September 16, 2021 Prof. Lary Vea

Google Drive Link:

https://drive.google.com/drive/folders/1korQBh5uqhYpaVudq3xCPB0VurI7O0b ?usp=sharing

PRE-REQUISITE INSTALLATION FOR THE QUIZ

File version used for this project:

Java: java jdk 1.8.0 111

Python: python 2.7.18

Zookeeper: apache zookeeper 3.7.0

Cassandra: apache Cassandra 3.11.11

Spark: Spark 2.4.6-bin-hadoop 2.7

Hadoop: Hadoop 2.7.2

Instructions: Make a working Big Data Streaming Pipeline and Integration Platform (please refer to the uploaded architecture).

1. Make at least 10 instances of the producer that I uploaded. Make the sensor names from sensor_01 to sensor_10.

To make 10 instances, we must first create topic in our kafka. First we must navigate in our kafka window folder using the command "cd C:\opt\kafka_2.12-2.8.0\bin\windows", then using the command "kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic count" we will create the topic count in our kafka.

```
C:\Users\Ivan Gil Mercano>cd C:\opt\kafka_2.12-2.8.0\bin\windows
C:\opt\kafka_2.12-2.8.0\bin\windows>kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partit
ions 1 --topic count
Created topic count.
```

The topic and server should be the same in both virtual_consumer and virtual_producer. Below I paste a snippet of the topic and server code for virtual_producer.

```
KAFKA_TOPIC_NAME_CONS = "counts" #topic
KAFKA_BOOTSTRAP_SERVERS_CONS = '127.0.0.1:9092'

consumer = KafkaConsumer(
    'counts', #cctv_vehicle_counts
    bootstrap_servers = ['127.0.0.1:9092'],
    auto_offset_reset = 'latest',
    enable_auto_commit = True
)
```

After setting up the topic for the consumer and producer. I copy and pasted the file virtual_consumer 10 times and rename the file from virtual_consumer_1 to virtual_consumer_10.

idea .idea	16/09/2021 3:04 am	File folder	
checkpoint	16/09/2021 2:41 am	File folder	
iar jar	16/09/2021 2:32 am	File folder	
venv	16/09/2021 2:34 am	File folder	
data_streaming.py	16/09/2021 3:39 am	Python File	4 KB
🏧 jar-20210915T180010Z-001.zip	16/09/2021 2:01 am	WinRAR ZIP archive	9,352 KB
€ main.py	16/09/2021 2:33 am	Python File	1 KB
₽ virtual_consumer.py	16/09/2021 2:18 am	Python File	1 KB
₽ virtual_producer _2.py	12/09/2021 4:10 pm	Python File	3 KB
₽ virtual_producer _3.py	12/09/2021 4:10 pm	Python File	3 KB
₽ virtual_producer _4.py	12/09/2021 4:10 pm	Python File	3 KB
₽ virtual_producer _5.py	12/09/2021 4:10 pm	Python File	3 KB
₽ virtual_producer _6.py	12/09/2021 4:11 pm	Python File	3 KB
R virtual_producer _7.py	12/09/2021 4:11 pm	Python File	3 KB
R virtual_producer _8.py	12/09/2021 4:11 pm	Python File	3 KB
R virtual_producer _9.py	12/09/2021 4:11 pm	Python File	3 KB
R virtual_producer _10.py	12/09/2021 4:11 pm	Python File	3 KB
→ virtual_producer_1.py	09/09/2021 6:42 pm	Python File	3 KB

After renaming each file, I renamed the sensor_id of each virtual producer according to the number indicated in the file name (ie. Virtual_producer_1 -> sensor_01). This process is the same for each virtual_producer file.

```
total = car + bus + truck + jeepney + bike + tryke + others

#send data

#send data

#sessage["timeuvid_id"] = str(time_uvid.utctime())

#sessage["lgu_code"] = '1200'

#sessage["sensor_id"] = [sensor_01']

#sessage["date_saved"] = str(date_today.strftime('%m/%d/%Y'))

#sessage["time_saved"] = str(date_today.strftime("%X"))

#sessage["total"] = total

#sessage["bus"] = bus

#sessage["bus"] = bus

#sessage["truck"] = truck

#sessage["jeepney"] = jeepney

#sessage["bus"] = bike

#sessage["tryke"] = tryke

#sessage["tryke"] = others
```

```
#send data
message["timeuvid_id"] = str(time_uvid.utctime())
message["lgu_code"] = '1200'

message["date_saved"] = 'sensor_03'

message["time_saved"] = str(date_today.strftime('%m/%d/%Y'))
message["total"] = total
message["car"] = car
message["bus"] = bus
message["truck"] = truck
message["ieepney"] = jeepney
message["bike"] = bike
message["tryke"] = tryke
message["tryke"] = tryke
message["others"] = others
```

```
message["timeuuid_id"] = str(time_uuid.utctime())

message["lgu_code"] = '1200'

message["sensor_id"] = 'sensor_04'

message["date_saved"] = str(date_today.strftime('%m/%d/%Y'))

message["time_saved"] = str(date_today.strftime("%X"))

message["total"] = total

message["car"] = car

message["bus"] = bus

message["truck"] = truck

message["truck"] = jeepney

message["jeepney"] = jeepney

message["bike"] = bike

message["tryke"] = tryke

message["tryke"] = tryke

message["others"] = others
```

```
message["timeuuid_id"] = str(time_uuid.utctime())

message["lgu_code"] = '1200'

message["sensor_id"] = 'sensor_05'

message["date_saved"] = str(date_today.strftime('%m/%d/%Y'))

message["time_saved"] = str(date_today.strftime("%X"))

message["total"] = total

message["car"] = car

message["bus"] = bus

message["truck"] = truck

message["jeepney"] = jeepney

message["jeepney"] = jeepney

message["tryke"] = tryke

message["tryke"] = tryke

message["others"] = others
```

```
message["timeuuid_id"] = str(time_uuid.utctime())

message["lgu_code"] = '1200'

message["sensor_id"] = 'sensor_06'

message["date_saved"] = str(date_today.strftime('%m/%d/%Y'))

message["time_saved"] = str(date_today.strftime("%X"))

message["total"] = total

message["car"] = car

message["bus"] = bus

message["truck"] = truck

message["truck"] = jeepney

message["jeepney"] = jeepney

message["bike"] = bike

message["tryke"] = tryke

message["tryke"] = tryke

message["others"] = others
```

```
message["timeuuid_id"] = str(time_uuid.utctime())

message["lgu_code"] = '1200'

message["sensor_id"] = 'sensor_07'

message["date_saved"] = str(date_today.strftime('%m/%d/%Y'))

message["time_saved"] = str(date_today.strftime("%X"))

message["total"] = total

message["car"] = car

message["bus"] = bus

message["truck"] = truck

message["truck"] = jeepney

message["bike"] = bike

message["tryke"] = tryke

message["others"] = others
```

```
### virtual_producer_8.py \timessage["timeuvid_id"] = str(time_uvid.utctime())

### description of the producer_8.py \timessage["timeuvid_id"] = str(time_uvid.utctime())

#### description of the producer_8.py \timessage["duv_code"] = '1200'

### description of the producer_8.py \timessage["sensor_id"] = '1200'

### description of the producer_8.py \timessage["sensor_id"] = '1200'

### description of the producer_8.py \timessage["sensor_id"] = str(date_today.strftime("%m/%d/%Y'))

### description of the producer_8.py \timessage["models average of the producer_8.py \timessage["models average of the producer_8.py \timessage["sensor_id"] = str(date_today.strftime("%X"))

### description of the producer_8.py \timessage["models average of the producer_9.py \timessage["models average of the producer_8.py \timessage["models average of the producer_9.py \timessage["models average of the pro
```

```
message["timeuuid_id"] = str(time_uuid.utctime())

message["lgu_code"] = '1200'

message["sensor_id"] = [sensor_09']

message["date_saved"] = str(date_today.strftime('%m/%d/%Y'))

message["time_saved"] = str(date_today.strftime("%X"))

message["total"] = total

message["car"] = car

message["bus"] = bus

message["truck"] = truck

message["jeepney"] = jeepney

message["jeepney"] = jeepney

message["bike"] = bike

message["tryke"] = tryke

message["others"] = others
```

```
message["timeuvid_id"] = str(time_uvid.utctime())

message["tyu_code"] = '1200'

message["sensor_id"] = 'sensor_10'

message["date_saved"] = str(date_today.strftime('%m/%d/%Y'))

message["time_saved"] = str(date_today.strftime("%X"))

message["total"] = total

message["car"] = car

message["bus"] = bus

message["truck"] = truck

message["truck"] = truck

message["jeepney"] = jeepney

message["bike"] = bike

message["tryke"] = tryke

message["tryke"] = tryke

message["others"] = others
```

Virtual_consumer

```
tivirtual_consumer.py ×

# -*- coding: utf-8 -*-

from kafka import KafkaConsumer

consumer = KafkaConsumer(
    'counts', #cctv_vehicle_counts
    bootstrap_servers = ['127.0.0.1:9092'],
    auto_offset_reset = 'latest',
    enable_auto_commit = True

    )

for message in consumer:
    message = message.value

print(message)
```

2. Run these producers simultaneously so that the data produced will be sent to a topic in Kafka.

To properly run virtual producer and consumer, we must first ensure that both zookeeper and kafka are running in the background.

To run the zookeeper, I used the command "zkserver" in cmd.

```
C:\Users\Ivan Gil Mercano>zkserver
C:\Users\Ivan Gil Mercano>call "C:\Progra~1\Java\jdk1.8.0_111"\bin\java "-Dzookeeper.log.dir=C:\opt\zookeeper\bin\..\log
s" "-Dzookeeper.root.logger=INFO,CONSOLE" "-Dzookeeper.log.file=zookeeper-Ivan Gil Mercano-server-IVAN-PC.log" "-XX:+Hea
pDumpOnOutOfMemoryError" "-XX:OnOutOfMemoryError=cmd /c taskkill /pid %%p /t /f" -cp "C:\opt\zookeeper\bin\..\build\clas ses;C:\opt\zookeeper\bin\..\build\lib\*;C:\opt\zookeeper\bin\..\taskkill /pid %%p /t /f" -cp "C:\opt\zookeeper\bin\..\build\lib\*;C:\opt\zookeeper\bin\..\taskkill /pid %%p /t /f" -cp "C:\opt\zookeeper\bin\..\build\lib\*;C:\opt\zookeeper\bin\..\taskkill /f" -cp "C:\opt\zookeeper\bin\..\lib\*;C:\opt\zookeeper\bin\..\taskkill /f" -cp "C:\opt\zookeeper\bin\..\lib\*;C:\opt\zookeeper\bin\..\conf\zoo.cfg" 2021-09-16 18:35:44,987 [myid:] - INFO [main:QuorumPeerConfig@174] - Reading configuration from: C:\opt\zookeeper\bin\..
.\conf\zoo.cfg
2021-09-16 18:35:45,009 [myid:] - WARN [main:VerifyingFileFactory@65] - C:optzookeeperdata is relative. Prepend .\ to i
2021-09-16 18:35:45,009 [my10:] - WARN
ndicate that you're sure!
2021-09-16 18:35:45,021 [myid:] - INFO
2021-09-16 18:35:45,023 [myid:] - INFO
2021-09-16 18:35:45,023 [myid:] - INFO
2021-09-16 18:35:45,024 [myid:] - INFO
                                                                               [main:QuorumPeerConfig@444] - clientPortAddress is 0.0.0.0:2181
                                                                               [main:QuorumPeerConfig@448] - secureClientPort is not set
[main:QuorumPeerConfig@464] - observerMasterPort is not se
                                                                                                                                      - observerMasterPort is not set
                                                                               [main:QuorumPeerConfig@481] - metricsProvider.className is org.apache.zookeeper.
 metrics.impl.DefaultMetricsProvider
2021-09-16 18:35:45,025 [myid:] - INFO
2021-09-16 18:35:45,026 [myid:] - INFO
2021-09-16 18:35:45,026 [myid:] - INFO
                                                                               [main:DatadirCleanupManager@78] - autopurge.snapRetainCount set to 3
[main:DatadirCleanupManager@79] - autopurge.purgeInterval set to 0
[main:DatadirCleanupManager@101] - Purge task is not scheduled.
2021-09-16 18:35:45,026 [myid:] - WARN
                                                                               [main:QuorumPeerMain@139] - Either no config or no quorum defined in config, run
ning in standalone mode
2021-09-16 18:35:45,029 [myid:] - INFO [main:ManagedUtil@44] - Log4j 1.2 jmx support found and enabled.
2021-09-16 18:35:45,078 [myid:] - INFO [main:QuorumPeerConfig@174] - Reading configuration from: C:\opt\zookeeper\bin\.
.\conf\zoo.cfg
2021-09-16 18:35:45,079 [myid:] - WARN [main:VerifyingFileFactory@65] - C:optzookeeperdata is relative. Prepend .\ to i
2021-09-16 18:35:45,079 [myid:] - INFO [main:QuorumPeerConfig@444] - clientPortAddress is 0.0.0.0:2181
```

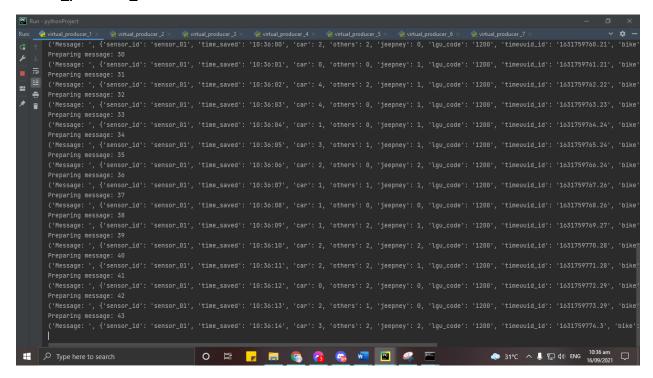
For the kafka server, we must first navigate to our kafka path yung "cd C:\opt\kafka_2.12-2.8.0" then after that, we must run the command ".\bin\windows\kafka-server-start.bat .\config\server.properties".

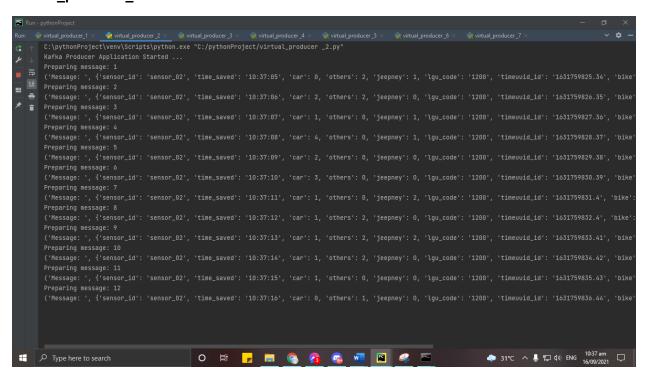
```
C:\Users\Ivan Gil Mercano>cd C:\opt\kafka_2.12-2.8.0

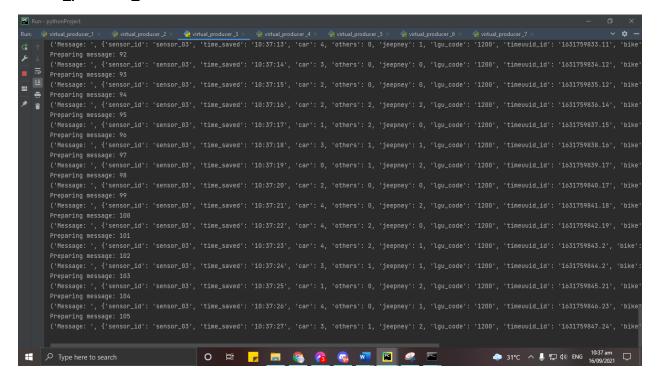
C:\opt\kafka_2.12-2.8.0>.\bin\windows\kafka_server-start.bat .\config\server.properties
[2021-09-16 18:37:45,675] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistratio n$)
[2021-09-16 18:37:46,190] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TL S renegotiation (org.apache.zookeeper.common.X509Util)
[2021-09-16 18:37:46,209] INFO Starting (kafka.server.KafkaServer)
[2021-09-16 18:37:46,270] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2021-09-16 18:37:46,300] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zooke eper.ZooKeeperClient)
[2021-09-16 18:37:46,448] INFO Client environment:zookeeper.version=3.5.9-83df9301aa5c2a5d284a9940177808c01bc35cef, buil to n01/06/2021 20:03 GMT (org.apache.zookeeper.ZooKeeper)
[2021-09-16 18:37:46,449] INFO Client environment:java.version=1.8.0 111 (org.apache.zookeeper.ZooKeeper)
[2021-09-16 18:37:46,449] INFO Client environment:java.version=1.8.0 111 (org.apache.zookeeper.ZooKeeper)
[2021-09-16 18:37:46,449] INFO Client environment:java.vendor=Oracle Corporation (org.apache.zookeeper.ZooKeeper)
[2021-09-16 18:37:46,449] INFO Client environment:java.vendor=Oracle Corporation (org.apache.zookeeper.ZooKeeper)
[2021-09-16 18:37:46,449] INFO Client environment:java.loms=C:\Progna-1\Java\jdk1.8.0_111\jre (org.apache.zookeeper.ZooKeeper)
[2021-09-16 18:37:46,449] INFO Client e
```

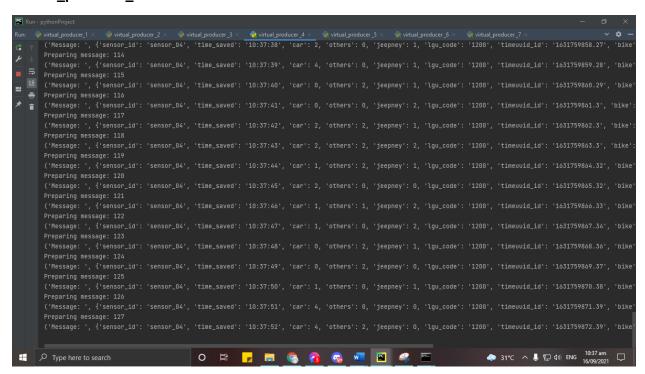
After that ensuring that the zkserver and Kafka are running we can now run each producer file and consumer file.

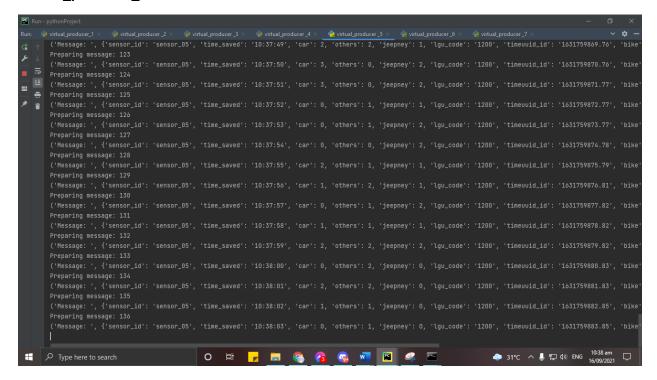
Virtual producer 1

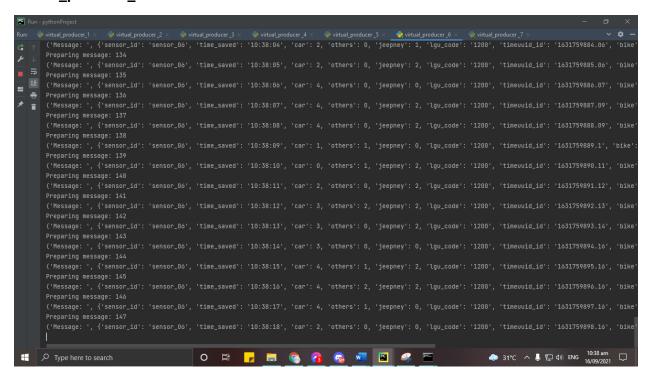


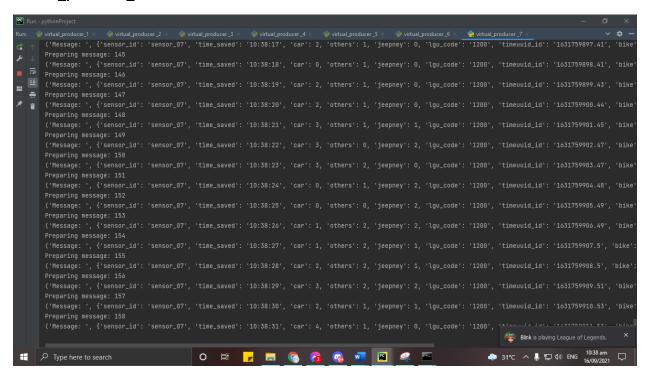


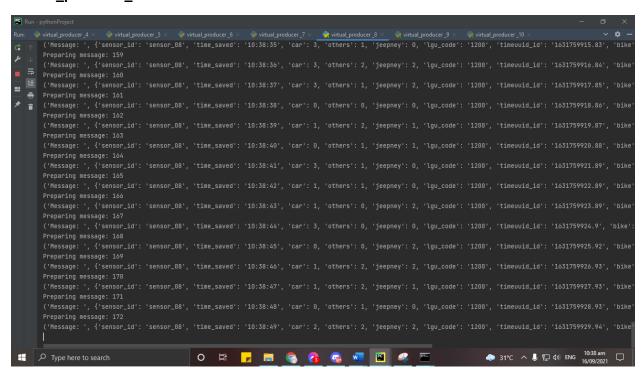


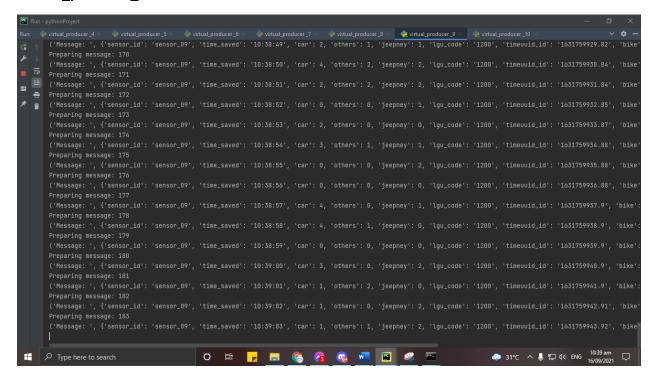


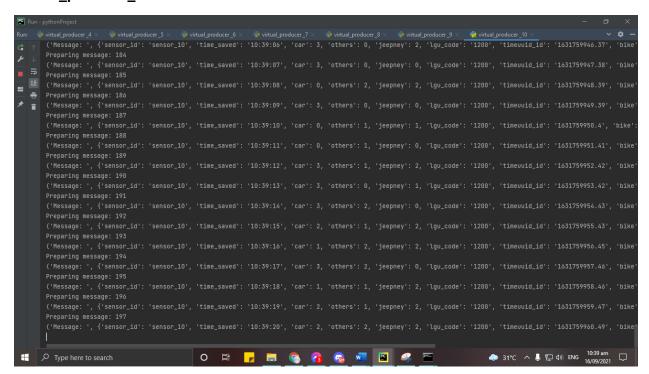




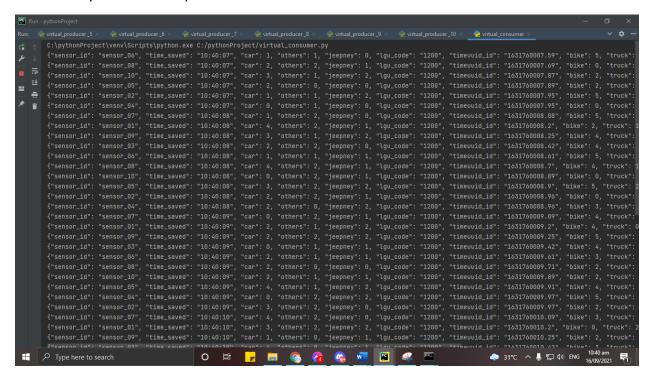








3. Check if Kafka is properly working by running the consumer. The consumer should display the data sent by the different producers.



4. Create a database and a table in Apache Hive (&/or Cassandra) that will be used as a storage of the data from the producers.

Creating keyspace count db and describing it to see if it is created properly.

```
© Command Prompt-py -2-m cqlsh

cqlsh> CREATE KEYSPACE IF NOT EXISTS Count_db WITH REPLICATION = {'class':'SimpleStrategy','replication_factor':'1'};

cqlsh> DESCRIBE KEYSPACE Count_db;

CREATE KEYSPACE count_db WITH replication = {'class': 'SimpleStrategy', 'replication_factor': '1'} AND durable_writes = true;
```

Creating table count_tbl and describing it to see if the table is created and the columns are properly named

```
os. Command Prompt - py -2 -m cqlsh
                                                                                                                                                                              cqlsh> CREATE TABLE IF NOT EXISTS Count_db.count_tbl(
   ... timeuuid_id text PRIMARY KEY,
    ... lgu_code text,
... sensor_id text,
    ... date_saved text,
... time_saved text,
... total int,
   ... car int,
... bus int,
... truck int,
   ... jeepney int,
... bike int,
   ... tryke int,
... others int
... );
cqlsh> DESCRIBE TABLE Count_db.count_tbl;
CREATE TABLE count_db.count_tbl (
timeuuid_id text PRIMARY KEY,
     bike int,
     bus int,
     date_saved text,
jeepney int,
lgu_code text,
     others int,
sensor_id text,
      time_saved text,
      total int,
```

I also tried to insert data manually in the table to see the table with values inside

After checking, I drop the values of the table I manually inserted using *TRUNCATE Count_db.count_tbl;* command.

```
cqlsh> TRUNCATE Count_db.count_tbl;
cqlsh> select * from Count_db.count_tbl;
timeuuid_id | bike | bus | car | date_saved | jeepney | lgu_code | others | sensor_id | time_saved | total | truck | tryke

(0 rows)
cqlsh> ___
```

5. Create another Python program that fetches the latest batch of data that arrived in Kafka and then sends (saves) this batch of data to the Hive (&/or Cassandra).

```
print("Printing before Cassandra table save: " + str(epoc_id))

| print("Printing before Cassandra table save: " + str(epoc_id))
| print("Brinting before Cassandra table save: " + str(epoc_id))
| print("Bata Processing Application Started ...")
| print("Data Processing Application Started ...")
| print(time.strftime("%Y-%m-%d %H:%M:%S"))
| #Connect to Spark | spar
```

Below are the jre files I used in my python file.

Name	Date modified	Туре	Size
isr166e-1.1.0.jar	12/09/2021 8:32 am	Executable Jar File	61 KB
kafka-clients-1.1.0.jar	12/09/2021 8:32 am	Executable Jar File	1,682 KB
spark-cassandra-connector-2.4.0-s_2.11.jar	12/09/2021 8:33 am	Executable Jar File	8,341 KB
spark-sql-kafka-0-10_2.11-2.4.5.jar	12/09/2021 8:32 am	Executable Jar File	523 KB

To run the python file without error, we must first open our Cassandra in our background by simply typing the command "Cassandra" in cmd.

```
C:\Users\Ivan Gil Mercano>cassandra

Detected powershell execution permissions. Running with enhanced startup scripts.

*****

WARNING! Automatic page file configuration detected.

It is recommended that you disable swap when running Cassandra
for performance and stability reasons.

***

WARNING! Detected a power profile other than High Performance.
Performance of this node will suffer.
Modify conf\cassandra.env.ps1 to suppress this warning.

***

C:\Users\Ivan Gil Mercano>CompilerOracle: dontinline org/apache/cassandra/db/Columns;I)Lorg/apache/cassandra/db/Columns;
CompilerOracle: dontinline org/apache/cassandra/db/Columns;Serializer.serializeLargeSubset (Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;I)Lorg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;Serializer.serializeLargeSubsetSize (Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;Serializer.serializeLargeSubsetSize (Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;Serializer.serializeLargeSubsetSize (Ljava/util/Collection;ILorg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns;IIOrg/apache/cassandra/db/Columns
```

After making sure that the Cassandra is running properly, we can now run the python file for data streaming.

Below is a snippet of the running data streaming python file.

```
C:\pythonProject\venv\Scripts\python.exe C:/pythonProject/data_streaming.py
Data Processing Application Started ...
2021-09-16 11:05:41
Data Processing Application Started ...
Printing Schema of orders_df:
 |-- key: binary (nullable = true)
 |-- value: binary (nullable = true)
 |-- topic: string (nullable = true)
 |-- partition: integer (nullable = true)
 |-- offset: long (nullable = true)
 |-- timestamp: timestamp (nullable = true)
 |-- timestampType: integer (nullable = true)
Printing epoc_id:
Printing before Cassandra table save: 13
Printing before Cassandra table save: 13
Printing epoc_id:
Printing before Cassandra table save: 14
Printing before Cassandra table save: 14
```

6. Check if the data were saved by querying the Hive (&/or Cassandra) table.

To check the if the data were save by Cassandra, I used the command "SELECT * FROM count_db.count_tbl;"

