Yes Yes NULL
25713677	3.4.12	Android 89.943435	127.313415		b328297e-17ae-11eb-adc1-0242ac120002									No No Yes No NULL
83474293	3.1.8	Android -69.93987	-36.45167		e7bc5f52-1231-11eb-adc1-0242ac120002									Yes No Yes No NULL
63727887	2.2.9	iOS	64.887188	-81.822878	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes Yes Yes NULL
73737987	4.3.19	Android -18.895988	-116.358375		b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
36927433	3.2.26	iOS	-84.6857245	-146.867678	de545711-3914-4458-8c11-b17b8dabb5e1									Yes Yes No Yes NULL
12691383	3.1.10	Android 54.3882925	-31.430441		de545711-3914-4458-8c11-b17b8dabb5e1									No No No No NULL
22635621	4.4.36	iOS	-8.891675	150.90565	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
23035644	1.2.16	Android 8.891675	-83.929878		de545711-3914-4458-8c11-b17b8dabb5e1									No No No No NULL
16929816	3.2.25	Android 34.8576385	-136.639835		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
41929865	1.1.14	iOS	-34.3413155	-137.467833	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No No NULL
68949320	1.1.17	Android 81.738239	-194.207743		de545711-3914-4458-8c11-b17b8dabb5e1									No Yes No Yes NULL
95495928	1.2.18	Android 44.186527	78.125488		e7bc5f52-1231-11eb-adc1-0242ac120002									No No No No NULL
65235869	4.1.16	Android -61.178487	8.534213		de545711-3914-4458-8c11-b17b8dabb5e1									No Yes Yes Yes NULL
81758845	4.4.38	Android 78.994984	167.315596		e7bc5f52-1231-11eb-adc1-0242ac120002									Yes Yes No Yes NULL
46822667	1.2.26	Android 45.8939385	135.895118		de545711-3914-4458-8c11-b17b8dabb5e1									No Yes Yes Yes NULL
37623867	3.2.27	iOS	-74.1131585	-178.715421	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No No NULL
14571693	3.4.21	iOS	-28.8329855	161.132369	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No No NULL
62605529	2.1.31	Android 44.196239	-35.3564515		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
83457253	2.1.2	iOS	38.7836905	-59.081929	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
47886145	3.3.16	Android 62.2958945	47.37952		de545711-3914-4458-8c11-b17b8dabb5e1									No No No Yes NULL
40925212	4.4.12	iOS	-87.4678555	-139.785288	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
19283684	2.4.23	iOS	47.447583	31.679116	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
21563520	1.3.31	iOS	-24.616214	-66.741365	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
35705068	2.2.13	iOS	-11.2298655	8.018661	de545711-3914-4458-8c11-b17b8dabb5e1									No Yes No Yes NULL
12252116	2.1.13	iOS	-7.7118045	-49.999898	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
88636479	3.2.22	iOS	-46.236733	-21.683563	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
41664687	4.3.9	Android -0.2919165	-77.952798		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes Yes Yes NULL
72578848	3.3.25	Android 49.2911795	-152.594458		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
17535587	4.2.13	iOS	-43.4282565	24.324168	de545711-3914-4458-8c11-b17b8dabb5e1									No Yes No Yes NULL
14734838	4.2.29	Android -25.862119	65.408339		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
68194999	2.3.4	iOS	51.4876125	73.588584	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
96812436	3.1.29	iOS	28.057742	-81.253891	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
47631988	2.1.23	iOS	48.842616	-66.847399	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
64811321	2.2.38	iOS	-11.1664985	171.128929	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
53911958	2.2.39	iOS	5.9886555	179.08221	de545711-3914-4458-8c11-b17b8dabb5e1									No No Yes Yes NULL
98568652	2.1.12	Android 5.6975445	37.819333		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
83587969	3.2.1	iOS	37.872972	-49.466494	de545711-3914-4458-8c11-b17b8dabb5e1									No No Yes Yes NULL
49512687	4.4.4	Android 18.586405	88.6866545		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
22724557	3.1.32	iOS	-69.16196	-69.081245	de545711-3914-4458-8c11-b17b8dabb5e1									No Yes No Yes NULL
86678389	2.4.29	Android -28.988465	-55.018245		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
68422739	3.1.21	iOS	14.5942888	-164.817167	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
85346419	3.3.36	Android 31.731629	11.394737		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
66719552	3.3.28	iOS	-69.3938215	-77.618869	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
11316378	1.4.11	Android 28.478736	158.8967		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
83156659	4.1.14	iOS	48.2892126	95.714311	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
29702837	1.1.12	Android 1.1578665	-152.366834		b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
35776187	1.3.3	iOS	29.3831835	-156.45416	b328297e-17ae-11eb-adc1-0242ac120002									No Yes No Yes NULL
44822486	2.1.12	iOS	-39.698884	19.972828	de545711-3914-4458-8c11-b17b8dabb5e1									No Yes No Yes NULL
72311629	1.2.25	iOS	-15.094244	-58.16225	de545711-3914-4458-8c11-b17b8dabb5e1									No Yes Yes Yes NULL
99713267	1.2.13	Android -63.011675	-168.394573		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes No Yes NULL
66968218	3.2.3	iOS	-71.376849	-141.492428	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes Yes Yes NULL
22783844	1.1.28	Android -25.972366	17.318069		e7bc5f52-1231-11eb-adc1-0242ac120002									No No No No NULL
92178756	3.3.29	Android -85.98985	-12.773762		de545711-3914-4458-8c11-b17b8dabb5e1									No Yes Yes Yes NULL
93617994	1.1.24	iOS	-85.7651895	-116.523285	de545711-3914-4458-8c11-b17b8dabb5e1									No Yes Yes Yes NULL
93846678	1.4.19	iOS	10.996818	-154.33489	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes Yes Yes NULL
65044856	3.4.38	Android -86.735896	-168.343227		e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes Yes Yes NULL
20086795	1.4.11	Android 44.376255	-15.33288		de545711-3914-4458-8c11-b17b8dabb5e1									No Yes Yes Yes NULL
86984724	4.3.15	Android -3.0758015	-97.024893		de545711-3914-4458-8c11-b17b8dabb5e1									No Yes No No NULL
91588778	3.4.5	Android -18.2869335	84.429953		de545711-3914-4458-8c11-b17b8dabb5e1									No No No Yes NULL
62561362	3.3.10	iOS	28.044871	-16.287224	e7bc5f52-1231-11eb-adc1-0242ac120002									No Yes Yes No NULL
84536540	4.4.12	iOS	36.1815295	-126.499135	b328297e-17ae-11eb-adc1-0242ac120002									No Yes 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
```