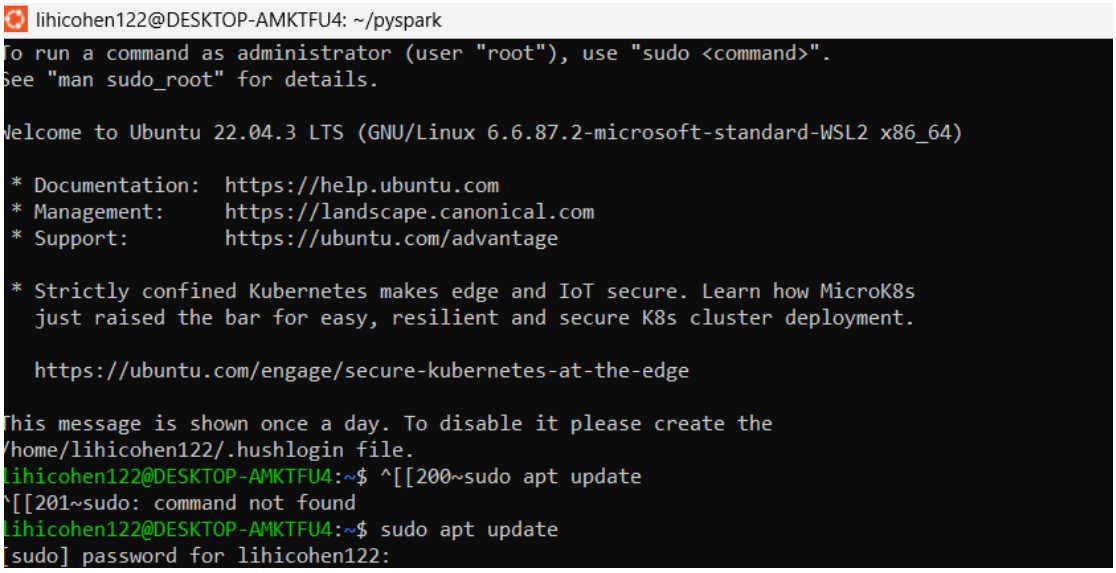
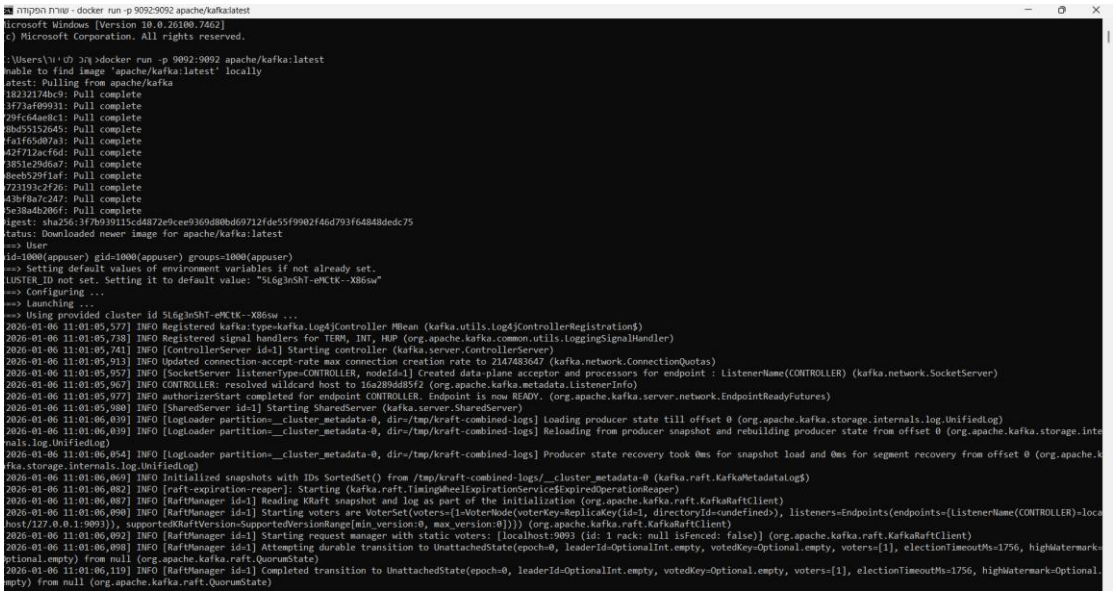
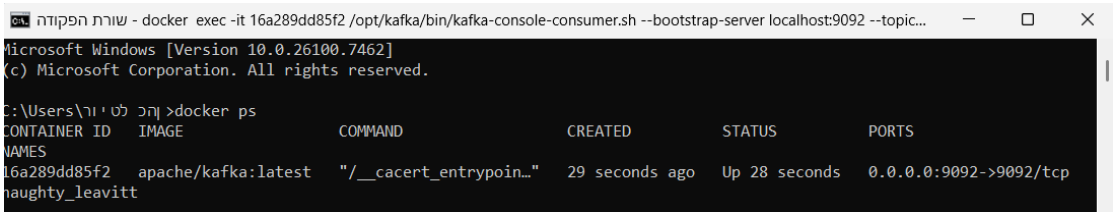
Working with ubuntu:



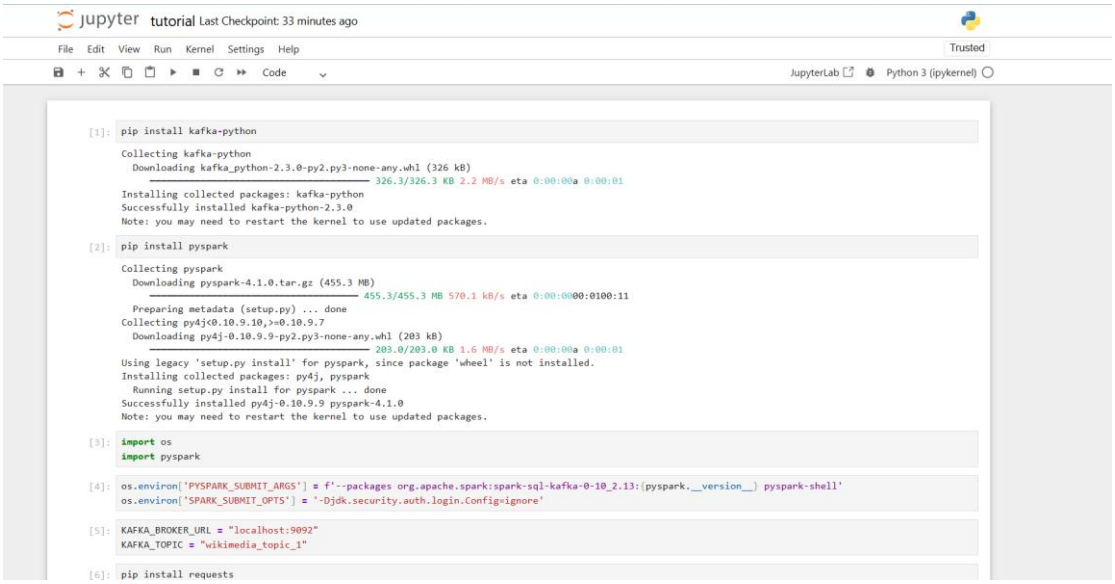
Docker:





```
C:\Users\לטיור והכ >docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-topics.sh --bootstrap-server localhost:9092 --list
wikimedia_topic_1

What's next?
  Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug 16a289dd85f2
  Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\לטיור והכ >docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-topics.sh --bootstrap-server localhost:9092 --descr
ibe --topic wikimedia_topic_1
Topic: wikimedia_topic_1        TopicId: Rr1x_ZjvSM2Vp1m7ZNg8Xg PartitionCount: 1       ReplicationFactor: 1    Configs:
 segment.bytes=1073741824
        Topic: wikimedia_topic_1        Partition: 0    Leader: 1       Replicas: 1     Isr: 1  Elr:    LastKnownElr:

What's next?
  Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug 16a289dd85f2
  Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\לטיור והכ >docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-console-consumer.sh --bootstrap-server localhost:90
92 --topic wikimedia_topic_1 [--from-beginning]
```

Jupiter (Tutorial):




```
[6]: pip install requests

Requirement already satisfied: requests in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (2.32.5)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests) (3.4.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests) (2.6.2)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests) (3.11)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests) (2026.1.4)
Note: you may need to restart the kernel to use updated packages.

[7]: pip install sseclient

Collecting sseclient
  Downloading sseclient-0.0.27.tar.gz (7.5 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: requests>=2.9 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (2.32.5)
Requirement already satisfied: six in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (1.17.0)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests>=2.9->sseclient) (2.6.2)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests>=2.9->sseclient)
(3.4.4)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests>=2.9->sseclient) (2026.1.
4)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicohen122/.venv3.11/lib/python3.11/site-packages (from requests>=2.9->sseclient) (3.11)
Using legacy 'setup.py install' for sseclient, since package 'wheel' is not installed.
Installing collected packages: sseclient
  Running setup.py install for sseclient ... done
Successfully installed sseclient-0.0.27
Note: you may need to restart the kernel to use updated packages.

[8]: import threading
import time
from kafka import KafkaConsumer
from kafka import KafkaProducer
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, from_json, expr
from pyspark.sql.types import StructType, StringType, IntegerType
from pyspark.streaming import StreamingContext
```

[9]:
```
producer = KafkaProducer(bootstrap_servers=KAFKA_BROKER_URL)
```

[10]:
```
import requests
from sseclient import SSEClient
URL = 'https://stream.wikimedia.org/v2/stream/recentchange'
headers = {
    "User-Agent": "MyLocalScript",
    "Authorization": "eyJ0eXAiOiJKV1QiLCJhbGciOiJSUzI1NiJ9.eyJhdWQiOiI3ODk5N2EzNzllMmE2NzIwNTJlNTg4NWE4NjkxZTBkMiIsImp0aSI6IjIxZTE2MTJkYzZiOTkyY2U4YmU0Z
}
def relay():
    events = SSEClient(URL, headers=headers, timeout=30)
    for i in range(100):
        for event in events:
            if event.event == 'message' and event.data != None:
                message = event.data.encode("utf-8")
                producer.send(KAFKA_TOPIC, value=message)
                break
threading.Thread(target=relay).start()
```

[11]:
```
spark = SparkSession.builder \
    .appName("PySpark-jupyter-streaming") \
    .config("spark.jars.packages", "org.apache.spark:spark-sql-kafka-0-10_2.13:{pyspark.__version__}") \
    .config("spark.sql.streaming.checkpointLocation", "./checkpoint") \
    .getOrCreate()
```

```
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
26/01/06 13:23:45 WARN Utils: Your hostname, DESKTOP-AMKTFU4, resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead (on interface lo)
26/01/06 13:23:45 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
:: loading settings :: url = jar:file:/home/lihicohen122/.venv3.11/lib/python3.11/site-packages/pyspark/jars/ivy-2.5.3.jar!/org/apache/ivy/core/setting
s/ivysettings.xml
Ivy Default Cache set to: /home/lihicohen122/.ivy2.5.2/cache
```

```
downloading https://repo1.maven.org/maven2/org/lz4/lz4-java/1.8.0/lz4-java-1.8.0.jar ...
        [SUCCESSFUL ] org.lz4#lz4-java;1.8.0!lz4-java.jar (197ms)
downloading https://repo1.maven.org/maven2/org/xerial/snappy/snappy-java/1.1.10.8/snappy-java-1.1.10.8.jar ...
        [SUCCESSFUL ] org.xerial.snappy#snappy-java;1.1.10.8!snappy-java.jar(bundle) (505ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/2.0.17/slf4j-api-2.0.17.jar ...
        [SUCCESSFUL ] org.slf4j#slf4j-api;2.0.17!slf4j-api.jar (64ms)
downloading https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-client-api/3.4.2/hadoop-client-api-3.4.2.jar ...
        [SUCCESSFUL ] org.apache.hadoop#hadoop-client-api;3.4.2!hadoop-client-api.jar (4567ms)
:: resolution report :: resolve 15623ms :: artifacts dl 17467ms
        :: modules in use:
        com.google.code.findbugs#jsr305;3.0.0 from central in [default]
        org.apache.commons#commons-pool2;2.12.1 from central in [default]
        org.apache.hadoop#hadoop-client-api;3.4.2 from central in [default]
        org.apache.hadoop#hadoop-client-runtime;3.4.2 from central in [default]
        org.apache.kafka#kafka-clients;3.9.1 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.13;4.1.0 from central in [default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.13;4.1.0 from central in [default]
        org.lz4#lz4-java;1.8.0 from central in [default]
        org.scala-lang.modules#scala-parallel-collections_2.13;1.2.0 from central in [default]
        org.slf4j#slf4j-api;2.0.17 from central in [default]
        org.xerial.snappy#snappy-java;1.1.10.8 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   11  |   11  |   11  |   0   ||   11  |   11  |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-35a7d01e-2e4c-45b2-aece-40f5b47ff6d0
        confs: [default]
        11 artifacts copied, 0 already retrieved (62936kB/107ms)
26/01/06 13:24:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
```

[12]:
```
kafka_df = spark.readStream \
  .format("kafka") \
  .option("kafka.bootstrap.servers", KAFKA_BROKER_URL) \
  .option("subscribe", KAFKA_TOPIC) \
  .option("startingOffsets", "earliest") \
  .load()
```

[13]:
```
schema = StructType() \
    .add("id", IntegerType()) \
    .add("type", StringType()) \
    .add("comment", StringType()) \
    .add("user", StringType()) \
    .add("title", StringType()) \
    .add("server_name", StringType())

# Transform data to dataframe of json format
parsed_df = kafka_df.selectExpr("CAST(value AS STRING)") \
    .select(from_json(col("value"), schema).alias("data")) \
    .select("data.*")
```

[14]:
```
parsed_df.writeStream \
    .outputMode("append") \
    .format("console") \
    .start()
```

```
26/01/06 13:26:14 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
```

```
26/01/06 13:26:14 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
```

[14]:
```
<pyspark.sql.streaming.query.StreamingQuery at 0x72b42f4c6590>
```

```
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+----------+--------------------+---------------+--------------------+--------------------+
|        id|      type|             comment|           user|               title|         server_name|
+----------+----------+--------------------+---------------+--------------------+--------------------+
|      NULL|      NULL|                NULL|           NULL|                NULL|                NULL|
| 341730927|categorize|      [[:José Mingoranc...|     Werthercito|      Categoría:Falleci...|    es.wikipedia.org|
|  17682112|categorize|      [[:Gleitufer]] vo...|Alexander Gamauf|      Kategorie:Wartung...|   de.wiktionary.org|
|      NULL|      edit|      /* wbeditentity-u...|      Rkieferbot|      File:Sori - Chies...|commons.wikimedia...|
|      NULL|      edit|      /* wbcreateclaim-...|         Chabe01|      File:Plaque Allée...|commons.wikimedia...|
|      NULL|      edit|      /* wbsetclaim-upd...|          Sporti|              Q137710176|    www.wikidata.org|
| 133285614|       log|      Bot: Mass deletin...|       Ladsgroup|      Thảo luận Thành v...|    vi.wikipedia.org|
| 142109988|      edit|      /* Verb */ litera...|         Jberkel|                    gall|   en.wiktionary.org|
|  32084861|      edit|        /* Библиография */|      Al Silonov|                  скарн|   ru.wiktionary.org|
|      NULL|      edit|      /* wbeditentity-u...|      M2k~dewiki|              Q137673651|    www.wikidata.org|
|  11705348|       new|      "[[:en:Special:Re...|     Wikipidiya1|                  ताप्ती|    hi.wikipedia.org|
|1983163400|categorize|      [[:Wikipedia:Arti...|       MatrixBot|      Category:Noindexe...|    en.wikipedia.org|
|  70248355|categorize|      Nampiana tamin'ny...|        HJAOVELO|      Sokajy:Anarana io...|   mg.wiktionary.org|
|      NULL|      edit|      /* wbcreateclaim-...|         Chabe01|      File:Plaque Allée...|commons.wikimedia...|
|      NULL|      edit|      /* wbsetsitelink-...|    Robinscarlet|                Q49510664|    www.wikidata.org|
|      NULL|      edit|      /* wbsetqualifier...|       Sakretsu|                  Q40917|    www.wikidata.org|
|  91406231|      edit|붓: 문자열 변경 ((== *외...|       A.TedBot|                 키케온|    ko.wikipedia.org|
|      NULL|       log|                        |  Paul Gascoigne|              Niky Savage|    it.wikipedia.org|
|      NULL|      edit|      /* wbeditentity-u...|   SchlurcherBot|File:01.03 總統「洲美三...|commons.wikimedia...|
|   2420052|      edit|      додав посилання н...|         Balakun|      Вікіджерела:Кален...|   uk.wikisource.org|
+----------+----------+--------------------+---------------+--------------------+--------------------+
only showing top 20 rows
```

[15]:
```
parsed_df.createOrReplaceTempView("parsed_df")

spark.sql("select user, count(*) as count from parsed_df group by user") \
.writeStream \
.outputMode("complete") \
.format("console") \
.start()
```

```
26/01/06 13:26:53 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
```

[15]:
```
<pyspark.sql.streaming.query.StreamingQuery at 0x72b42d740650>
```

```
26/01/06 13:26:54 WARN MicroBatchExecution: Disabling AQE since AQE is not supported in stateful workloads.

-------------------------------------------
Batch: 0
-------------------------------------------
+---------------+-----+
|           user|count|
+---------------+-----+
|  The Anomebot2|    3|
|    Quetzal1964|    1|
|       A.TedBot|    1|
|        RussBot|    1|
|       Bergenga|    1|
|      MatrixBot|    2|
|           NULL|    1|
|    Marklar2007|    1|
|       HJAOVELO|    1|
|          Alexf|    1|
|  SwarabaktiBot|    3|
|         巫宗霖|    1|
|Alexander Gamauf|    1|
|     Rkieferbot|    3|
|        Balakun|    1|
| Paul Gascoigne|    1|
|    Werthercito|    1|
|        Ivtorov|    1|
|       Sakretsu|    3|
|  17387349L8764|    1|
+---------------+-----+
only showing top 20 rows
```

[16]:
```
spark.sql("select type, count(*) as count from parsed_df group by type") \
.writeStream \
.outputMode("complete") \
.format("console") \
.start()
```

```
26/01/06 13:28:57 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
```

[16]:
```
<pyspark.sql.streaming.query.StreamingQuery at 0x72b42d7459d0>
```

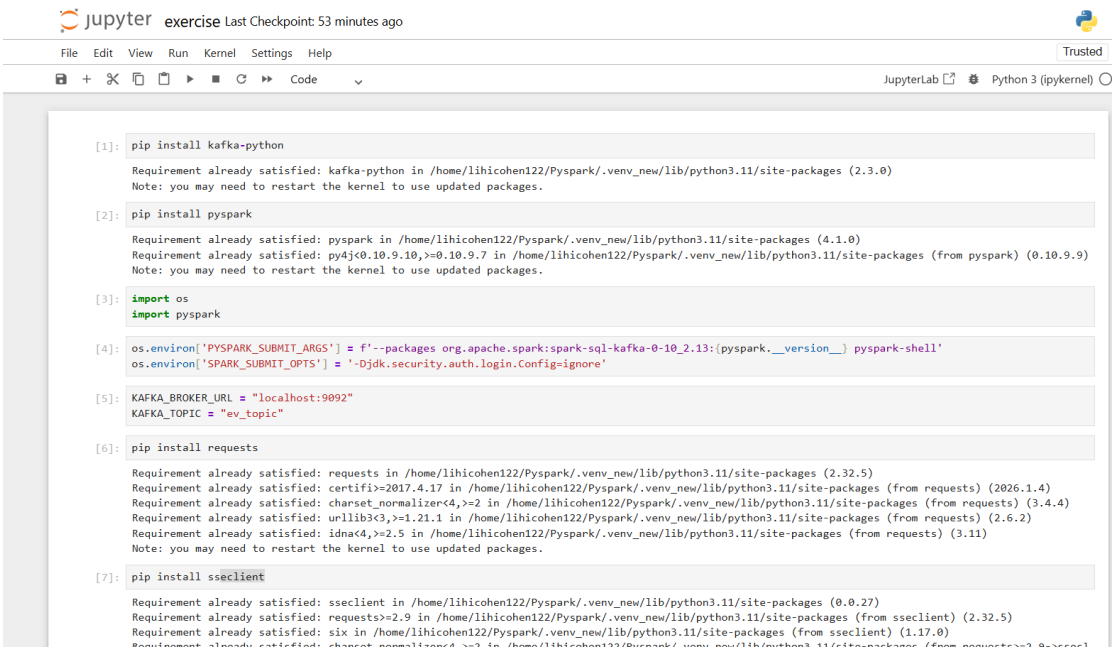
```
26/01/06 13:28:57 WARN MicroBatchExecution: Disabling AQE since AQE is not supported in stateful workloads.

-------------------------------------------
Batch: 0
-------------------------------------------
+----------+-----+
|      type|count|
+----------+-----+
|       new|    2|
|       log|    9|
|      NULL|    1|
|      edit|   44|
|categorize|   44|
+----------+-----+
```

[ ]:

Docker (Exercise):

```
C:\Users\לט'ור וכמ>docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-topics.sh --bootstrap-server localhost:9092 --list
__consumer_offsets
ev_topic
wikimedia_topic_1

What's next?
  Try Docker Debug for seamless, persistent debugging tools in any container or image → docker debug 16a289dd85f2
  Learn more at https://docs.docker.com/go/debug-cli/

C:\Users\לט'ור וכמ>
```

Jupiter (Exercise):



[7]:
```
pip install sseclient
```

```
Requirement already satisfied: sseclient in /home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages (0.0.27)
Requirement already satisfied: requests>=2.9 in /home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages (from sseclient) (2.32.5)
Requirement already satisfied: six in /home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages (from sseclient) (1.17.0)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages (from requests>=2.9->sseclient) (3.4.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages (from requests>=2.9->sseclient) (2.6.2)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages (from requests>=2.9->sseclient) (2026.1.4)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages (from requests>=2.9->sseclient) (3.11)
Note: you may need to restart the kernel to use updated packages.

Note: you may need to restart the kernel to use updated packages.
```

[8]:
```
import csv
import json
import time
from kafka import KafkaProducer

KAFKA_BROKER_URL = "localhost:9092"
KAFKA_TOPIC = "ev_topic"

producer = KafkaProducer(
    bootstrap_servers=KAFKA_BROKER_URL,
    value_serializer=lambda x: json.dumps(x).encode('utf-8')
)

file_path = '../input/Electric_Vehicle_Population_Data.csv'
print(f"Streaming data to {KAFKA_TOPIC}...")

with open(file_path, mode='r', encoding='utf-8') as csvfile:
    reader = csv.DictReader(csvfile)
    for row in reader:
        producer.send(KAFKA_TOPIC, value=row)

producer.flush()
```

```
Streaming data to ev_topic...
```

[9]:
```python
from pyspark.sql import SparkSession
import pyspark

# Initialize the Spark Session with Kafka support
spark = SparkSession.builder \
    .appName("EV-Streaming-Exercise") \
    .config("spark.jars.packages", f"org.apache.spark:spark-sql-kafka-0-10_2.13:{pyspark.__version__}") \
    .config("spark.sql.streaming.checkpointLocation", "./checkpoint") \
    .getOrCreate()
```

```
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
26/01/06 16:09:32 WARN Utils: Your hostname, DESKTOP-AMKTFU4, resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead (on interface lo)
26/01/06 16:09:32 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
:: loading settings :: url = jar:file:/home/lihicohen122/Pyspark/.venv_new/lib/python3.11/site-packages/pyspark/jars/ivy-2.5.3.jar!/org/apache/ivy/cor
e/settings/ivysettings.xml
Ivy Default Cache set to: /home/lihicohen122/.ivy2.5.2/cache
The jars for the packages stored in: /home/lihicohen122/.ivy2.5.2/jars
org.apache.spark#spark-sql-kafka-0-10_2.13 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-e0c83b49-b5c2-46a3-a2df-8af6c8a5602f;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.13;4.1.0 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.13;4.1.0 in central
        found org.apache.kafka#kafka-clients;3.9.1 in central
        found org.lz4#lz4-java;1.8.0 in central
        found org.xerial.snappy#snappy-java;1.1.10.8 in central
        found org.slf4j#slf4j-api;2.0.17 in central
        found org.apache.hadoop#hadoop-client-runtime;3.4.2 in central
        found org.apache.hadoop#hadoop-client-api;3.4.2 in central
        found com.google.code.findbugs#jsr305;3.0.0 in central
        found org.scala-lang.modules#scala-parallel-collections_2.13;1.2.0 in central
        found org.apache.commons#commons-pool2;2.12.1 in central
:: resolution report :: resolve 3842ms :: artifacts dl 128ms
        :: modules in use:
        com.google.code.findbugs#jsr305;3.0.0 from central in [default]
        org.apache.commons#commons-pool2;2.12.1 from central in [default]
        org.apache.hadoop#hadoop-client-api;3.4.2 from central in [default]
        org.apache.hadoop#hadoop-client-runtime;3.4.2 from central in [default]
        org.apache.kafka#kafka-clients;3.9.1 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.13;4.1.0 from central in [default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.13;4.1.0 from central in [default]
        org.lz4#lz4-java;1.8.0 from central in [default]
        org.scala-lang.modules#scala-parallel-collections_2.13;1.2.0 from central in [default]
        org.slf4j#slf4j-api;2.0.17 from central in [default]
        org.xerial.snappy#snappy-java;1.1.10.8 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   11  |   0   |   0   |   0   ||   11  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-e0c83b49-b5c2-46a3-a2df-8af6c8a5602f
        confs: [default]
        0 artifacts copied, 11 already retrieved (0kB/63ms)
26/01/06 16:09:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
```

[10]:
```python
from pyspark.sql.types import StructType, StringType
from pyspark.sql.functions import col, from_json

# 1. Define the schema
schema = StructType() \
    .add("City", StringType()) \
    .add("Model Year", StringType())

# 2. Connect to the Kafka stream
kafka_df = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", KAFKA_BROKER_URL) \
    .option("subscribe", KAFKA_TOPIC) \
    .option("startingOffsets", "earliest") \
    .load()

# 3. Transform the raw Kafka 'value' into columns
parsed_df = kafka_df.selectExpr("CAST(value AS STRING)") \
    .select(from_json(col("value"), schema).alias("data")) \
    .select("data.*")
```

[11]:
```
from pyspark.sql.functions import desc
# Find top 3 cities for 2023
top_cities_2023 = parsed_df.filter(col("Model Year") == "2023") \
    .groupBy("City") \
    .count() \
    .orderBy(desc("count")) \
    .limit(3)

# Start the stream
query = top_cities_2023.writeStream \
    .outputMode("complete") \
    .format("console") \
    .trigger(availableNow=True) \
    .start()

query.awaitTermination()
```

```
26/01/06 16:10:03 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
26/01/06 16:10:04 WARN MicroBatchExecution: Disabling AQE since AQE is not supported in stateful workloads.

-------------------------------------------
Batch: 0
-------------------------------------------
+--------+-----+
|    City|count|
+--------+-----+
| Seattle|22708|
| Tukwila|10517|
|Bellevue| 9706|
+--------+-----+
```

```
26/01/06 16:10:04 WARN MicroBatchExecution: Disabling AQE since AQE is not supported in stateful workloads.
[I 2026-01-06 16:10:41.829 ServerApp] Saving file at /wikimedia/exercise.ipynb1]
-------------------------------------------
Batch: 0
-------------------------------------------
+--------+-----+
|    City|count|
+--------+-----+
| Seattle|22708|
| Tukwila|10517|
|Bellevue| 9706|
+--------+-----+
```

Running the python script in ubuntu:

```
(.venv_new) lihicohen122@DESKTOP-AMKTFU4:~/Pyspark$ python3 producer.py
Streaming data to ev_topic...
```