

Working with ubuntu:

```
lihicohen122@DESKTOP-AMKTFU4: ~/pyspark
```

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 6.6.87.2-microsoft-standard-WSL2 x86_64)

* Documentation: <https://help.ubuntu.com>
* Management: <https://landscape.canonical.com>
* Support: <https://ubuntu.com/advantage>

* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s just raised the bar for easy, resilient and secure K8s cluster deployment.

<https://ubuntu.com/engage/secure-kubernetes-at-the-edge>

This message is shown once a day. To disable it please create the
/home/lihicohen122/.hushlogin file.

```
lihicohen122@DESKTOP-AMKTFU4:~$ ^[[200~sudo apt update
^[[201~sudo: command not found
lihicohen122@DESKTOP-AMKTFU4:~$ sudo apt update
[sudo] password for lihicohen122:
```

Docker:

```
Microsoft Windows [Version 10.0.26100.7462]
(c) Microsoft Corporation. All rights reserved.

C:\Users\lihicohen122> docker run -p 9092:9092 apache/kafka:latest
Unable to find image 'apache/kafka:latest' locally
latest: Pulling from apache/kafka
18232174bc9: Pull complete
3f73af89931: Pull complete
29f6d4a8d1: Pull complete
8b055152645: Pull complete
f41f65d07a3: Pull complete
42f712ac6d: Pull complete
3851c290e07: Pull complete
8e2b529f1af: Pull complete
723193c2f26: Pull complete
43b8fa7c2d7: Pull complete
5c28a4b260f: Pull complete
Digest: sha256:3f7b939115cd872e9ce936d88bd69712fde55f9902f46d793f64848dedc75
Status: Downloaded newer image for apache/kafka:latest
=> User
uid=1000(appuser) gid=1000(appuser) groups=1000(appuser)
=> Setting default values of environment variables if not already set.
KUSTLER_ID not set. Setting it to default value: "5163e387-ePCTK-X86sw"
=> Configuring ...
=> Launching ...
=> Using provided cluster id 5163e387-ePCTK-X86sw ...
2026-01-06 11:01:05,577] INFO Registered kafka:type=kafka.log4jController MBean (kafka.utils.log4jControllerRegistration)
2026-01-06 11:01:05,738] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
2026-01-06 11:01:05,741] INFO [ControllerServer id=1] Starting controller (kafka.server.ControllerServer)
2026-01-06 11:01:05,913] INFO Updated connection-accept-rate max connection creation rate to 2147483647 (kafka.network.ConnectionQuotas)
2026-01-06 11:01:05,957] INFO [SocketServer listenerType=CONTROLLER, nodeId=1] Created data-plane acceptor and processors for endpoint : ListenerName(CONTROLLER) (kafka.network.SocketServer)
2026-01-06 11:01:05,967] INFO CONTROLLER: resolved wildcard host to 16a289dd85f2 (org.apache.kafka.metadata.ListenerInfo)
2026-01-06 11:01:05,977] INFO [authorizerStart completed for endpoint CONTROLLER. Endpoint is now READY. (org.apache.kafka.server.network.EndpointReadyFutures)
2026-01-06 11:01:05,980] INFO [SharedServer id=1] Starting SharedServer (kafka.server.SharedServer)
2026-01-06 11:01:06,039] INFO [LogLoader partition=__cluster_metadata-0, dir=/tmp/kraft-combined-logs] Loading producer state till offset 0 (org.apache.kafka.storage.internals.log.UnifiedLog)
2026-01-06 11:01:06,039] INFO [LogLoader partition=__cluster_metadata-0, dir=/tmp/kraft-combined-logs] Reloading from producer snapshot and rebuilding producer state from offset 0 (org.apache.kafka.storage.internals.log.UnifiedLog)
2026-01-06 11:01:06,054] INFO [LogLoader partition=__cluster_metadata-0, dir=/tmp/kraft-combined-logs] Producer state recovery took 0ms for snapshot load and 0ms for segment recovery from offset 0 (org.apache.kafka.storage.internals.log.UnifiedLog)
2026-01-06 11:01:06,069] INFO [raft-expiration-reaper]: Starting (kafka.raft.TimingWheelExpirationServiceExpiredOperationReaper)
2026-01-06 11:01:06,087] INFO [RaftManager id=1] Reading Raft snapshot and log as part of the initialization (org.apache.kafka.raft.KafkaRaftClient)
2026-01-06 11:01:06,090] INFO [RaftManager id=1] Starting voters are VoterSet(voters=[1-VoterNode(voterKey=ReplicaKey(id=1, directoryId=undefined), listeners=Endpoints(endpoints=[ListenerName(CONTROLLER)-localhost/127.0.0.1:9092]), supportedRaftVersion=SupportedVersionRange(min_version=0, max_version=0))]) (org.apache.kafka.raft.KafkaRaftClient)
2026-01-06 11:01:06,092] INFO [RaftManager id=1] Starting request manager with static voters: [localhost:9092 (id: 1 rack: null isFenced: false)] (org.apache.kafka.raft.KafkaRaftClient)
2026-01-06 11:01:06,093] INFO [RaftManager id=1] Attempting durable transition to UnattachedState(epoch=0, leaderId=OptionalInt.empty, votedKey=Optional.empty, voters=[1], electionTimeoutMs=1756, highWatermark=Optional.empty) from null (org.apache.kafka.raft.QuorumState)
2026-01-06 11:01:06,110] INFO [RaftManager id=1] Completed transition to UnattachedState(epoch=0, leaderId=OptionalInt.empty, votedKey=Optional.empty, voters=[1], electionTimeoutMs=1756, highWatermark=Optional.empty) from null (org.apache.kafka.raft.QuorumState)
```

```
Microsoft Windows [Version 10.0.26100.7462]
(c) Microsoft Corporation. All rights reserved.

C:\Users\lihicohen122> docker ps
CONTAINER ID   IMAGE          COMMAND                  CREATED        STATUS        PORTS
16a289dd85f2   apache/kafka:latest   "/_cacert_entrypoint..."   29 seconds ago   Up 28 seconds   0.0.0.0:9092->9092/tcp
naughty_leavitt
```

```
C:\Users\יור לט יורל>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\יור לט יורל>docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-topics.sh --bootstrap-server localhost:9092 --describe
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\יור לט יורל>docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic wikimedia_topic_1 [--from-beginning]
```

Jupyter (Tutorial):


Jupyter tutorial Last Checkpoint: 33 minutes ago


File Edit View Run Kernel Settings Help

Trusted


JupyterLab Python 3 (ipykernel)

```
[1]: pip install kafka-python
Collecting kafka-python
  Downloading kafka_python-2.3.0-py2.py3-none-any.whl (326 kB)
    326.3/326.3 KB 2.2 MB/s eta 0:00:00a 0:00:01
Installing collected packages: kafka-python
Successfully installed kafka-python-2.3.0
Note: you may need to restart the kernel to use updated packages.

[2]: pip install pyspark
Collecting pyspark
  Downloading pyspark-4.1.0.tar.gz (455.3 MB)
    455.3/455.3 MB 570.1 kB/s eta 0:00:0000:0100:11
  Preparing metadata (setup.py) ... done
Collecting py4j<0.10.9.10,>=0.10.9.7
  Downloading py4j-0.10.9.9-py2.py3-none-any.whl (203 kB)
    203.0/203.0 KB 1.6 MB/s eta 0:00:00a 0:00:01
Using legacy 'setup.py install' for pyspark, since package 'wheel' is not installed.
Installing collected packages: py4j, pyspark
Running setup.py install for pyspark ... done
Successfully installed py4j-0.10.9.9 pyspark-4.1.0
Note: you may need to restart the kernel to use updated packages.

[3]: import os
import pyspark

[4]: os.environ['PYSPARK_SUBMIT_ARGS'] = f'--packages org.apache.spark:spark-sql-kafka-0-10-2.13:{pyspark.__version__} pyspark-shell'
os.environ['SPARK_SUBMIT_OPTS'] = '-Djdk.security.auth.login.Config=ignore'

[5]: KAFKA_BROKER_URL = "localhost:9092"
KAFKA_TOPIC = "wikimedia_topic_1"

[6]: pip install requests

Requirement already satisfied: requests in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (2.32.5)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (from requests) (3.4.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (from requests) (2.6.2)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (from requests) (3.11)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicochen122/.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/lihicochen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (2.32.5)
Requirement already satisfied: six in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (1.17.0)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (from requests>=2.9->sseclient) (2.6.2)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (from requests>=2.9->sseclient) (3.4.4)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicochen122/.venv3.11/lib/python3.11/site-packages (from requests>=2.9->sseclient) (2026.1.4)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicochen122/.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
```

```
[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/lihichohen122/.venv3.11/lib/python3.11/site-packages/pyspark/jars/ivy-2.5.3.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/lihichohen122/.ivy2.5.2/cache
```

```
[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	categoryize	[[[:José Mingoranc...	Werthercito	Categoria:Falleci...	es.wikipedia.org	
17682112	categoryize	[[[:Gleituferr]] vo...	Alexander Gamauf	Kategorie:Wartung...	de.wiktionary.org	
NULL	edit	/# wbeditentity-u...	Rkieferrbot	File:Sori - Chies...	commons.wikimedia...	
NULL	edit	/# wbeditentity-u...	Chabe01	File:Plaque Allée...	commons.wikimedia...	
NULL	edit	/# wbeditentity-u...	Sporti	Q137710176	www.wikidata.org	
133285614	log	Bot: Mass deletin...	Ladsgroup	Thảo luận Thành v...	vi.wikipedia.org	
142109988	edit	/# Verb */ litera...	Überkel	gall	en.wiktionary.org	
32084861	edit	/# Библиография */	Al Silonov	скари	ru.wiktionary.org	
NULL	edit	/# wbeditentity-u...	M2k~dewiki	Q137673651	www.wikidata.org	
11705348	new	"[[[:en:Special:Re...	Wikipidiya1	कराव	hi.wikipedia.org	
1983163400	categoryize	[[[:Wikipedia:Arti...	MatrixBot	Category:Noindexe...	en.wikipedia.org	
70248355	categoryize	Nampiana tamin'ny...	HJAOVELO	Sokajy:Anarana io...	mg.wiktionary.org	
NULL	edit	/# wbeditentity-u...	Chabe01	File:Plaque Allée...	commons.wikimedia...	
NULL	edit	/# wbeditentity-u...	Robinscarlet	Q49510664	www.wikidata.org	
NULL	edit	/# wbeditentity-u...	Sakretsul	Q40917	www.wikidata.org	
91406231	edit	문자열 변경 ((= *외...	A.TedBot	키케온	ko.wikipedia.org	
NULL	log		Paul Gascoigne	Niky Savage	it.wikipedia.org	
NULL	edit	/# wbeditentity-u...	SchlurcheBot	File:01.03 總統「洲美三...	commons.wikimedia...	
2420052	edit	додав посилення н...	Balakun	Біліжхрелна:Канен...	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
必宗森	1
Alexander Gamauf	1
Rkieferrbot	3
Balakun	1
Paul Gascoigne	1
Werthercito	1
Ivtorov	1
Sakretsul	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
categoryize	44

[]:



Docker (Exercise):

```
C:\Users\יורן לט > 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\יורן לט >
```

Jupyter (Exercise):

 **jupyter** exercise Last Checkpoint: 53 minutes ago 

File Edit View Run Kernel Settings Help Trusted

JupyterLab Python 3 (ipykernel)

```
[1]: pip install kafka-python
Requirement already satisfied: kafka-python in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (2.3.0)
Note: you may need to restart the kernel to use updated packages.

[2]: pip install pyspark
Requirement already satisfied: pyspark in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (4.1.0)
Requirement already satisfied: py4j<0.10.9.10,>=0.10.9.7 in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (from pyspark) (0.10.9.9)
Note: you may need to restart the kernel to use updated packages.

[3]: import os
import pyspark

[4]: os.environ['PYSPARK_SUBMIT_ARGS'] = f'--packages org.apache.spark:spark-sql-kafka-0-10-2.13:{pyspark.__version__} pyspark-shell'
os.environ['SPARK_SUBMIT_OPTS'] = '-Djdk.security.auth.login.Config=ignore'

[5]: KAFKA_BROKER_URL = "localhost:9092"
KAFKA_TOPIC = "ev_topic"

[6]: pip install requests
Requirement already satisfied: requests in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (2.32.5)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (from requests) (2026.1.4)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (from requests) (3.4.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (from requests) (2.6.2)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (from requests) (3.11)
Note: you may need to restart the kernel to use updated packages.

[7]: pip install sseclient
Requirement already satisfied: sseclient in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (0.0.27)
Requirement already satisfied: requests>=2.9 in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (from sseclient) (2.32.5)
Requirement already satisfied: six in /home/lihicochen122/PySpark/.venv_new/lib/python3.11/site-packages (from sseclient) (1.17.0)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicochen122/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/lihicochen122/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/lihicochen122/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/lihicochen122/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]: 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/lihicochen122/Pyspark/.venv_new/lib/python3.11/site-packages/pyspark/jars/ivy-2.5.3.jar!/org/apache/ivy/core/
e/settings/ivysettings.xml
Ivy Default Cache set to: /home/lihicochen122/.ivy2.5.2/cache
The jars for the packages stored in: /home/lihicochen122/.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]
```

```
    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]
-----
|               |             |             |             |             |             |
|      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).
```

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

```
[10]: 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...
```