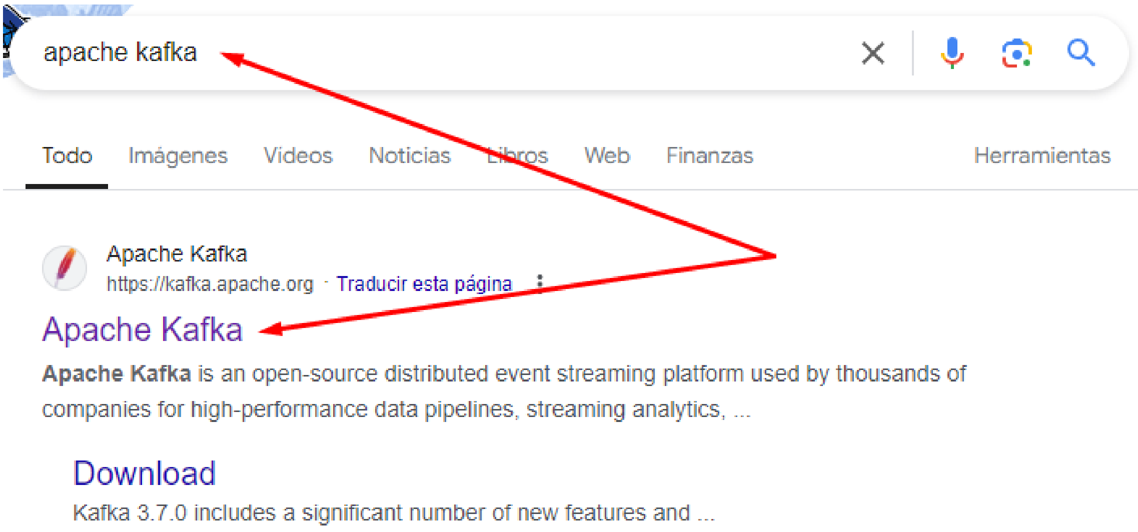
I.E.S. Fernando Aguilar Quignon
C/Conil de la Frontera, 3
CP 11010, Cádiz

5075. Big Data aplicado - 1ª Evaluación (RA 1 – CE b)

Unidad Didáctica 1. Estrategias de ingestión y almacenamiento de datos en Big Data

Práctica 7. Practicando la ingesta de datos en Streaming.

Paso 1. Descargamos e instalamos Apache Kafka.



apache kafka

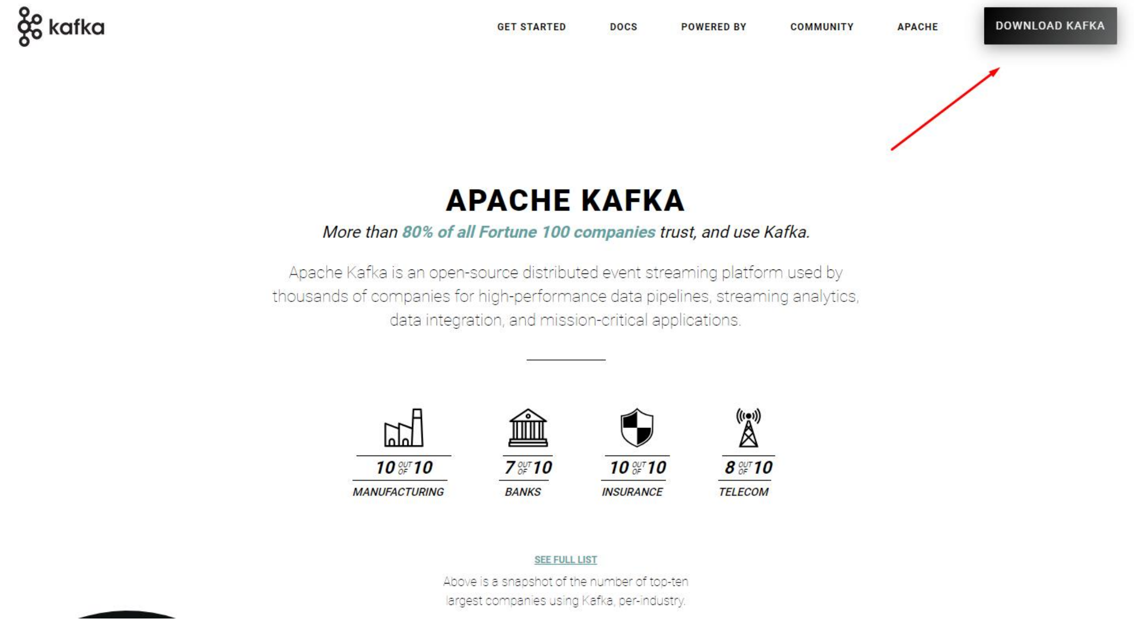
Apache Kafka
https://kafka.apache.org · Traducir esta página

Apache Kafka

Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, ...

Download

Kafka 3.7.0 includes a significant number of new features and ...

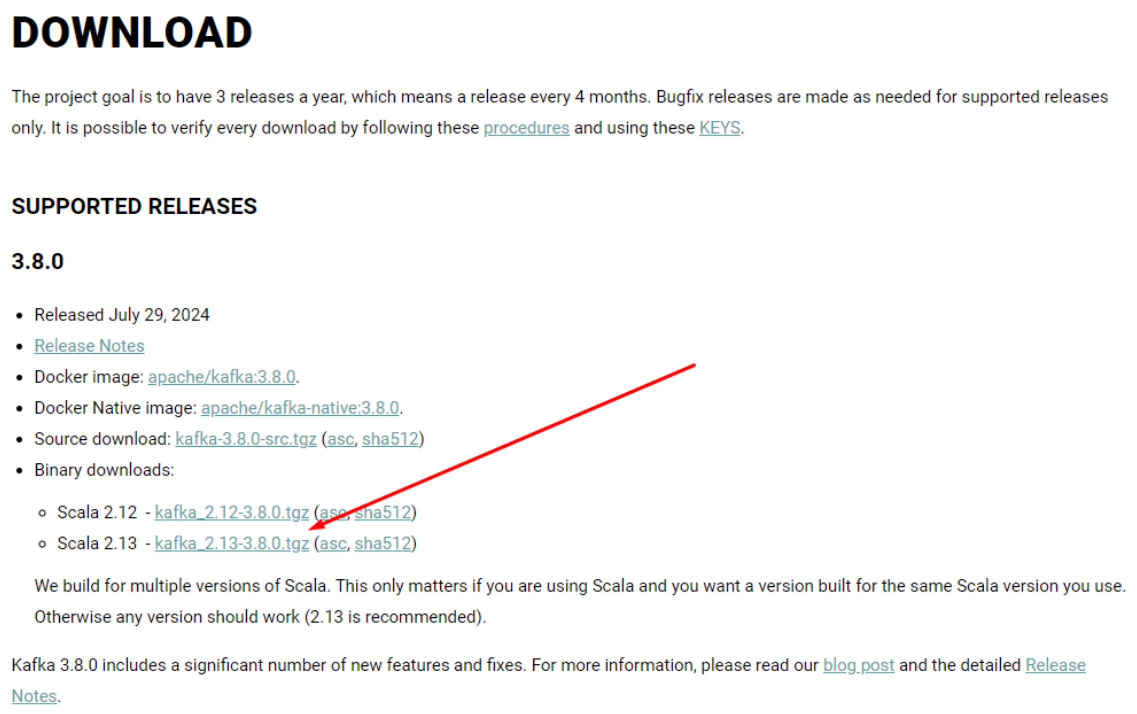


APACHE KAFKA

More than 80% of all Fortune 100 companies trust, and use Kafka.

Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications.

Above is a snapshot of the number of top-ten largest companies using Kafka, per-industry.



DOWNLOAD

The project goal is to have 3 releases a year, which means a release every 4 months. Bugfix releases are made as needed for supported releases only. It is possible to verify every download by following these procedures and using these KEYS.

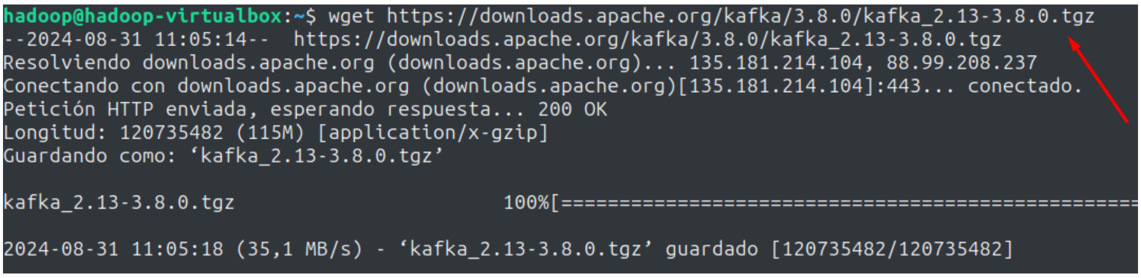
SUPPORTED RELEASES

3.8.0

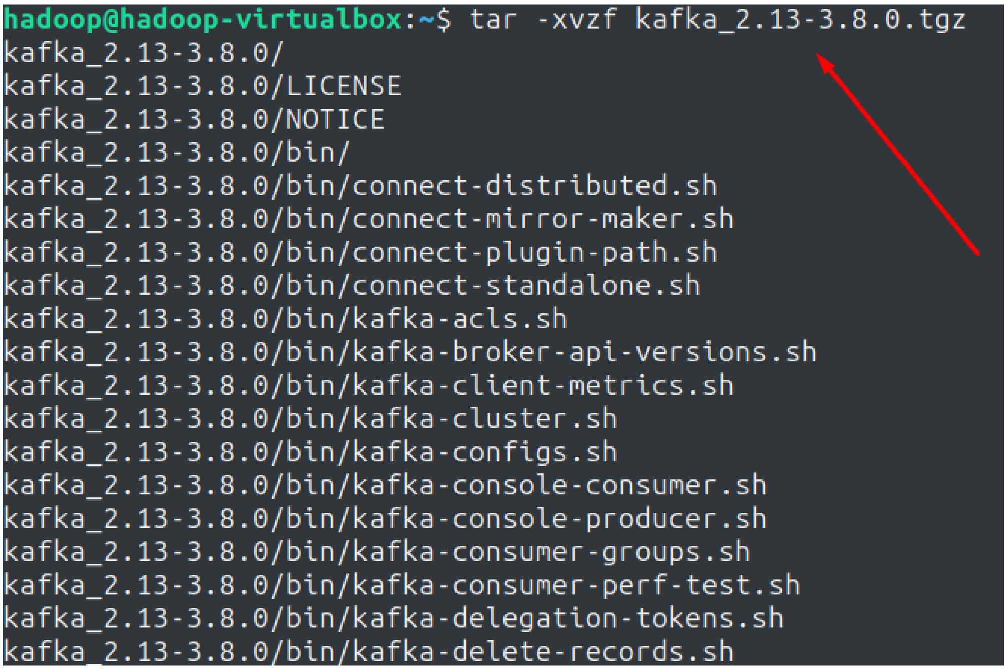
- Released July 29, 2024
- Release Notes
- Docker image: apache/kafka:3.8.0.
- Docker Native image: apache/kafka-native:3.8.0.
- Source download: kafka-3.8.0-src.tgz (asc, sha512)
- Binary downloads:
  - Scala 2.12 - kafka_2.12-3.8.0.tgz (asc, sha512)
  - Scala 2.13 - kafka_2.13-3.8.0.tgz (asc, sha512)

  We build for multiple versions of Scala. This only matters if you are using Scala and you want a version built for the same Scala version you use. Otherwise any version should work (2.13 is recommended).

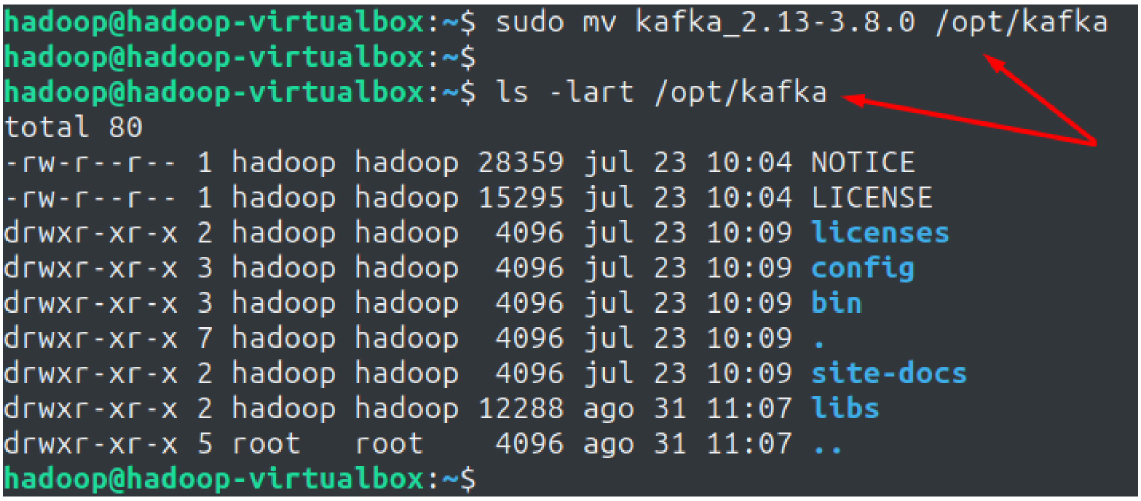
Kafka 3.8.0 includes a significant number of new features and fixes. For more information, please read our blog post and the detailed Release Notes.



```
hadoop@hadoop-virtualbox:~$ wget https://downloads.apache.org/kafka/3.8.0/kafka_2.13-3.8.0.tgz
--2024-08-31 11:05:14--  https://downloads.apache.org/kafka/3.8.0/kafka_2.13-3.8.0.tgz
Resolviendo downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.208.237
Conectando con downloads.apache.org (downloads.apache.org)[135.181.214.104]:443... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 120735482 (115M) [application/x-gzip]
Guardando como: ‘kafka_2.13-3.8.0.tgz’

kafka_2.13-3.8.0.tgz                   100%[==================================================

2024-08-31 11:05:18 (35,1 MB/s) - ‘kafka_2.13-3.8.0.tgz’ guardado [120735482/120735482]
```

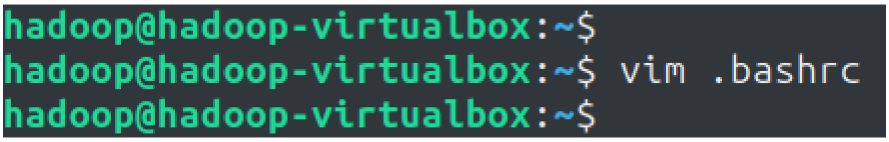


```
hadoop@hadoop-virtualbox:~$ tar -xvzf kafka_2.13-3.8.0.tgz
kafka_2.13-3.8.0/
kafka_2.13-3.8.0/LICENSE
kafka_2.13-3.8.0/NOTICE
kafka_2.13-3.8.0/bin/
kafka_2.13-3.8.0/bin/connect-distributed.sh
kafka_2.13-3.8.0/bin/connect-mirror-maker.sh
kafka_2.13-3.8.0/bin/connect-plugin-path.sh
kafka_2.13-3.8.0/bin/connect-standalone.sh
kafka_2.13-3.8.0/bin/kafka-acls.sh
kafka_2.13-3.8.0/bin/kafka-broker-api-versions.sh
kafka_2.13-3.8.0/bin/kafka-client-metrics.sh
kafka_2.13-3.8.0/bin/kafka-cluster.sh
kafka_2.13-3.8.0/bin/kafka-configs.sh
kafka_2.13-3.8.0/bin/kafka-console-consumer.sh
kafka_2.13-3.8.0/bin/kafka-console-producer.sh
kafka_2.13-3.8.0/bin/kafka-consumer-groups.sh
kafka_2.13-3.8.0/bin/kafka-consumer-perf-test.sh
kafka_2.13-3.8.0/bin/kafka-delegation-tokens.sh
kafka_2.13-3.8.0/bin/kafka-delete-records.sh
```

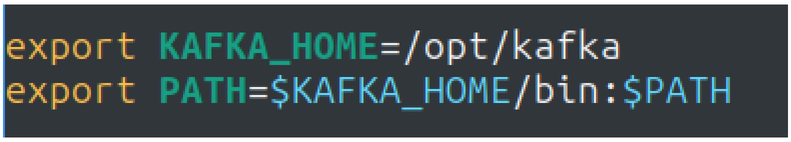


```
hadoop@hadoop-virtualbox:~$ sudo mv kafka_2.13-3.8.0 /opt/kafka
hadoop@hadoop-virtualbox:~$
hadoop@hadoop-virtualbox:~$ ls -lart /opt/kafka
total 80
-rw-r--r-- 1 hadoop hadoop 28359 jul 23 10:04 NOTICE
-rw-r--r-- 1 hadoop hadoop 15295 jul 23 10:04 LICENSE
drwxr-xr-x 2 hadoop hadoop  4096 jul 23 10:09 licenses
drwxr-xr-x 3 hadoop hadoop  4096 jul 23 10:09 config
drwxr-xr-x 3 hadoop hadoop  4096 jul 23 10:09 bin
drwxr-xr-x 7 hadoop hadoop  4096 jul 23 10:09 .
drwxr-xr-x 2 hadoop hadoop  4096 jul 23 10:09 site-docs
drwxr-xr-x 2 hadoop hadoop 12288 ago 31 11:07 libs
drwxr-xr-x 5 root   root    4096 ago 31 11:07 ..
hadoop@hadoop-virtualbox:~$
```

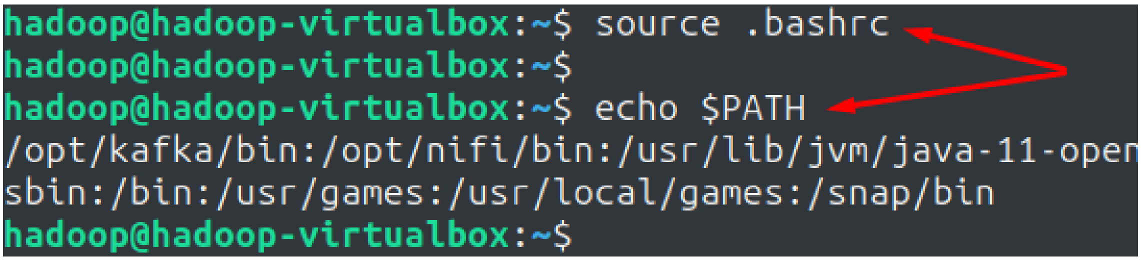
Paso 2. Definimos las variables de entorno para Apache Kafka y lo arrancamos.



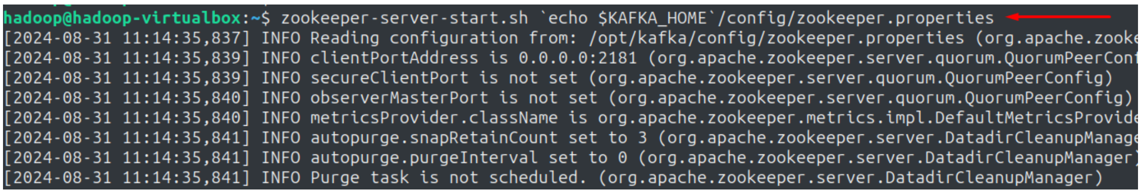
```
hadoop@hadoop-virtualbox:~$
hadoop@hadoop-virtualbox:~$ vim .bashrc
hadoop@hadoop-virtualbox:~$
```



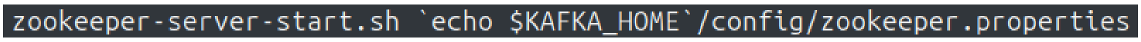
```
export KAFKA_HOME=/opt/kafka
export PATH=$KAFKA_HOME/bin:$PATH
```



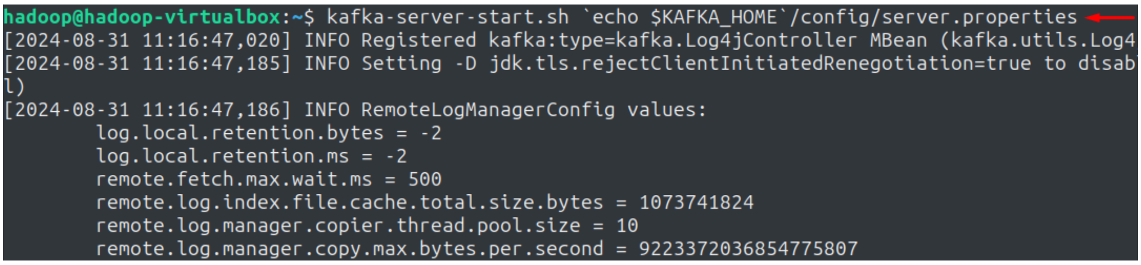
```
hadoop@hadoop-virtualbox:~$ source .bashrc
hadoop@hadoop-virtualbox:~$
hadoop@hadoop-virtualbox:~$ echo $PATH
/opt/kafka/bin:/opt/nifi/bin:/usr/lib/jvm/java-11-open
sbin:/bin:/usr/games:/usr/local/games:/snap/bin
hadoop@hadoop-virtualbox:~$
```



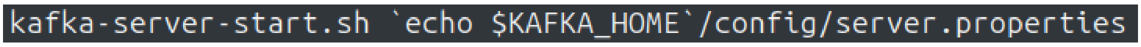
```
hadoop@hadoop-virtualbox:~$ zookeeper-server-start.sh `echo $KAFKA_HOME`/config/zookeeper.properties
[2024-08-31 11:14:35,837] INFO Reading configuration from: /opt/kafka/config/zookeeper.properties (org.apache.zooke
[2024-08-31 11:14:35,839] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConf
[2024-08-31 11:14:35,839] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-08-31 11:14:35,840] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-08-31 11:14:35,840] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvide
[2024-08-31 11:14:35,841] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManage
[2024-08-31 11:14:35,841] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2024-08-31 11:14:35,841] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
```



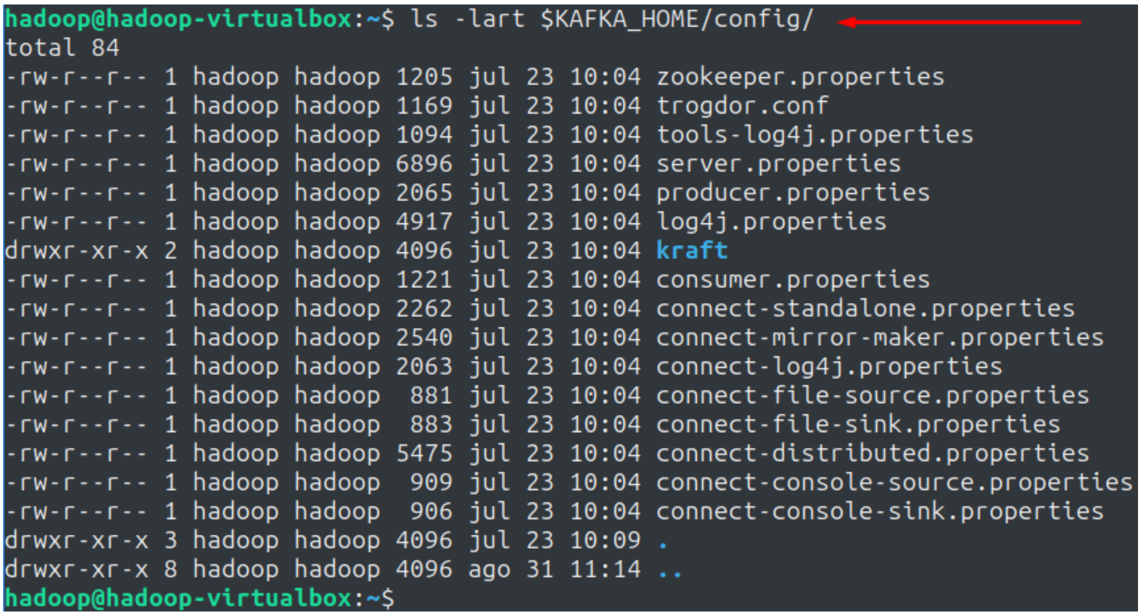
```
zookeeper-server-start.sh `echo $KAFKA_HOME`/config/zookeeper.properties
```



```
hadoop@hadoop-virtualbox:~$ kafka-server-start.sh `echo $KAFKA_HOME`/config/server.properties
[2024-08-31 11:16:47,020] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4
[2024-08-31 11:16:47,185] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disab
l)
[2024-08-31 11:16:47,186] INFO RemoteLogManagerConfig values:
        log.local.retention.bytes = -2
        log.local.retention.ms = -2
        remote.fetch.max.wait.ms = 500
        remote.log.index.file.cache.total.size.bytes = 1073741824
        remote.log.manager.copier.thread.pool.size = 10
        remote.log.manager.copy.max.bytes.per.second = 9223372036854775807
```

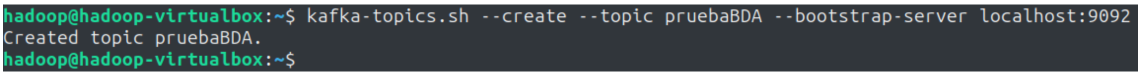


```
kafka-server-start.sh `echo $KAFKA_HOME`/config/server.properties
```

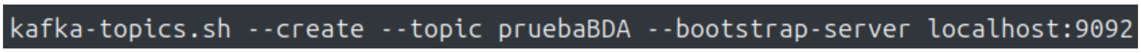


```
hadoop@hadoop-virtualbox:~$ ls -lart $KAFKA_HOME/config/
total 84
-rw-r--r-- 1 hadoop hadoop 1205 jul 23 10:04 zookeeper.properties
-rw-r--r-- 1 hadoop hadoop 1169 jul 23 10:04 trogdor.conf
-rw-r--r-- 1 hadoop hadoop 1094 jul 23 10:04 tools-log4j.properties
-rw-r--r-- 1 hadoop hadoop 6896 jul 23 10:04 server.properties
-rw-r--r-- 1 hadoop hadoop 2065 jul 23 10:04 producer.properties
-rw-r--r-- 1 hadoop hadoop 4917 jul 23 10:04 log4j.properties
drwxr-xr-x 2 hadoop hadoop 4096 jul 23 10:04 kraft
-rw-r--r-- 1 hadoop hadoop 1221 jul 23 10:04 consumer.properties
-rw-r--r-- 1 hadoop hadoop 2262 jul 23 10:04 connect-standalone.properties
-rw-r--r-- 1 hadoop hadoop 2540 jul 23 10:04 connect-mirror-maker.properties
-rw-r--r-- 1 hadoop hadoop 2063 jul 23 10:04 connect-log4j.properties
-rw-r--r-- 1 hadoop hadoop  881 jul 23 10:04 connect-file-source.properties
-rw-r--r-- 1 hadoop hadoop  883 jul 23 10:04 connect-file-sink.properties
-rw-r--r-- 1 hadoop hadoop 5475 jul 23 10:04 connect-distributed.properties
-rw-r--r-- 1 hadoop hadoop  909 jul 23 10:04 connect-console-source.properties
-rw-r--r-- 1 hadoop hadoop  906 jul 23 10:04 connect-console-sink.properties
drwxr-xr-x 3 hadoop hadoop 4096 jul 23 10:09 .
drwxr-xr-x 8 hadoop hadoop 4096 ago 31 11:14 ..
hadoop@hadoop-virtualbox:~$
```

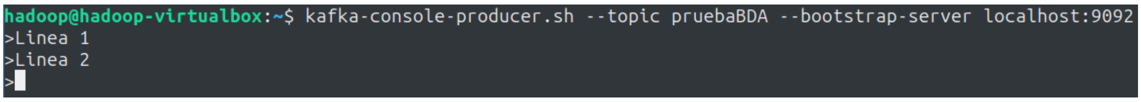
Paso 3. Creamos un tópico para poder publicar y consumir del mismo.



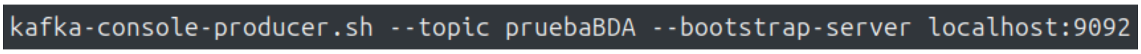
```
hadoop@hadoop-virtualbox:~$ kafka-topics.sh --create --topic pruebaBDA --bootstrap-server localhost:9092
Created topic pruebaBDA.
hadoop@hadoop-virtualbox:~$
```



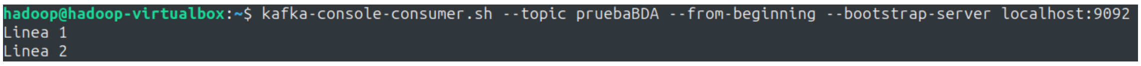
```
kafka-topics.sh --create --topic pruebaBDA --bootstrap-server localhost:9092
```



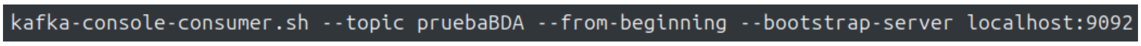
```
hadoop@hadoop-virtualbox:~$ kafka-console-producer.sh --topic pruebaBDA --bootstrap-server localhost:9092
>Linea 1
>Linea 2
>
```



```
kafka-console-producer.sh --topic pruebaBDA --bootstrap-server localhost:9092
```

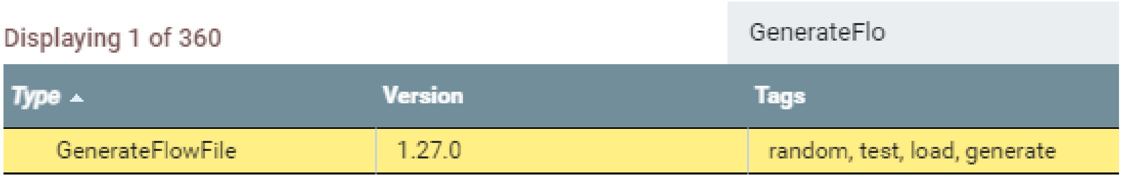


```
hadoop@hadoop-virtualbox:~$ kafka-console-consumer.sh --topic pruebaBDA --from-beginning --bootstrap-server localhost:9092
Linea 1
Linea 2
```



```
kafka-console-consumer.sh --topic pruebaBDA --from-beginning --bootstrap-server localhost:9092
```

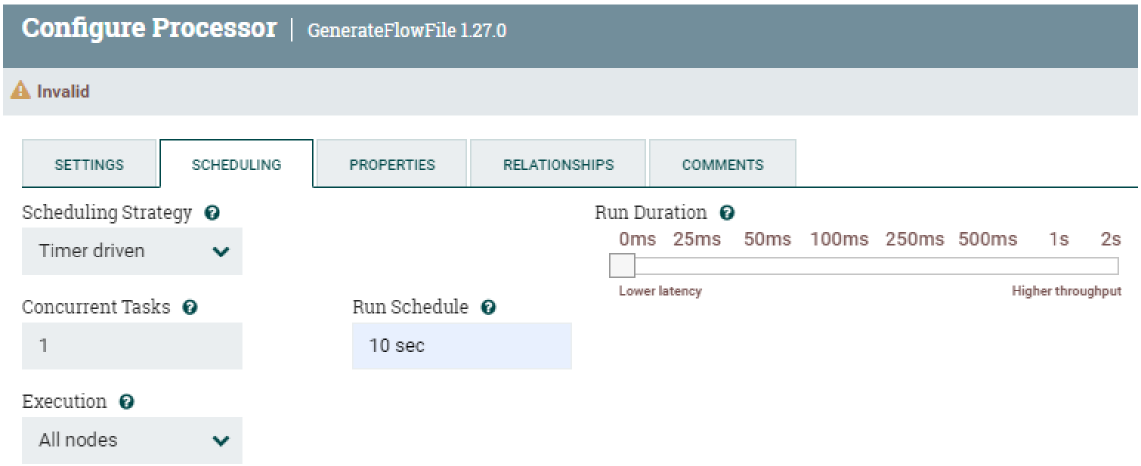
Paso 4. Añadimos los procesos para publicar en Kafka desde NiFi.



Displaying 1 of 360

GenerateFlo

Type ▲	Version	Tags
GenerateFlowFile	1.27.0	random, test, load, generate



Configure Processor | GenerateFlowFile 1.27.0

⚠ Invalid

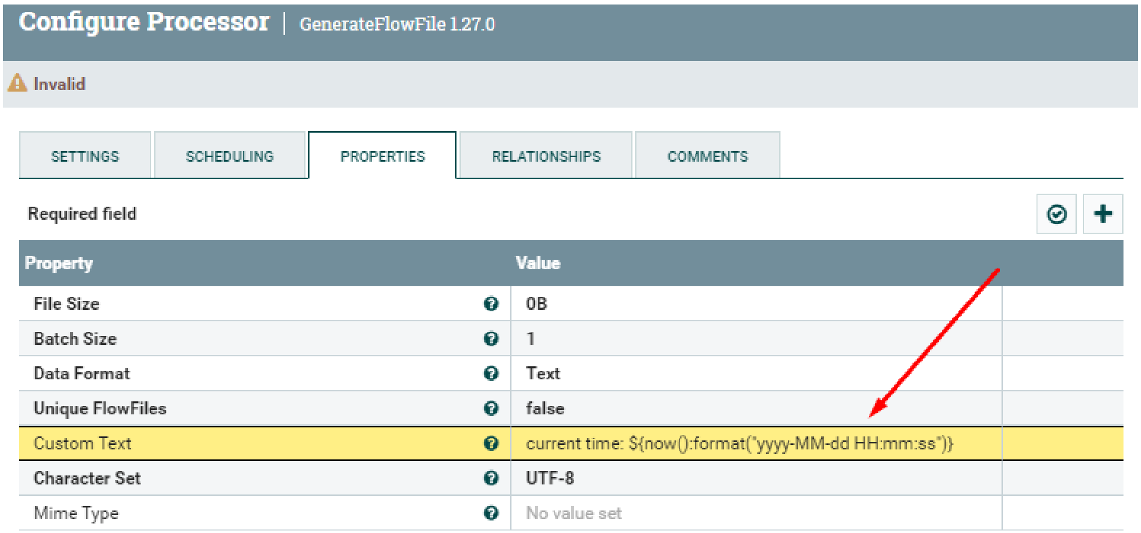
Scheduling Strategy: Timer driven

Concurrent Tasks: 1

Run Schedule: 10 sec

Execution: All nodes

Run Duration: 0ms 25ms 50ms 100ms 250ms 500ms 1s 2s

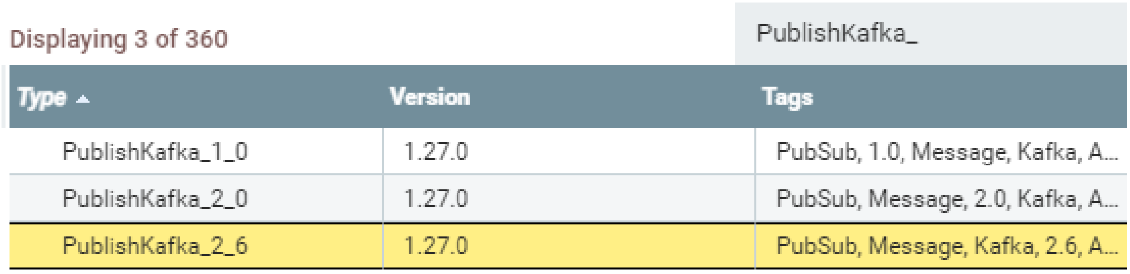


Configure Processor | GenerateFlowFile 1.27.0

⚠ Invalid

Required field

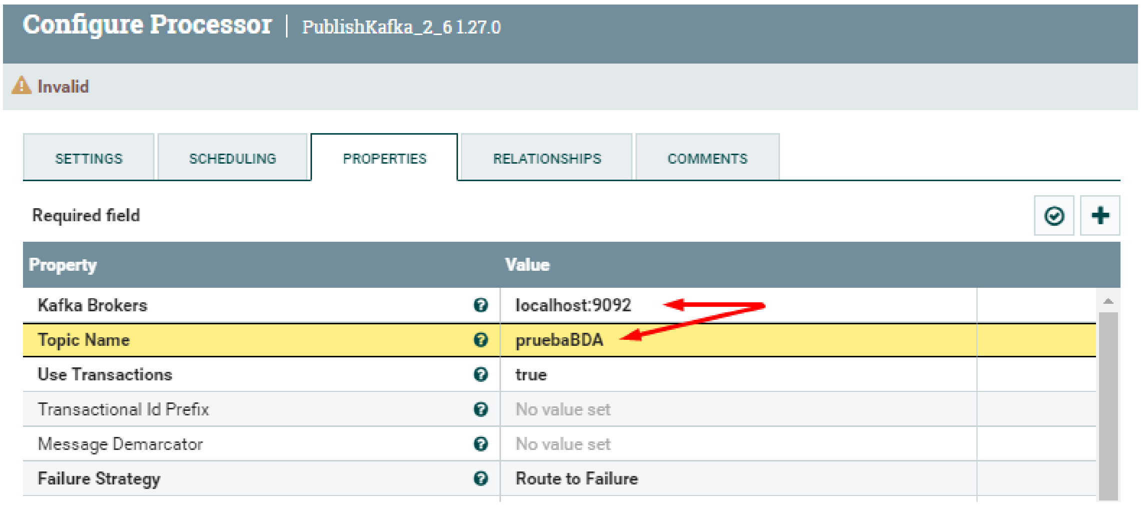
Property	Value
File Size	0B
Batch Size	1
Data Format	Text
Unique FlowFiles	false
Custom Text	current time: ${now():format("yyyy-MM-dd HH:mm:ss")}
Character Set	UTF-8
Mime Type	No value set



Displaying 3 of 360

PublishKafka_

Type ▲	Version	Tags
PublishKafka_1_0	1.27.0	PubSub, 1.0, Message, Kafka, A...
PublishKafka_2_0	1.27.0	PubSub, Message, 2.0, Kafka, A...
PublishKafka_2_6	1.27.0	PubSub, Message, Kafka, 2.6, A...

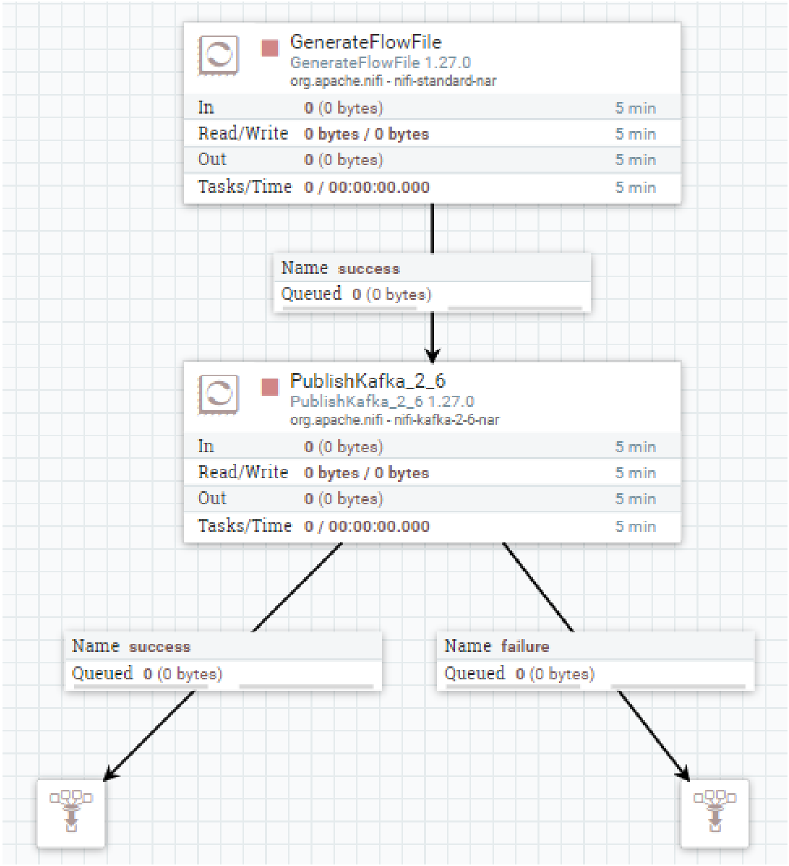


Configure Processor | PublishKafka_2_6 1.27.0

⚠ Invalid

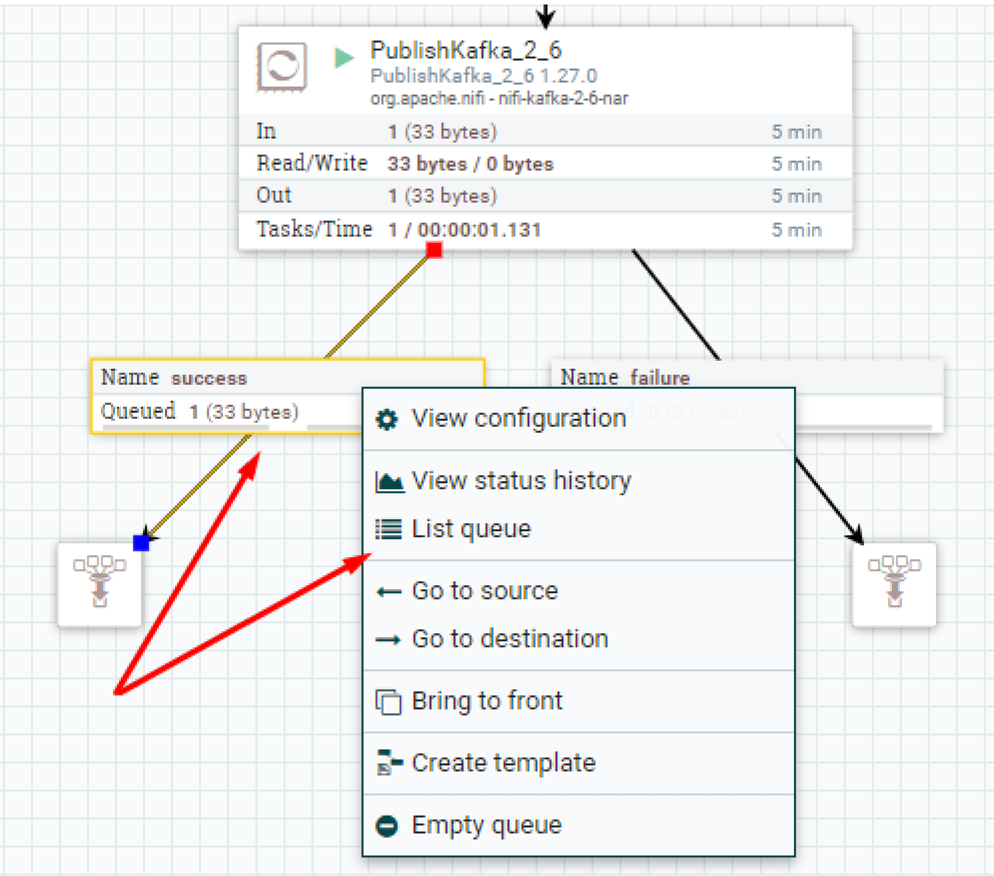
Required field

Property	Value
Kafka Brokers	localhost:9092
Topic Name	pruebaBDA
Use Transactions	true
Transactional Id Prefix	No value set
Message Demarcator	No value set
Failure Strategy	Route to Failure



```
GenerateFlowFile
GenerateFlowFile 1.27.0
org.apache.nifi - nifi-standard-nar
In           0 (0 bytes)            5 min
Read/Write   0 bytes / 0 bytes      5 min
Out          0 (0 bytes)            5 min
Tasks/Time   0 / 00:00:00.000       5 min

Name  success
Queued  0 (0 bytes)

PublishKafka_2_6
PublishKafka_2_6 1.27.0
org.apache.nifi - nifi-kafka-2-6-nar
In           0 (0 bytes)            5 min
Read/Write   0 bytes / 0 bytes      5 min
Out          0 (0 bytes)            5 min
Tasks/Time   0 / 00:00:00.000       5 min

Name  success
Queued  0 (0 bytes)

Name  failure
Queued  0 (0 bytes)
```

Paso 5. Comprobamos el flujo.

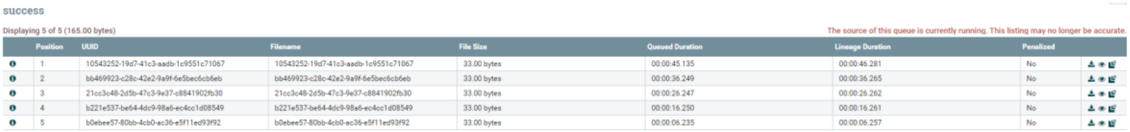


PublishKafka_2_6
PublishKafka_2_6 1.27.0
org.apache.nifi - nifi-kafka-2-6-nar

```
In            1 (33 bytes)                5 min
Read/Write    33 bytes / 0 bytes          5 min
Out           1 (33 bytes)                5 min
Tasks/Time    1 / 00:00:01.131            5 min
```

Name success
Queued 1 (33 bytes)

Name failure

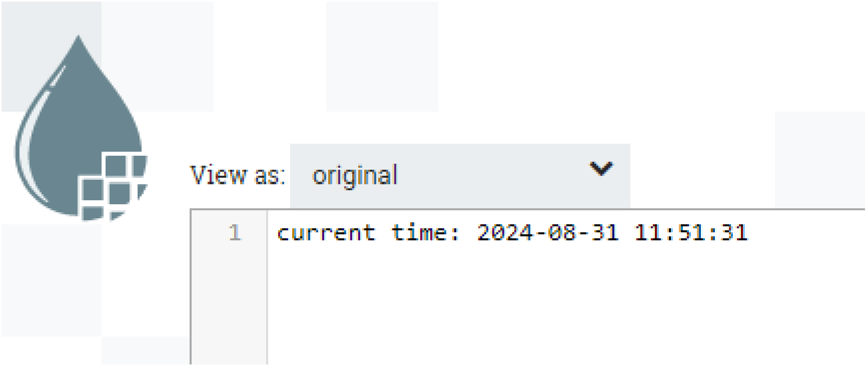


success

Displaying 5 of 5 (165.00 bytes)

The source of this queue is currently running. This listing may no longer be accurate.

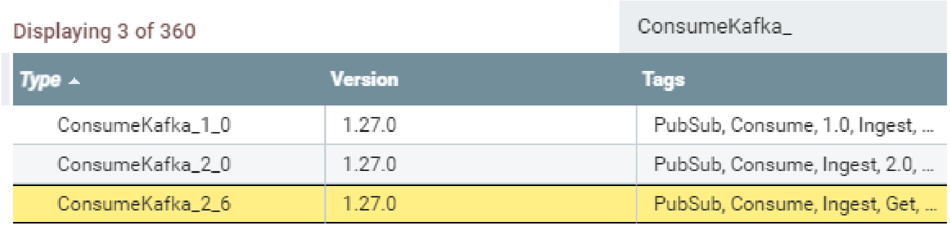
	Position	UUID	Filename	File Size	Queued Duration	Lineage Duration	Penalized	
ℹ	1	10543252-19d7-41c3-aadb-1c9551c71067	10543252-19d7-41c3-aadb-1c9551c71067	33.00 bytes	00:00:45.135	00:00:46.281	No	
ℹ	2	bb469923-c28c-42e2-9a9f-6e5bec6cb6eb	bb469923-c28c-42e2-9a9f-6e5bec6cb6eb	33.00 bytes	00:00:36.249	00:00:36.265	No	
ℹ	3	21cc3c48-2d5b-47c3-9e37-c8841902fb30	21cc3c48-2d5b-47c3-9e37-c8841902fb30	33.00 bytes	00:00:26.247	00:00:26.262	No	
ℹ	4	b221e537-be64-4dc9-98a6-ec4cc1d08549	b221e537-be64-4dc9-98a6-ec4cc1d08549	33.00 bytes	00:00:16.250	00:00:16.261	No	
ℹ	5	b0ebee57-80bb-4cb0-ac36-e5f11ed93f92	b0ebee57-80bb-4cb0-ac36-e5f11ed93f92	33.00 bytes	00:00:06.235	00:00:06.257	No	



View as: original

```
current time: 2024-08-31 11:51:31
```

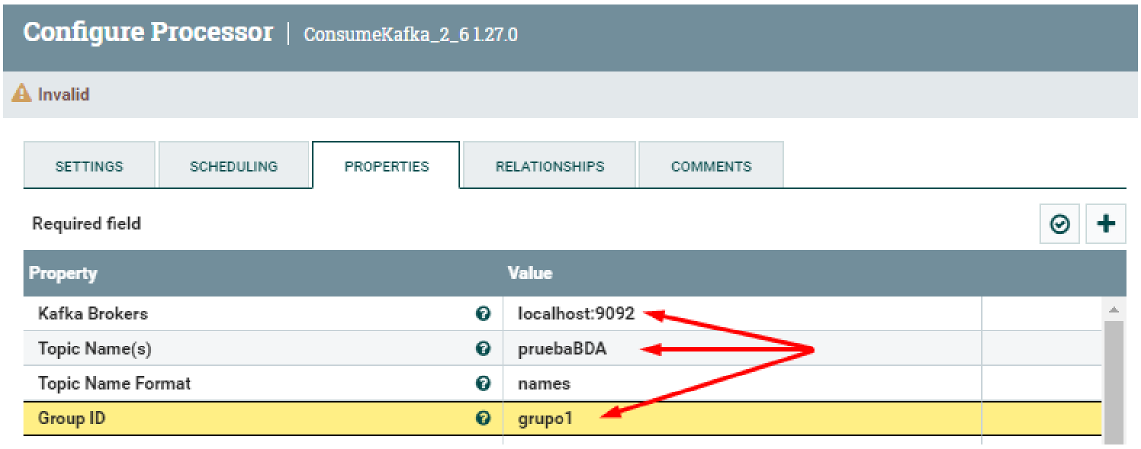
Paso 6. Añadimos los procesos para consumir en Kafka desde NiFi.



Displaying 3 of 360

ConsumeKafka_

Type ▲	Version	Tags
ConsumeKafka_1_0	1.27.0	PubSub, Consume, 1.0, Ingest, ...
ConsumeKafka_2_0	1.27.0	PubSub, Consume, Ingest, 2.0, ...
ConsumeKafka_2_6	1.27.0	PubSub, Consume, Ingest, Get, ...

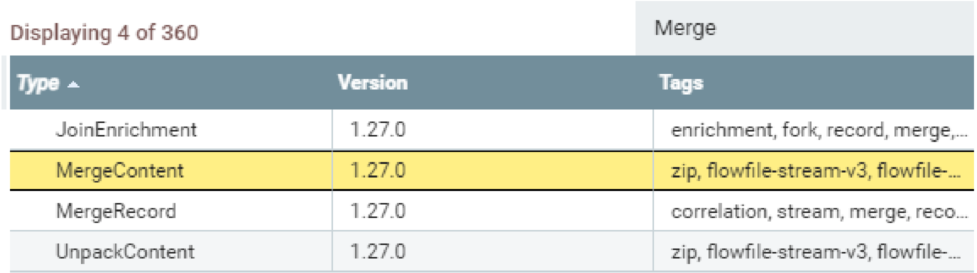


Configure Processor | ConsumeKafka_2_6 1.27.0

⚠ Invalid

Required field

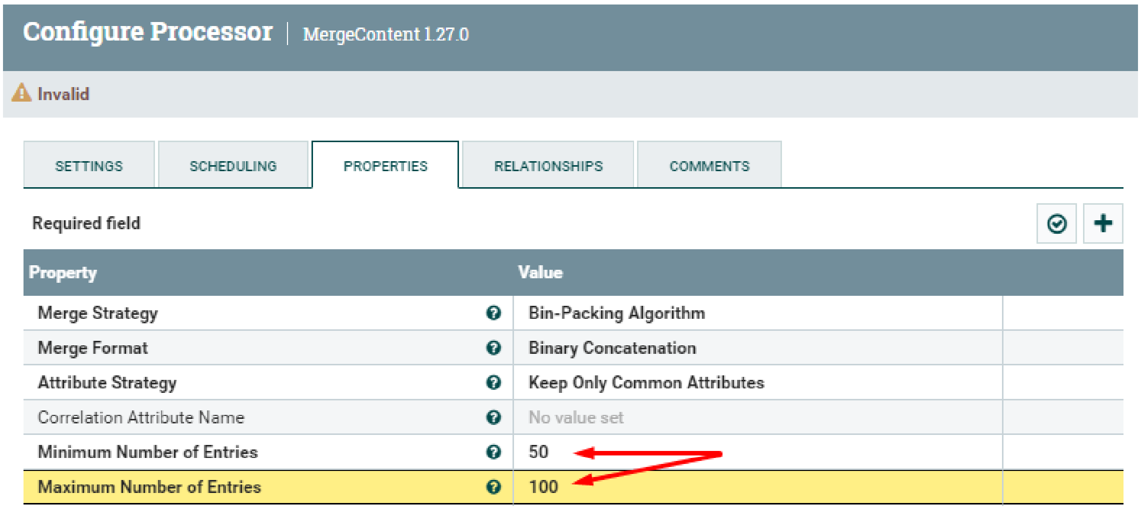
Property	Value
Kafka Brokers	localhost:9092
Topic Name(s)	pruebaBDA
Topic Name Format	names
Group ID	grupo1



Displaying 4 of 360

Merge

Type ▲	Version	Tags
JoinEnrichment	1.27.0	enrichment, fork, record, merge,...
MergeContent	1.27.0	zip, flowfile-stream-v3, flowfile-...
MergeRecord	1.27.0	correlation, stream, merge, reco...
UnpackContent	1.27.0	zip, flowfile-stream-v3, flowfile-...

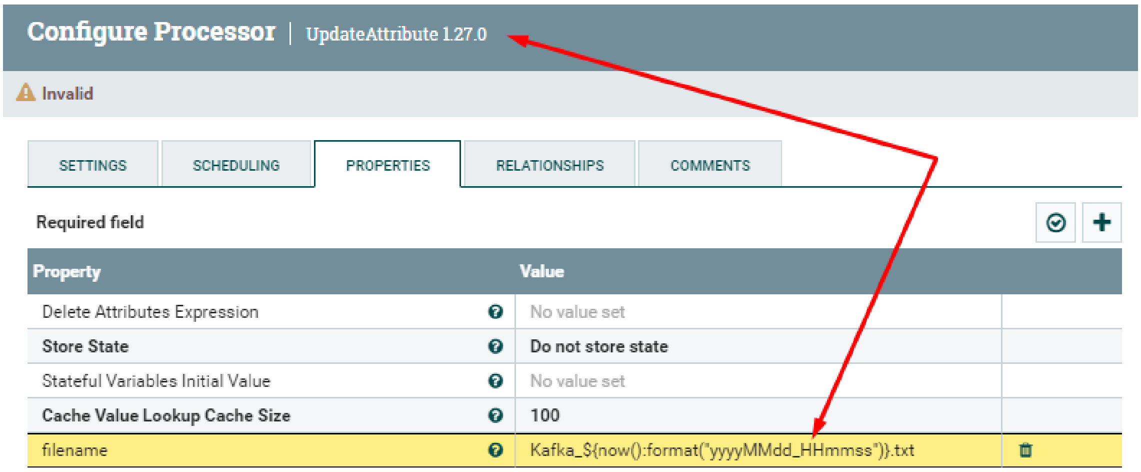


Configure Processor | MergeContent 1.27.0

⚠ Invalid

Required field

Property	Value
Merge Strategy	Bin-Packing Algorithm
Merge Format	Binary Concatenation
Attribute Strategy	Keep Only Common Attributes
Correlation Attribute Name	No value set
Minimum Number of Entries	50
Maximum Number of Entries	100



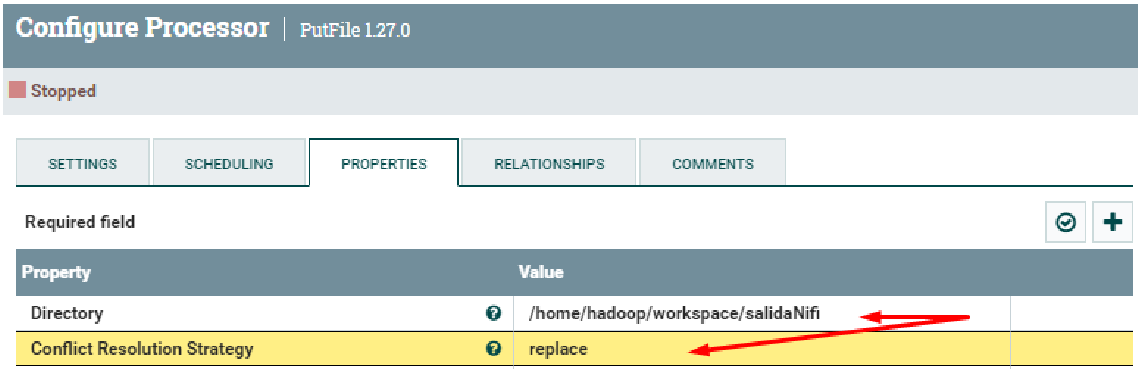
Configure Processor | UpdateAttribute 1.27.0

⚠ Invalid

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Delete Attributes Expression	No value set
Store State	Do not store state
Stateful Variables Initial Value	No value set
Cache Value Lookup Cache Size	100
filename	Kafka_${now():format("yyyyMMdd_HHmmss")}.txt



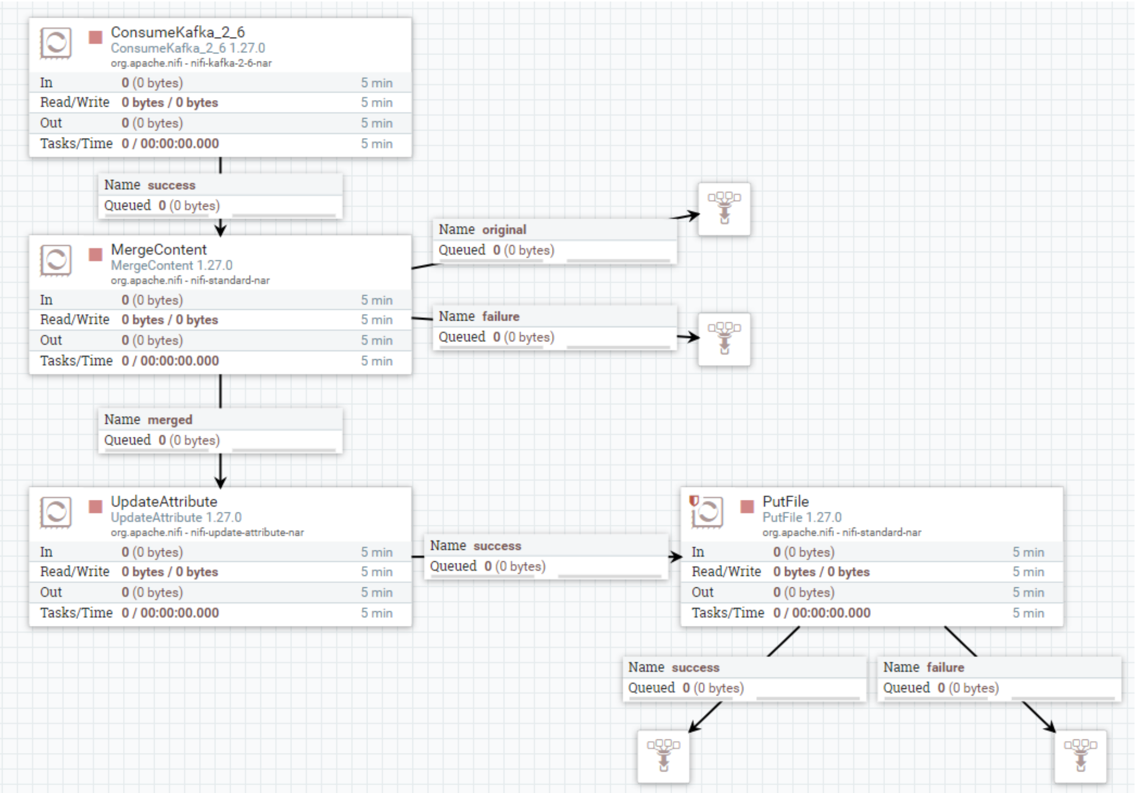
Configure Processor | PutFile 1.27.0

Stopped

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Directory	/home/hadoop/workspace/salidaNifi
Conflict Resolution Strategy	replace



ConsumeKafka_2_6
ConsumeKafka_2_6 1.27.0
org.apache.nifi - nifi-kafka-2-6-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

MergeContent
MergeContent 1.27.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name original
Queued 0 (0 bytes)

Name failure
Queued 0 (0 bytes)

Name merged
Queued 0 (0 bytes)

UpdateAttribute
UpdateAttribute 1.27.0
org.apache.nifi - nifi-update-attribute-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

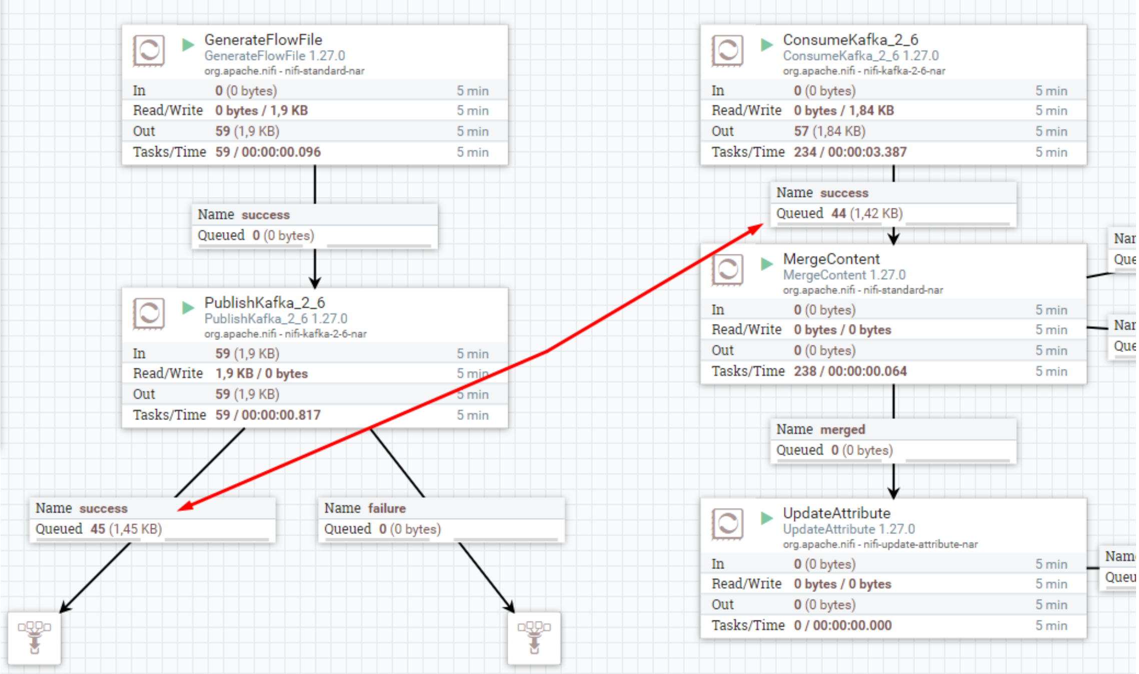
PutFile
PutFile 1.27.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

Name failure
Queued 0 (0 bytes)

Paso 7. Comprobamos el flujo.



GenerateFlowFile
GenerateFlowFile 1.27.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 1,9 KB 5 min
Out 59 (1,9 KB) 5 min
Tasks/Time 59 / 00:00:00.096 5 min

Name success
Queued 0 (0 bytes)

PublishKafka_2_6
PublishKafka_2_6 1.27.0
org.apache.nifi - nifi-kafka-2-6-nar
In 59 (1,9 KB) 5 min
Read/Write 1,9 KB / 0 bytes 5 min
Out 59 (1,9 KB) 5 min
Tasks/Time 59 / 00:00:00.817 5 min

Name success
Queued 45 (1,45 KB)

Name failure
Queued 0 (0 bytes)

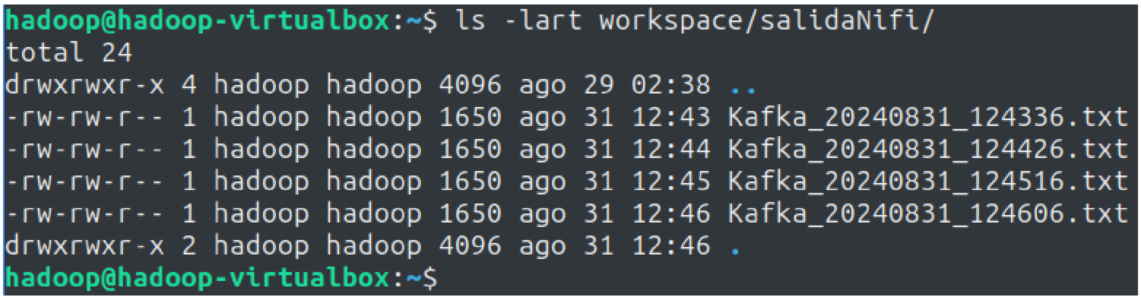
ConsumeKafka_2_6
ConsumeKafka_2_6 1.27.0
org.apache.nifi - nifi-kafka-2-6-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 1,84 KB 5 min
Out 57 (1,84 KB) 5 min
Tasks/Time 234 / 00:00:03.387 5 min

Name success
Queued 44 (1,42 KB)

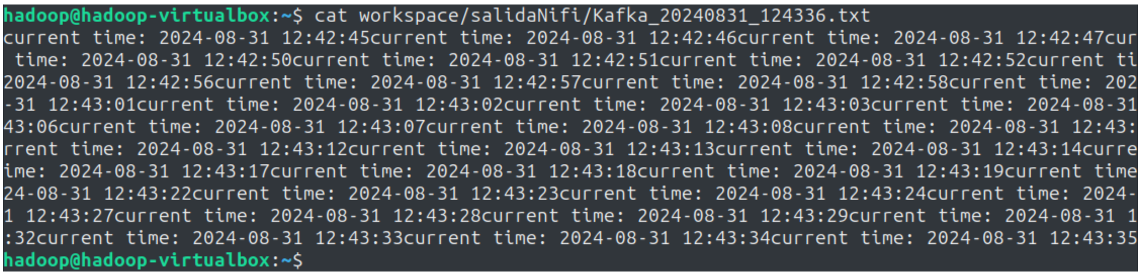
MergeContent
MergeContent 1.27.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 238 / 00:00:00.064 5 min

Name merged
Queued 0 (0 bytes)

UpdateAttribute
UpdateAttribute 1.27.0
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min



```
hadoop@hadoop-virtualbox:~$ ls -lart workspace/salidaNifi/
total 24
drwxrwxr-x 4 hadoop hadoop 4096 ago 29 02:38 ..
-rw-rw-r-- 1 hadoop hadoop 1650 ago 31 12:43 Kafka_20240831_124336.txt
-rw-rw-r-- 1 hadoop hadoop 1650 ago 31 12:44 Kafka_20240831_124426.txt
-rw-rw-r-- 1 hadoop hadoop 1650 ago 31 12:45 Kafka_20240831_124516.txt
-rw-rw-r-- 1 hadoop hadoop 1650 ago 31 12:46 Kafka_20240831_124606.txt
drwxrwxr-x 2 hadoop hadoop 4096 ago 31 12:46 .
hadoop@hadoop-virtualbox:~$
```



```
hadoop@hadoop-virtualbox:~$ cat workspace/salidaNifi/Kafka_20240831_124336.txt
current time: 2024-08-31 12:42:45current time: 2024-08-31 12:42:46current time: 2024-08-31 12:42:47cur
 time: 2024-08-31 12:42:50current time: 2024-08-31 12:42:51current time: 2024-08-31 12:42:52current ti
2024-08-31 12:42:56current time: 2024-08-31 12:42:57current time: 2024-08-31 12:42:58current time: 202
-31 12:43:01current time: 2024-08-31 12:43:02current time: 2024-08-31 12:43:03current time: 2024-08-31
43:06current time: 2024-08-31 12:43:07current time: 2024-08-31 12:43:08current time: 2024-08-31 12:43:
rrent time: 2024-08-31 12:43:12current time: 2024-08-31 12:43:13current time: 2024-08-31 12:43:14curre
ime: 2024-08-31 12:43:17current time: 2024-08-31 12:43:18current time: 2024-08-31 12:43:19current time
24-08-31 12:43:22current time: 2024-08-31 12:43:23current time: 2024-08-31 12:43:24current time: 2024-
1 12:43:27current time: 2024-08-31 12:43:28current time: 2024-08-31 12:43:29current time: 2024-08-31 1
:32current time: 2024-08-31 12:43:33current time: 2024-08-31 12:43:34current time: 2024-08-31 12:43:35
hadoop@hadoop-virtualbox:~$
```