

Working with ubuntu:

```
lihicoen122@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/lihicoen122/.hushlogin file.

lihicoen122@DESKTOP-AMKTFU4:~$ ^[[200~sudo apt update
^[[[201~sudo: command not found
lihicoen122@DESKTOP-AMKTFU4:~$ sudo apt update
[sudo] password for lihicoen122:
```

Docker:

```
Administrator: docker run -p 9092:9092 apache/kafka:latest
Microsoft Windows [Version 10.0.26100.7462]
(c) Microsoft Corporation. All rights reserved.

C:\Users\lihicoen122\Documents\GitHub\kafka-secure

;Users\lihicoen122\Documents\GitHub\kafka-secure> docker run -p 9092:9092 apache/kafka:latest
Pulling from apache/kafka
18232174bc9: Pull complete
3f73af0f931: Pull complete
42f165d045: Pull complete
4af1f65d07a3: Pull complete
42f712ecf6d: Pull complete
3851e29d6a7: Pull complete
8eeb5291faef: Pull complete
7232a2a033c: Pull complete
43bf8a7c47: Pull complete
5e384d4206f: Pull complete
Digest: sha256:3f7b939115cd4872e9ceef9369d88bd69712fde55ff902f46d793f648483dec75
Status: Downloaded newer image for apache/kafka:latest
status: Downloaded newer image for apache/kafka:latest
--> User
--> User
id=1000(gpuuser) gid=1000(gpuuser) groups=1000(gpuuser)
--> Setting default values of environment variables if not already set.
LUSTER_ID not set. Setting it to default value: "516g3nSh1-eMtk-X86sw"
--> Configuring ...
--> Launching ...
--> Using provided Cluster.id 516g3nSh1-eMtk-X86sw ...
2026-01-06 11:01:45.727] INFO Registered KafkaLog4jController $Bean (kafka.utils.Log4jControllerRegistration$)
2026-01-06 11:01:45.738] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
2026-01-06 11:01:45.913] INFO Updated connection-accept-rate max connection creation rate to 2147483647 (kafka.network.ConnectionQuotas)
2026-01-06 11:01:45.957] INFO [ControllerServer id=1] Starting controller (kafka.server.ControllerServer)
2026-01-06 11:01:45.967] INFO [ControllerServer id=1] resolved wildcard host to 16a289dd85f2 (org.apache.kafka.metadata.ListenerInfo)
2026-01-06 11:01:45.971] INFO [ControllerServer id=1] assigned controller id 1 (CONTROLLER) and endpoint localhost:9092 (org.apache.kafka.server.network.EndpointReadyFutures)
2026-01-06 11:01:45.980] INFO [SharedServer id=1] Starting SharedServer (kafka.server.ShardedServer)
2026-01-06 11:01:46.039] INFO [LogLoader partitions=_cluster_metadata-0, dir=/tmp/kraft-combined-logs] Loading producer state till offset 0 (org.apache.kafka.storage.internals.log.UnifiedLog)
2026-01-06 11:01:46.039] INFO [LogLoader partitions=_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 11:01:46.059] INFO [LogLoader partitions=_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 11:01:46.069] INFO Initialized snapshots with 10s SortedSet() from /tmp/kraft-combined-logs/_cluster_metadata-0 (kafka.raft.KafkaMetadataLog$)
2026-01-06 11:01:46.082] INFO [raft-expansion-reaper]: Starting (kafka.raft.TimingHeelExpirationService$ExpiredOperationReaper)
2026-01-06 11:01:46.087] INFO [RaftManager id=1] Reading Raft log and log as part of the initialization (org.apache.kafka.raft.RaftManagerClient)
2026-01-06 11:01:46.090] INFO [RaftManager id=1] Starting voters are VoterSet{voterId=(1-Watermark),replicaKey=(1-DirectoryId=undefined)}, listeners=Endpoints{endpoints=[(ListenerName(CONTROLLER)-localhost:9092)]}
2026-01-06 11:01:46.092] INFO [RaftManager id=1] Starting Raft manager with static voters: [localhost:9093 (id: 1 rack: null isFenced: false)] (org.apache.kafka.raft.RaftManagerClient)
2026-01-06 11:01:46.098] 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.QorunmState)
2026-01-06 11:01:46.119] INFO [RaftManager id=1] Completed transition to UnattachedState(epoch=0, leaderId=OptionalInt.empty(), votedKey=Optional.empty(), voters=[1], electionTimeoutMs=1756, highWatermark=Optional.empty())
```

```
Administrator: docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic...
Microsoft Windows [Version 10.0.26100.7462]
(c) Microsoft Corporation. All rights reserved.

C:\Users\lihicoen122\Documents\GitHub\kafka-secure> docker ps
CONTAINER ID   IMAGE          COMMAND           CREATED          STATUS          PORTS
NAMES
16a289dd85f2   apache/kafka:latest   "/__cacert_entrypoint..."   29 seconds ago   Up 28 seconds   0.0.0.0:9092->9092/tcp
haughty_leavitt
```

```
C:\Users\lih\Documents>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\lih\Documents>docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-topics.sh --bootstrap-server localhost:9092 --describe --topic wikimedia_topic_1
Topic: wikimedia_topic_1          TopicId: Rm1x_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\lih\Documents>docker exec -it 16a289dd85f2 /opt/kafka/bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic wikimedia_topic_1 [--from-beginning]
```

Jupyter (Tutorial):

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

[3]: import os
import pypar

[4]: os.environ['PYSPIKE_SUBMIT_ARGS'] = '--packages org.apache.spark:spark-sql-kafka-0-10_2.13:[pyspark.__version__] pyspark-shell'
os.environ['SPARK_SUBMIT_OPTS'] = '-Djdk.security.auth.login.Configignore'

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

[6]: pip install requests
Requirement already satisfied: requests in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (2.32.5)
Requirement already satisfied: charset_normalizer<4,>2 in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (from requests) (3.4.4)
Requirement already satisfied: six in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (from requests) (2.6.2)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (from requests) (3.11)
Requirement already satisfied: certifi<=2017.4.17 in /home/lihicoen122/.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/lihicoen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (2.32.5)
Requirement already satisfied: six in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (1.17.0)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (2.6.2)
Requirement already satisfied: charset_normalizer<4,>2 in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (3.4.4)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicoen122/.venv3.11/lib/python3.11/site-packages (from sseclient) (2026.1.4)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicoen122/.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": "eyJ0eXAiOiJKV1QiLCJhbGciOiJSUzI1NiJ9eyJhdWQiOiI3ODkSN2EzNzllMmE2NzIwNTJlNTg4NWE4NjkxZTBkMiIsImp0aSI6IjIxZTE2MTJkYzI0TkY2U4YmU0Z"
      }
      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/lihicochen122/.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/lihicochen122/.ivy2.5.2/cache
-----[ivysettings.xml]-----
<ivysettings>
  <ivy>
    <conf name="default"/>
    <modules>
      <module name="findbugs" revision="3.0.0" conf="default"/>
      <module name="commons-pool2" revision="2.12.1" conf="default"/>
      <module name="hadoop-client" revision="3.4.2" conf="default"/>
      <module name="hadoop-client-runtime" revision="3.4.2" conf="default"/>
      <module name="kafka" revision="2.1.0" conf="default"/>
      <module name="spark-sql-kafka-0-10_2.13" revision="4.1.0" conf="default"/>
      <module name="spark-token-provider-kafka-0-10_2.13" revision="4.1.0" conf="default"/>
      <module name="slf4j" revision="2.0.17" conf="default"/>
      <module name="hadoop-client-api" revision="3.4.2" conf="default"/>
    </modules>
  </ivy>
</ivysettings>
-----[ivysettings.xml]-----

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 :: resolving 15623ms :: artifacts dl 17467ms
-----[resolution report]-----
:: modules in use:
com.google.code.findbugs#jxr305;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.sparkspark-sql-kafka-0-10_2.13#spark-sql-kafka-0-10_2.13;4.1.0 from central in [default]
org.apache.sparkspark-token-provider-kafka-0-10_2.13#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]
-----[resolution report]-----
| conf | modules | artifacts |
| number | search|dwnlded|evicted|| number|dwnlded|
-----[resolution report]-----
| default | 11 | 11 | 11 | 0 || 11 | 11 |
-----[resolution report]-----

:: retrieving :: org.apache.spark#spark-submit-parent-35a7d01e-2e4c-45b2-aece-40f5b47ff6d0
conf: [default]
11 artifacts copied, 0 already retrieved (62936kB/107ms)
26/01/06 13:24:19 WARN CodeLoader: 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|Categoria:Falleci...| es.wikipedia.org|
| 17682112|categorize|[[:Gleituf...| 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...| Oberkel|gall| en.wiktionary.org|
| 32084861|    edit| /* Былompson */| Al Silonov|скарн| ru.wiktionary.org|
|   NULL|    edit| /* wbeditentity-u...| M2k-dewiki|Q137673651| www.wikidata.org|
| 11705348|    new|[[:en:Special:Re...| Wikipidyal|  | hi.wikipedia.org|
|1983163400|categorize|[[:Wikipedia:Arti...| MatrixBot|Category:Noindexe...| en.wikipedia.org|
| 70248355|categorize|[[:Nampiana tamin'ny...| HJAOVELO|Sokajv:Anarana io...| mg.wiktionary.org|
|   NULL|    edit| /* wbcreateclaim...| Chabe01|File:Plaque Allée...|commons.wikimedia...|
|   NULL|    edit| /* wbsetitelink...| Robinscarlet|Q49510664| www.wikidata.org|
|   NULL|    edit| /* wbsetqualifier...| Sakrets|Q48917| www.wikidata.org|
| 91406231|    edit|봇: 문자열 변경 ((== *외...| A.TedBot|  | 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|  | 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|
| 巫衆霖| 1|
| Alexander Gamauf| 1|
| Rkieferbot| 3|
| Balakun| 1|
| Paul Gascoigne| 1|
| Werthercito| 1|
| Ivtorov| 1|
| Sakrets| 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 0x72b42d745900>
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\lihicoen122\PycharmProjects>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\lihicoen122\PycharmProjects>
```

Jupyter (Exercise):



The screenshot shows a Jupyter Notebook interface with several code cells and their corresponding outputs. The notebook has a title bar "Jupyter exercise Last Checkpoint: 53 minutes ago" and a toolbar with File, Edit, View, Run, Kernel, Settings, Help, and Trusted. The Python version is listed as Python 3 (ipykernel).

Code cells and their outputs:

- [1]:

```
pip install kafka-python
Requirement already satisfied: kafka-python in /home/lihicoen122/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/lihicoen122/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/lihicoen122/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.Configignore'
```
- [5]:

```
KAFKA_BROKER_URL = "localhost:9092"
KAFKA_TOPIC = "ev_topic"
```
- [6]:

```
pip install requests
Requirement already satisfied: requests in /home/lihicoen122/Pyspark/.venv_new/lib/python3.11/site-packages (2.32.5)
Requirement already satisfied: certifi>=2017.4.17 in /home/lihicoen122/Pyspark/.venv_new/lib/python3.11/site-packages (from requests) (2026.1.4)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicoen122/Pyspark/.venv_new/lib/python3.11/site-packages (from requests) (3.4.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /home/lihicoen122/Pyspark/.venv_new/lib/python3.11/site-packages (from requests) (2.6.2)
Requirement already satisfied: idna<4,>=2.5 in /home/lihicoen122/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/lihicoen122/Pyspark/.venv_new/lib/python3.11/site-packages (0.0.27)
Requirement already satisfied: requests>=2.9 in /home/lihicoen122/Pyspark/.venv_new/lib/python3.11/site-packages (from sseclient) (2.32.5)
Requirement already satisfied: six in /home/lihicoen122/Pyspark/.venv_new/lib/python3.11/site-packages (from sseclient) (1.17.0)
Requirement already satisfied: charset_normalizer<4,>=2 in /home/lihicoen122/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/lihicoen122/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/lihicoen122/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/lihicoen122/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.
```
- [8]:

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

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/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 :: resolves 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|dwlded|evicted|| number|dwlded|
-----|       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...
```