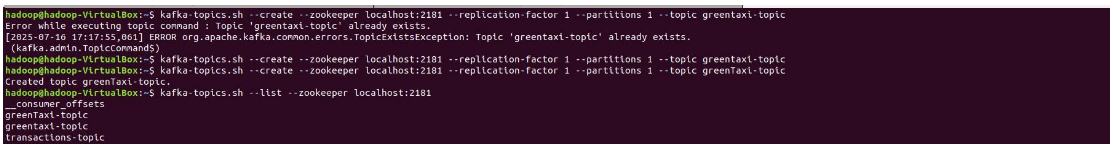
ASSIGNMENT -5 (Kafka + Spark)

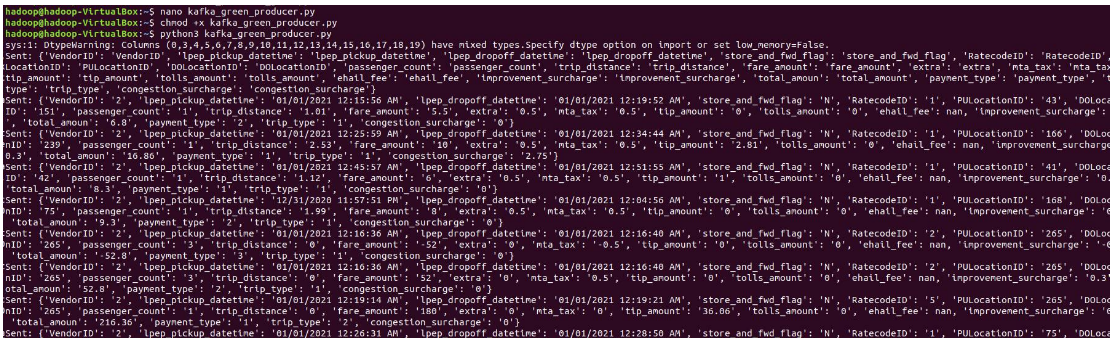
Sruthy Suji
23AD140
AI-DS C

1) Create a new topic – greentaxi-topic

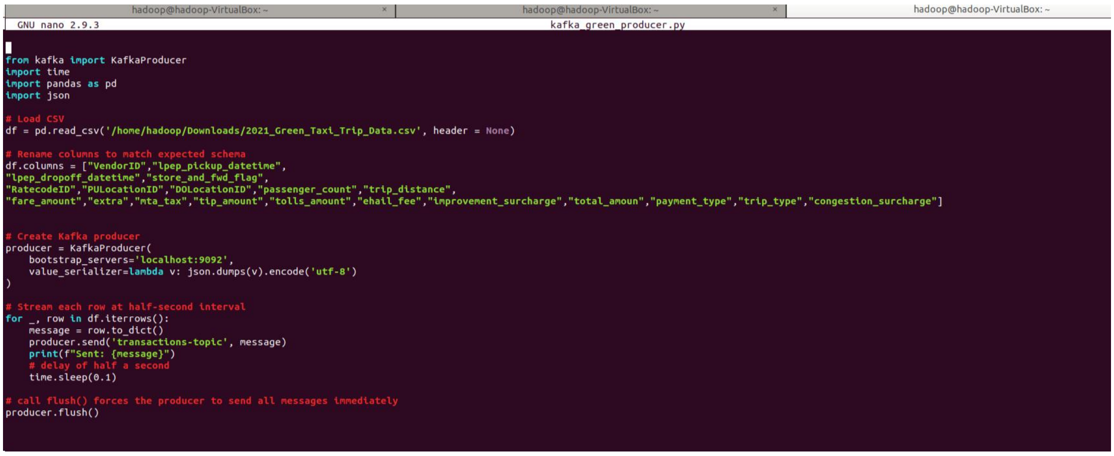


```
hadoop@hadoop-VirtualBox:~$ kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic greentaxi-topic
Error while executing topic command : Topic 'greentaxi-topic' already exists.
[2025-07-16 17:17:55,061] ERROR org.apache.kafka.common.errors.TopicExistsException: Topic 'greentaxi-topic' already exists.
 (kafka.admin.TopicCommand$)
hadoop@hadoop-VirtualBox:~$ kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic greentaxi-topic
hadoop@hadoop-VirtualBox:~$ kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic greenTaxi-topic
Created topic greenTaxi-topic.
hadoop@hadoop-VirtualBox:~$ kafka-topics.sh --list --zookeeper localhost:2181
__consumer_offsets
greenTaxi-topic
greentaxi-topic
transactions-topic
```

2) Stream Data @ 0.1 Sec Interval using Kafka_Taxi_Producer.py



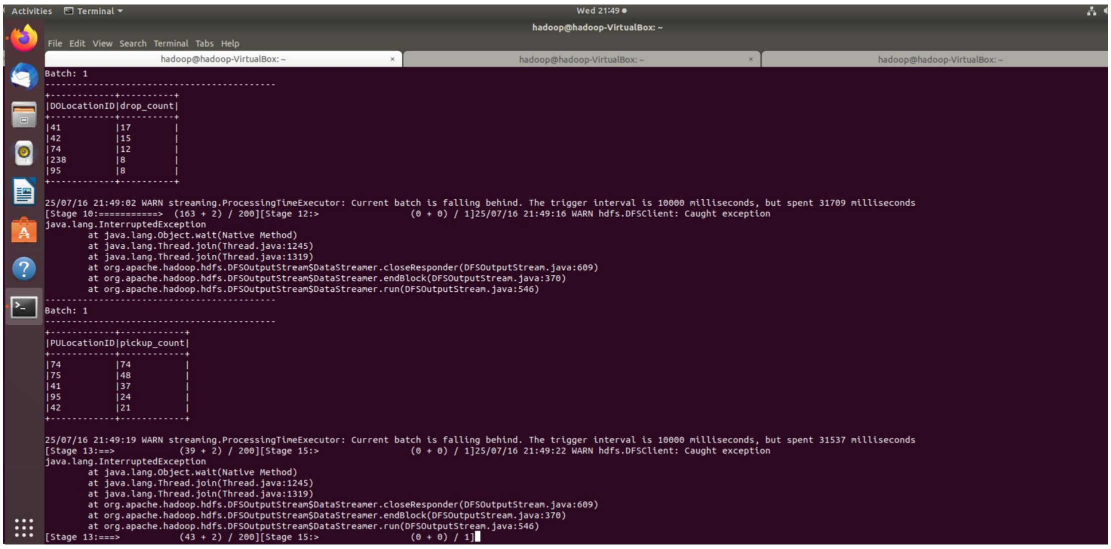
```
hadoop@hadoop-VirtualBox:~$ nano kafka_green_producer.py
hadoop@hadoop-VirtualBox:~$ chmod +x kafka_green_producer.py
hadoop@hadoop-VirtualBox:~$ python3 kafka_green_producer.py
sys:1: DtypeWarning: Columns (0,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19) have mixed types.Specify dtype option on import or set low_memory=False.
Sent: {'VendorID': 'VendorID', 'lpep_pickup_datetime': 'lpep_pickup_datetime', 'lpep_dropoff_datetime': 'lpep_dropoff_datetime', 'store_and_fwd_flag': 'store_and_fwd_flag', 'RatecodeID': 'RatecodeID',
LocationID': 'PULocationID', 'DOLocationID': 'DOLocationID', 'passenger_count': 'passenger_count', 'trip_distance': 'trip_distance', 'fare_amount': 'fare_amount', 'extra': 'extra', 'mta_tax': 'mta_tax
tip_amount': 'tip_amount', 'tolls_amount': 'tolls_amount', 'ehail_fee': 'ehail_fee', 'improvement_surcharge': 'improvement_surcharge', 'total_amoun': 'total_amount', 'payment_type': 'payment_type', 't
type': 'trip_type', 'congestion_surcharge': 'congestion_surcharge'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '01/01/2021 12:15:56 AM', 'lpep_dropoff_datetime': '01/01/2021 12:19:52 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '1', 'PULocationID': '43', 'DOLoca
ID': '151', 'passenger_count': '1', 'trip_distance': '1.01', 'fare_amount': '5.5', 'extra': '0.5', 'mta_tax': '0.5', 'tip_amount': '0', 'tolls_amount': '0', 'ehail_fee': nan, 'improvement_surcharge':
', 'total_amoun': '6.8', 'payment_type': '2', 'trip_type': '1', 'congestion_surcharge': '0'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '01/01/2021 12:25:59 AM', 'lpep_dropoff_datetime': '01/01/2021 12:34:44 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '1', 'PULocationID': '166', 'DOLoc
nID': '239', 'passenger_count': '1', 'trip_distance': '2.53', 'fare_amount': '10', 'extra': '0.5', 'mta_tax': '0.5', 'tip_amount': '2.81', 'tolls_amount': '0', 'ehail_fee': nan, 'improvement_surcharge
0.3', 'total_amoun': '16.86', 'payment_type': '1', 'trip_type': '1', 'congestion_surcharge': '2.75'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '01/01/2021 12:45:57 AM', 'lpep_dropoff_datetime': '01/01/2021 12:51:55 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '1', 'PULocationID': '41', 'DOLoca
ID': '42', 'passenger_count': '1', 'trip_distance': '1.12', 'fare_amount': '6', 'extra': '0.5', 'mta_tax': '0.5', 'tip_amount': '1', 'tolls_amount': '0', 'ehail_fee': nan, 'improvement_surcharge': '0.
 'total_amoun': '8.3', 'payment_type': '1', 'trip_type': '1', 'congestion_surcharge': '0'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '12/31/2020 11:57:51 PM', 'lpep_dropoff_datetime': '01/01/2021 12:04:56 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '1', 'PULocationID': '168', 'DOLoc
nID': '75', 'passenger_count': '1', 'trip_distance': '1.99', 'fare_amount': '8', 'extra': '0.5', 'mta_tax': '0.5', 'tip_amount': '0', 'tolls_amount': '0', 'ehail_fee': nan, 'improvement_surcharge': '
 'total_amoun': '9.3', 'payment_type': '2', 'trip_type': '1', 'congestion_surcharge': '0'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '01/01/2021 12:16:36 AM', 'lpep_dropoff_datetime': '01/01/2021 12:16:40 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '2', 'PULocationID': '265', 'DOLoc
nID': '265', 'passenger_count': '3', 'trip_distance': '0', 'fare_amount': '-52', 'extra': '0', 'mta_tax': '-0.5', 'tip_amount': '0', 'tolls_amount': '0', 'ehail_fee': nan, 'improvement_surcharge': '-
 'total_amoun': '-52.8', 'payment_type': '3', 'trip_type': '1', 'congestion_surcharge': '0'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '01/01/2021 12:16:36 AM', 'lpep_dropoff_datetime': '01/01/2021 12:16:40 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '2', 'PULocationID': '265', 'DOLoc
nID': '265', 'passenger_count': '3', 'trip_distance': '0', 'fare_amount': '52', 'extra': '0', 'mta_tax': '0.5', 'tip_amount': '0', 'tolls_amount': '0', 'ehail_fee': nan, 'improvement_surcharge': '0.3
otal_amoun': '52.8', 'payment_type': '2', 'trip_type': '1', 'congestion_surcharge': '0'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '01/01/2021 12:19:14 AM', 'lpep_dropoff_datetime': '01/01/2021 12:19:21 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '5', 'PULocationID': '265', 'DOLoc
nID': '265', 'passenger_count': '1', 'trip_distance': '0', 'fare_amount': '180', 'extra': '0', 'mta_tax': '0', 'tip_amount': '36.06', 'tolls_amount': '0', 'ehail_fee': nan, 'improvement_surcharge': '
 'total_amoun': '216.36', 'payment_type': '1', 'trip_type': '2', 'congestion_surcharge': '0'}
Sent: {'VendorID': '2', 'lpep_pickup_datetime': '01/01/2021 12:26:31 AM', 'lpep_dropoff_datetime': '01/01/2021 12:28:50 AM', 'store_and_fwd_flag': 'N', 'RatecodeID': '1', 'PULocationID': '75', 'DOLoca
```



hadoop@hadoop-VirtualBox: ~
hadoop@hadoop-VirtualBox: ~
hadoop@hadoop-VirtualBox: ~
GNU nano 2.9.3
kafka_green_producer.py

```python
from kafka import KafkaProducer
import time
import pandas as pd
import json

# Load CSV
df = pd.read_csv('/home/hadoop/Downloads/2021_Green_Taxi_Trip_Data.csv', header = None)

# Rename columns to match expected schema
df.columns = ["VendorID","lpep_pickup_datetime",
"lpep_dropoff_datetime","store_and_fwd_flag",
"RatecodeID","PULocationID","DOLocationID","passenger_count","trip_distance",
"fare_amount","extra","mta_tax","tip_amount","tolls_amount","ehail_fee","improvement_surcharge","total_amoun","payment_type","trip_type","congestion_surcharge"]


# Create Kafka producer
producer = KafkaProducer(
    bootstrap_servers='localhost:9092',
    value_serializer=lambda v: json.dumps(v).encode('utf-8')
)

# Stream each row at half-second interval
for _, row in df.iterrows():
    message = row.to_dict()
    producer.send('transactions-topic', message)
    print(f"Sent: {message}")
    # delay of half a second
    time.sleep(0.1)

# call flush() forces the producer to send all messages immediately
producer.flush()
```

3) Show Top 5 Trending Pickup Local - "PULocationID" [5] and Drop Location - "DOLocationID" [6]


Activities
Terminal
Wed 21:49
hadoop@hadoop-VirtualBox: ~
File
Edit
View
Search
Terminal
Tabs
Help
hadoop@hadoop-VirtualBox: ~
hadoop@hadoop-VirtualBox: ~
hadoop@hadoop-VirtualBox: ~

```
Batch: 1
-------------------------------------------
+------------+----------+
|DOLocationID|drop_count|
+------------+----------+
|41          |17        |
|42          |15        |
|74          |12        |
|238         |8         |
|95          |8         |
+------------+----------+

25/07/16 21:49:02 WARN streaming.ProcessingTimeExecutor: Current batch is falling behind. The trigger interval is 10000 milliseconds, but spent 31709 milliseconds
[Stage 10:===========>  (163 + 2) / 200][Stage 12:>                 (0 + 0) / 1]25/07/16 21:49:16 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1245)
        at java.lang.Thread.join(Thread.java:1319)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:609)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:370)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:546)
-------------------------------------------
Batch: 1
-------------------------------------------
+------------+------------+
|PULocationID|pickup_count|
+------------+------------+
|74          |74          |
|75          |48          |
|41          |37          |
|95          |24          |
|42          |21          |
+------------+------------+

25/07/16 21:49:19 WARN streaming.ProcessingTimeExecutor: Current batch is falling behind. The trigger interval is 10000 milliseconds, but spent 31537 milliseconds
[Stage 13:==>            (39 + 2) / 200][Stage 15:>                 (0 + 0) / 1]25/07/16 21:49:22 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1245)
        at java.lang.Thread.join(Thread.java:1319)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:609)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:370)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:546)
[Stage 13:===>           (43 + 2) / 200][Stage 15:>                 (0 + 0) / 1]
```