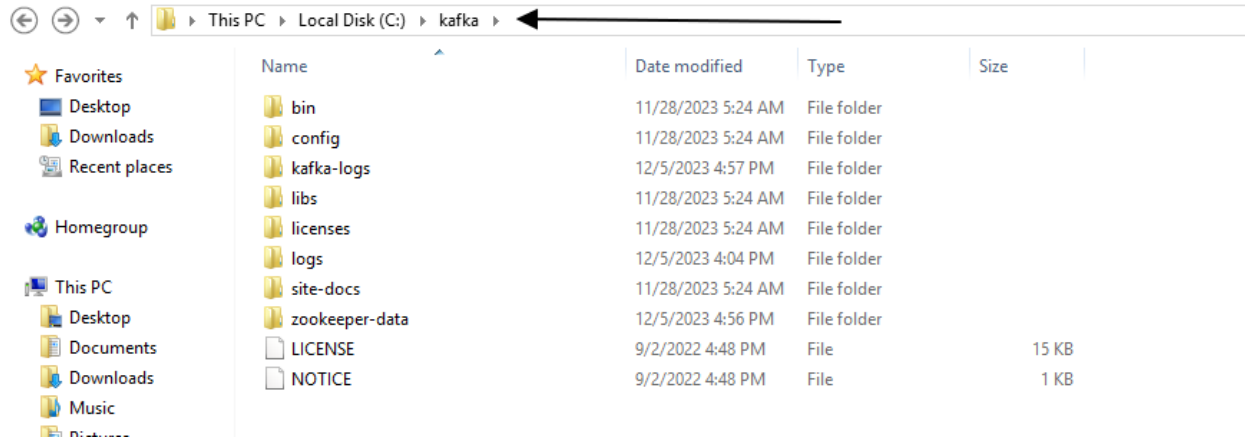
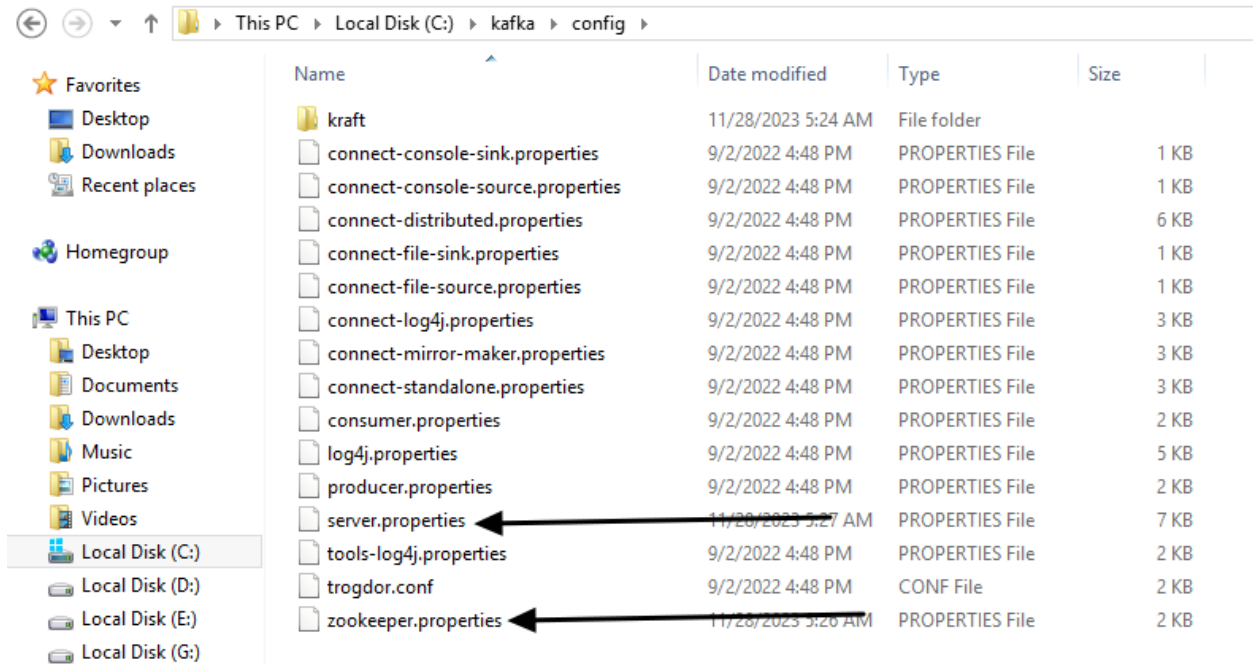


# Kafka

Extract “kafka\_2.12-2.8.2.tgz” and add it in “C:\kafka”



Go to config file to do some changes:



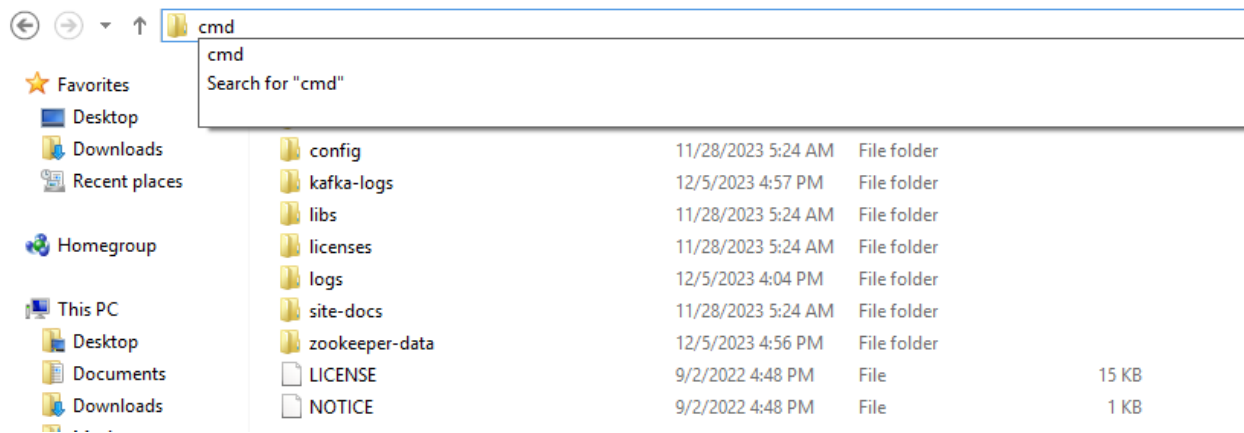
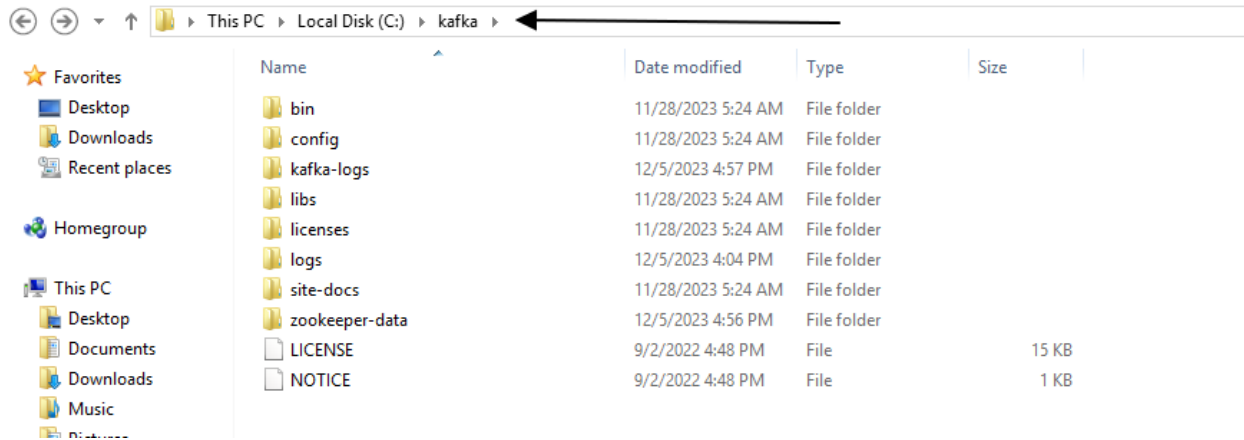
In zookeeper.properties, change line 16 to be like this:

```
9 #
10 # Unless required by applicable law or agreed
11 # distributed under the License is distributed
12 # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND
13 # See the License for the specific language
14 # limitations under the License.
15 # the directory where the snapshot is stored
16 dataDir=c:/kafka/zookeeper-data ←
17 # the port at which the clients will connect
18 clientPort=2181
19 # disable the per-ip limit on the number of
20 maxClientCnxns=0
```

In server.properties, change line 60 to be like this:

```
57 ##### Log Basics #####
58
59 # A comma separated list of directories under which
60 log.dirs=c:/kafka/kafka-logs ←
61
62 # The default number of log partitions per topic
63 # parallelism for consumption, but this will also
64 # the brokers.
65 num.partitions=1
66
```

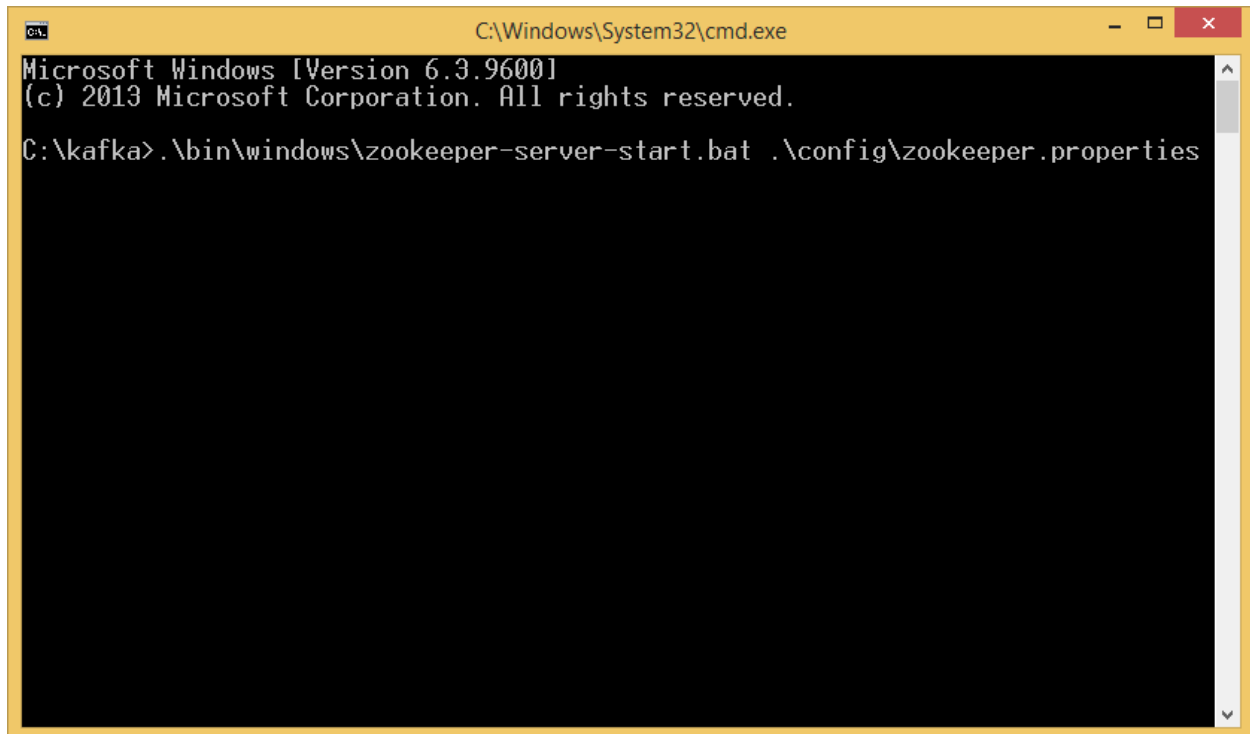
Navigate back to “C:\kafka” and open CMD:



Run zookeeper server first and keep it running (do not close the CMD)

Write this command and hit enter:

`.\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties`



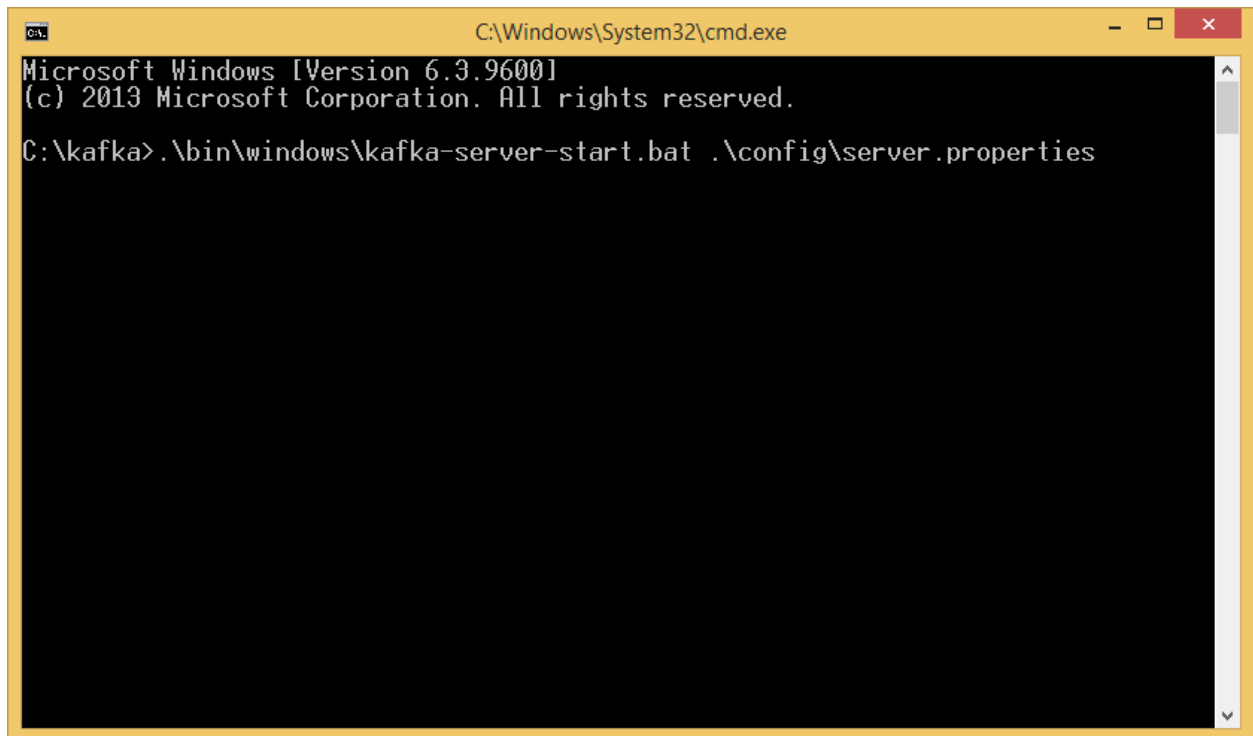
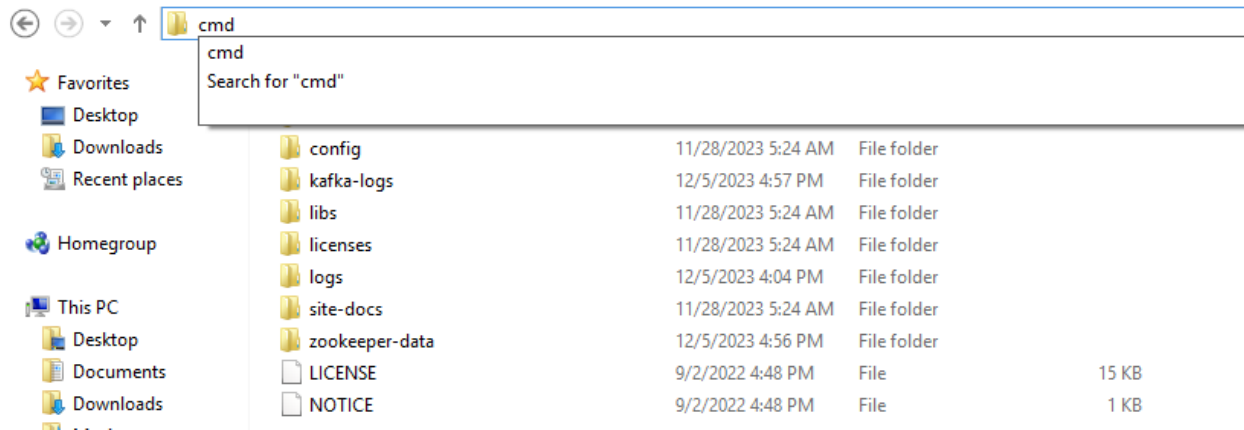
```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 6.3.9600]
(c) 2013 Microsoft Corporation. All rights reserved.

C:\kafka>.\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties
```

Open another CMD and Run kafka server and keep it running (do not close the CMD)

Write this command and hit enter:

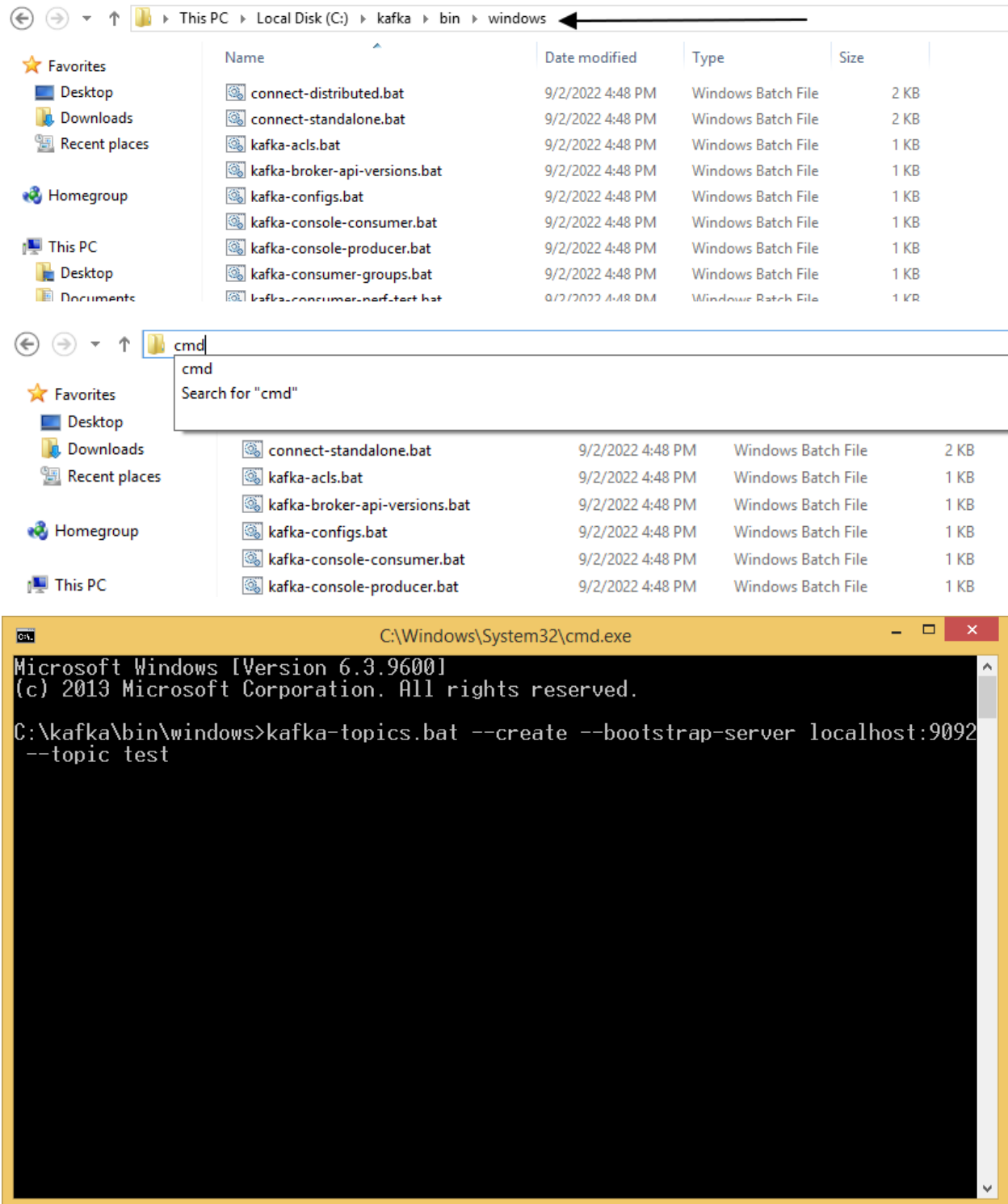
`.\bin\windows\kafka-server-start.bat .\config\server.properties`



Navigate to “C:\kafka\bin\windows” and open another CMD:

Create a Topic. write this command and hit enter:

kafka-topics.bat --create --bootstrap-server localhost:9092 --topic test



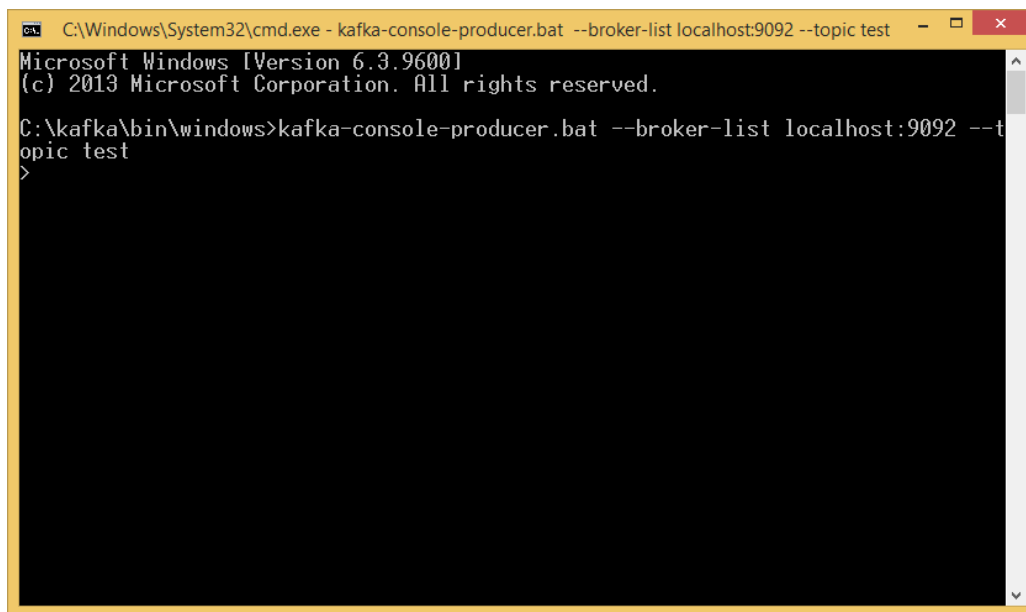
To check that everything is working, Open another 2 CMDs, one for producer and another for consumer

For producer write (in a cmd):

```
kafka-console-producer.bat --broker-list localhost:9092 --topic test
```

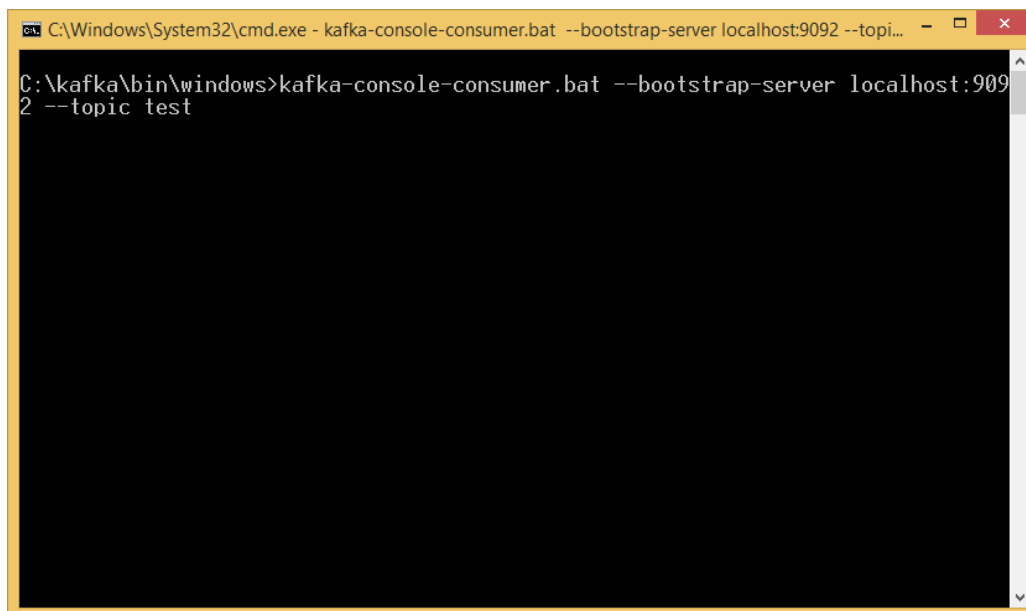
For consumer write (in another cmd):

```
kafka-console-consumer.bat --bootstrap-server localhost:9092 --topic test
```



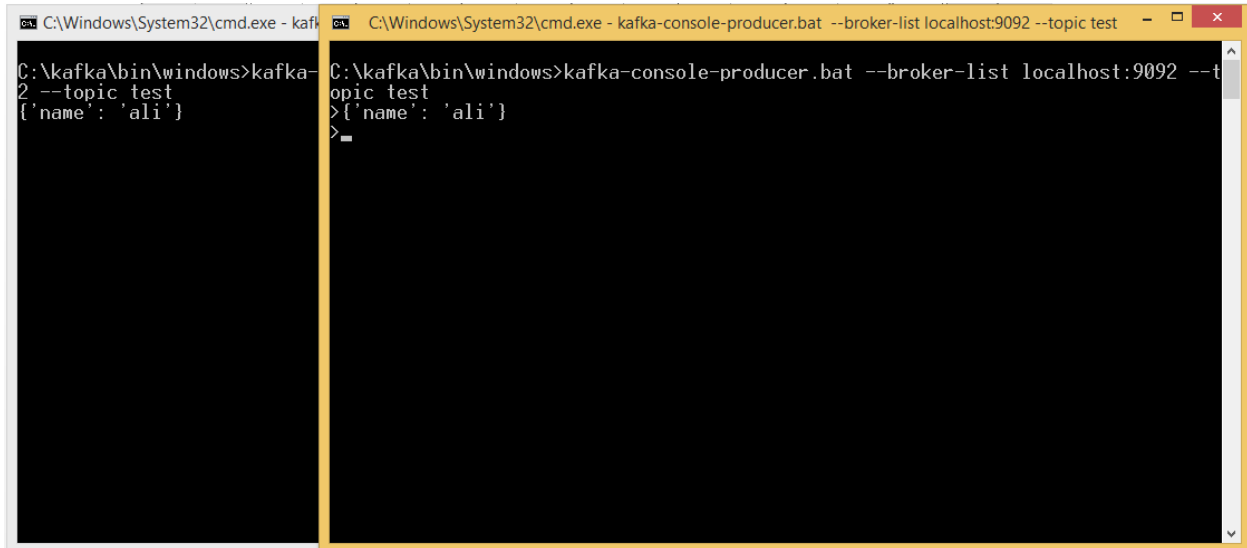
```
C:\Windows\System32\cmd.exe - kafka-console-producer.bat --broker-list localhost:9092 --topic test
Microsoft Windows [Version 6.3.9600]
(c) 2013 Microsoft Corporation. All rights reserved.

C:\kafka\bin\windows>kafka-console-producer.bat --broker-list localhost:9092 --topic test
>
```



```
C:\Windows\System32\cmd.exe - kafka-console-consumer.bat --bootstrap-server localhost:9092 --topic test
C:\kafka\bin\windows>kafka-console-consumer.bat --bootstrap-server localhost:9092 --topic test
```

Write a message in producer that should appear in consumer:

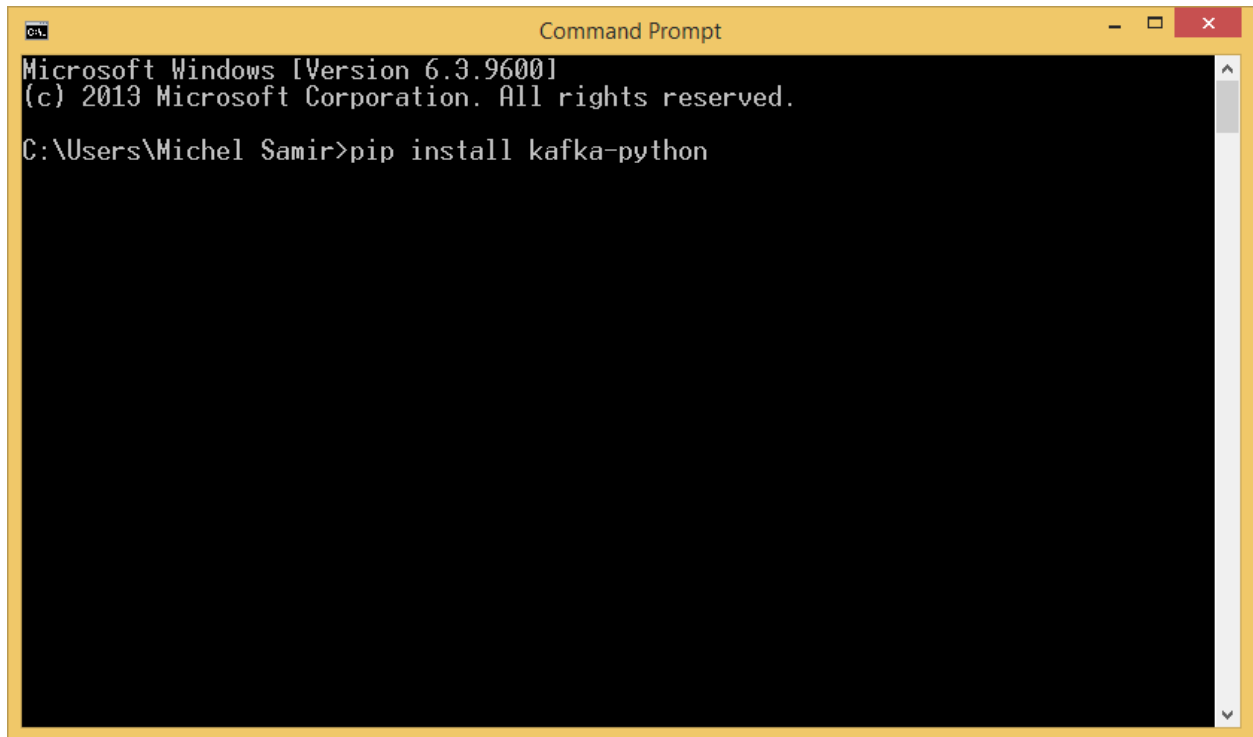


```
C:\Windows\System32\cmd.exe - kafka C:\Windows\System32\cmd.exe - kafka-console-producer.bat --broker-list localhost:9092 --topic test
C:\kafka\bin\windows>kafka-2 --topic test
{'name': 'ali'}
C:\kafka\bin\windows>kafka-console-producer.bat --broker-list localhost:9092 --t
opic test
>{'name': 'ali'}
>
```



To use kafka in python, install kafka-python package

pip install kafka-python

A screenshot of a Windows Command Prompt window. The title bar is yellow and says "Command Prompt". The window content is black with white text. It shows the Microsoft Windows version (6.3.9600) and copyright information (© 2013 Microsoft Corporation). The command prompt shows the user's location as C:\Users\Michel Samir and the command being executed is 'pip install kafka-python'.

## Two ways to connect kafka with spark

- 1- Use “spark-submit --packages <packages> <python file>”
- 2- Use “spark-submit --jars <packages paths> <python file>”

If use want to download the packages by yourself and use them offline,  
Use **--jars**

If you want spark to download them from online repository while  
running your first spark-kafka application, Use **--packages**

Recommended (I recommend using **--packages**)

Packages that we need (depends on the versions that you use):

org.apache.spark:spark-sql-kafka-0-10\_<scala\_version>:<spark\_version>

org.apache.spark:spark-streaming-kafka-0-10\_<scala\_version>:<spark\_version>

org.apache.kafka:kafka-clients:<kafka\_version>

Therefore, my versions will be:

org.apache.spark:spark-sql-kafka-0-10\_2.12:3.2.4

org.apache.spark:spark-streaming-kafka-0-10\_2.12:3.2.4

org.apache.kafka:kafka-clients:2.8.2

Therefore, the command will be for any python file containing spark and kafka:

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.2.4,org.apache.spark:spark-streaming-kafka-0-10_2.12:3.2.4,org.apache.kafka:kafka-clients:2.8.2 <python_file>
```

Note: if you want to use --jar, and download the packages by yourself, you can download them from maven repository (4 packages):

<https://mvnrepository.com/artifact/org.apache.kafka/kafka-clients/2.8.2>

[https://mvnrepository.com/artifact/org.apache.spark/spark-sql-kafka-0-10\\_2.12/3.2.4](https://mvnrepository.com/artifact/org.apache.spark/spark-sql-kafka-0-10_2.12/3.2.4)

[https://mvnrepository.com/artifact/org.apache.spark/spark-streaming-kafka-0-10-assembly\\_2.12/3.2.4](https://mvnrepository.com/artifact/org.apache.spark/spark-streaming-kafka-0-10-assembly_2.12/3.2.4)

<https://mvnrepository.com/artifact/org.apache.commons/commons-pool2/2.11.1>

Therefore, the command will be for any python file containing spark and kafka:

```
spark-submit --jars spark-sql-kafka-0-10_2.12-3.2.4.jar,spark-streaming-kafka-0-10-assembly_2.12-3.2.4.jar,kafka-clients-2.8.2.jar,commons-pool2-2.11.1.jar <python_file>
```

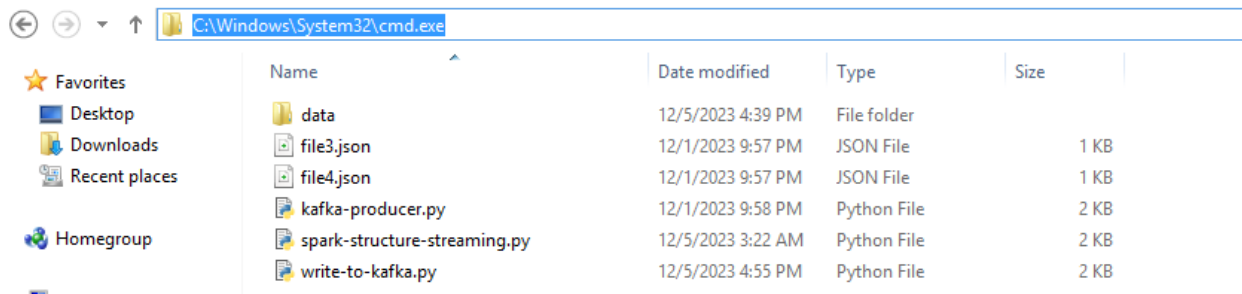
Note: these jars should be in the same folder with <python\_file>

2 examples are provided with this pdf:

- 1- Produce one message to kafka topic and spark consumes the message (folder 10)
- 2- Produce messages from files using spark to kafka topic and spark consumes the messages and store them in database. (folder 14)

For example (folder 14):

- 1- You will run the consumer (spark-structure-streaming.py)
- 2- You will run the producer (write-to-kafka.py)



Name	Date modified	Type	Size
data	12/5/2023 4:39 PM	File folder	
file3.json	12/1/2023 9:57 PM	JSON File	1 KB
file4.json	12/1/2023 9:57 PM	JSON File	1 KB
kafka-producer.py	12/1/2023 9:58 PM	Python File	2 KB
spark-structure-streaming.py	12/5/2023 3:22 AM	Python File	2 KB
write-to-kafka.py	12/5/2023 4:55 PM	Python File	2 KB

To run consumer, open cmd and write,

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.2.4,org.apache.spark:spark-streaming-kafka-0-10_2.12:3.2.4,org.apache.kafka:kafka-clients:2.8.2 spark-structure-streaming.py
```

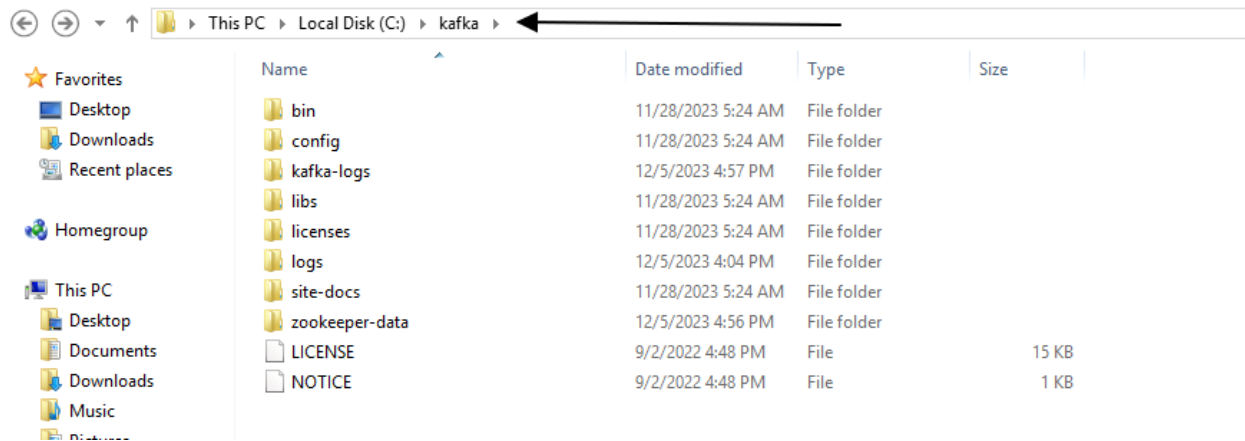
To run producer open another cmd and write:

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.2.4,org.apache.spark:spark-streaming-kafka-0-10_2.12:3.2.4,org.apache.kafka:kafka-clients:2.8.2 write-to-kafka.py
```

Note: change the directory path in code to match yours

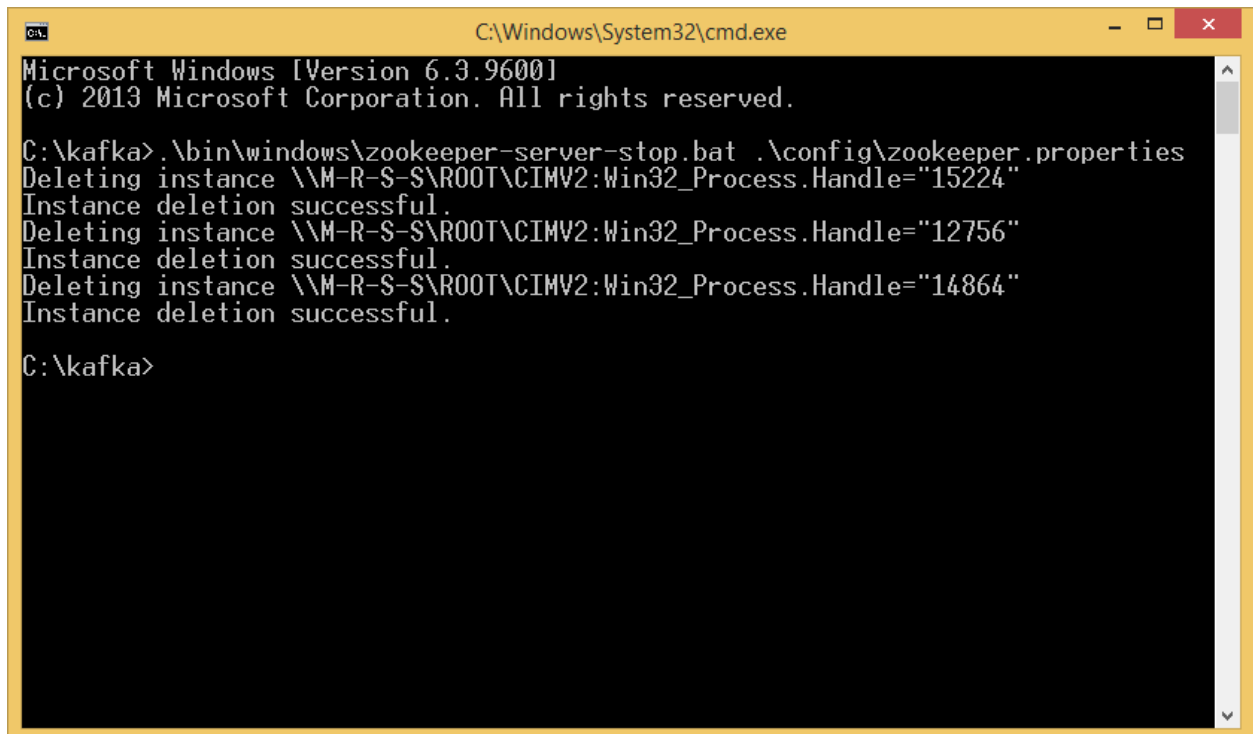
After Finishing, stop zookeeper and kafka servers

Navigate to “C:\kafka” and open cmd:



Write:

`.\bin\windows\zookeeper-server-stop.bat .\config\zookeeper.properties`



**Best Wishes**