

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

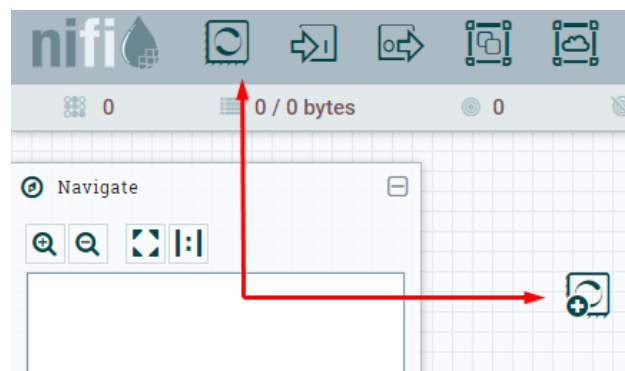
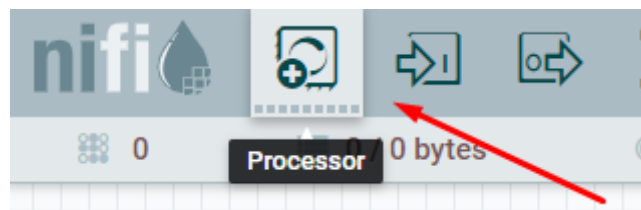
Unidad Didáctica 1. Estrategias de ingestión y almacenamiento de datos en Big Data

Práctica 3. Practicando el manejo de ficheros.

Paso 1. Preparamos los directorios de trabajo y seleccionamos el proceso adecuado para generar los ficheros.

```
hadoop@hadoop-virtualbox:~$ mkdir -p workspace/entradaNifi
hadoop@hadoop-virtualbox:~$ mkdir -p workspace/salidaNifi
hadoop@hadoop-virtualbox:~$
hadoop@hadoop-virtualbox:~$ ls -lart workspace/
total 16
drwxrwxr-x  2 hadoop hadoop 4096 ago 29 02:38 entradaNifi
drwxr-x--- 15 hadoop hadoop 4096 ago 29 02:38 ..
drwxrwxr-x  2 hadoop hadoop 4096 ago 29 02:38 salidaNifi
drwxrwxr-x  4 hadoop hadoop 4096 ago 29 02:38 .
hadoop@hadoop-virtualbox:~$
```

Para añadir un proceso colocaremos el ratón sobre el icono y lo arrastraremos al área de trabajo.



Add Processor

Source

all groups

amazon attributes
aws azure cloud
consume csv
delete fetch get
google ingest
json listen logs
message
microsoft put
query record
restricted source
storage text
update

Displaying 1 of 360

| Type | Version | Tags |
|------------------|---------|------------------------------|
| GenerateFlowFile | 1.27.0 | random, test, load, generate |

GenerateFlowFile 1.27.0

org.apache.nifi - nifi-standard-nar

This processor creates FlowFiles with random data or custom content. GenerateFlowFile is useful for load testing, configuration, and simulation. Also see DuplicateFlowFile for additional load testing.

CANCEL

ADD

nifi


0 0 / 0 bytes 0 0 0 0

Navigate

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 |

Configure Processor | GenerateFlowFile 1.27.0

 Invalid

SETTINGS

SCHEDULING

PROPERTIES

RELATIONSHIPS

COMMENTS

Scheduling Strategy ?
Timer driven ▼

Concurrent Tasks ?
1


Execution ?
All nodes ▼

Run Duration ?
0ms 25ms 50ms

Lower latency

Run Schedule ?
20 sec

Configure Processor | GenerateFlowFile 1.27.0

 Invalid



SETTINGS

SCHEDULING

PROPERTIES

RELATIONSHIPS

COMMENTS

Required field  

| Property | Value |
|--------------------|--------------|
| File Size ? | 0B |
| Batch Size ? | 1 |
| Data Format ? | Text |
| Unique FlowFiles ? | false |
| Custom Text ? | No value set |
| Character Set ? | UTF-8 |
| Mime Type ? | No value set |

EL ✓ PARAM ✓

1

Fichero creado: \${now():format("yyyy-MM-dd HH:mm:ss")}

☐ Set empty string

CANCEL

OK

Paso 2. Definimos el proceso para cambiar el atributo “filename” del Flowfile creado.

Displaying 1 of 360

updatea

| Type ^ | Version | Tags |
|-----------------|---------|------------------------------------|
| UpdateAttribute | 1.27.0 | Attribute Expression Language, ... |

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 |

Add Property

Property Name

filename

Sensitive Value ?

☐ Yes ☒ No

CANCEL

OK

EL ✓ PARAM ✓

1 ficheros_\${now():format("yyyyMMddHHmmss")}

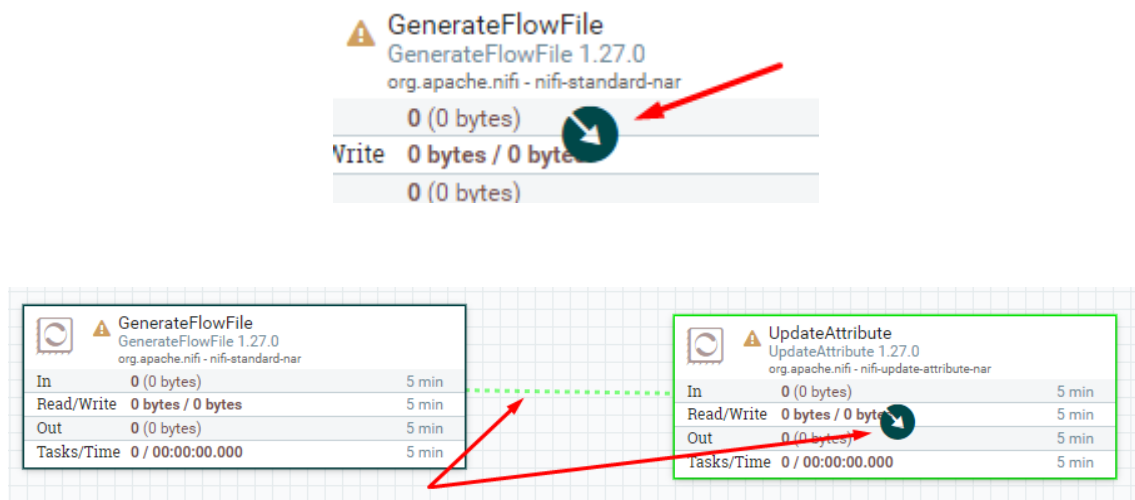
☐ Set empty string

CANCEL

OK

| 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 | ficheros_\${now()}.format("yyyyMMddHHmmss") |  |

Con los dos procesos creados, colocamos el puntero sobre el primer proceso y pulsando el botón izquierdo arrastramos la flecha hacia el segundo para enlazarlos.



Create Connection

DETAILS
SETTINGS

Name

Id
No value set

FlowFile Expiration
0 sec

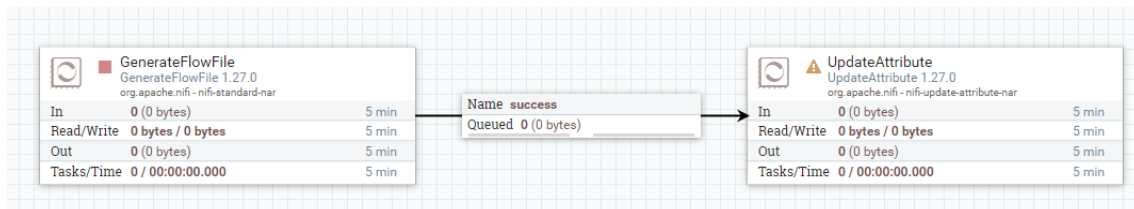
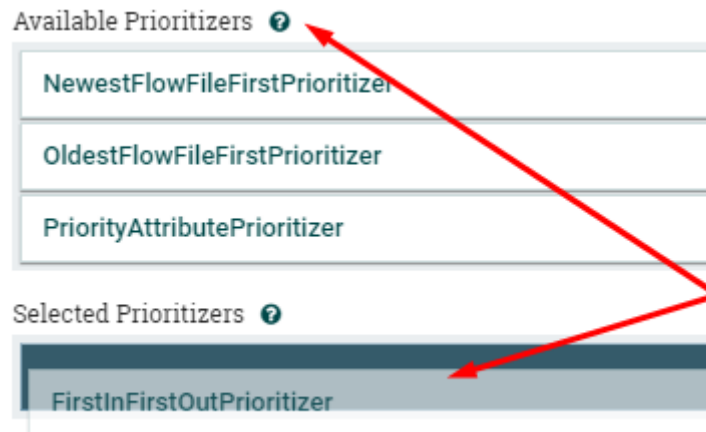
Back Pressure
Object Threshold
10000

Size Threshold
1 GB

Load Balance Strategy
Do not load balance

Available Prioritizers
FirstInFirstOutPrioritizer
NewestFlowFileFirstPrioritizer
OldestFlowFileFirstPrioritizer
PriorityAttributePrioritizer

Selected Prioritizers



Paso 3. Damos de alta un nuevo proceso para colocar el FlowFile creado en el directorio de entrada (posteriormente, en otro flujo, se moverá de entrada a salida).

Displaying 1 of 360

| Putfile | | |
|---------|---------|--|
| Type ^ | Version | Tags |
| PutFile | 1.27.0 | restricted, files, archive, copy, p... |

Configure Processor | PutFile 1.27.0

Invalid

| SETTINGS | | | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS |
|------------------------------|-------|--------------|------------|------------|---------------|----------|
| Required field | | | | | | |
| Property | Value | | | | | |
| Directory | ? | No value set | | | | |
| Conflict Resolution Strategy | ? | fail | | | | |
| Create Missing Directories | ? | true | | | | |

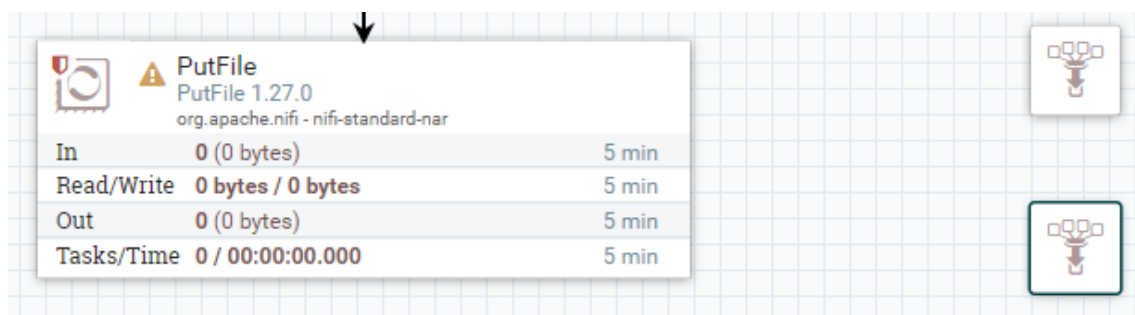
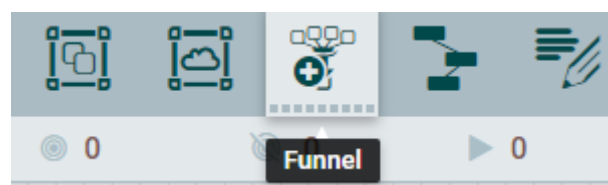
Required field

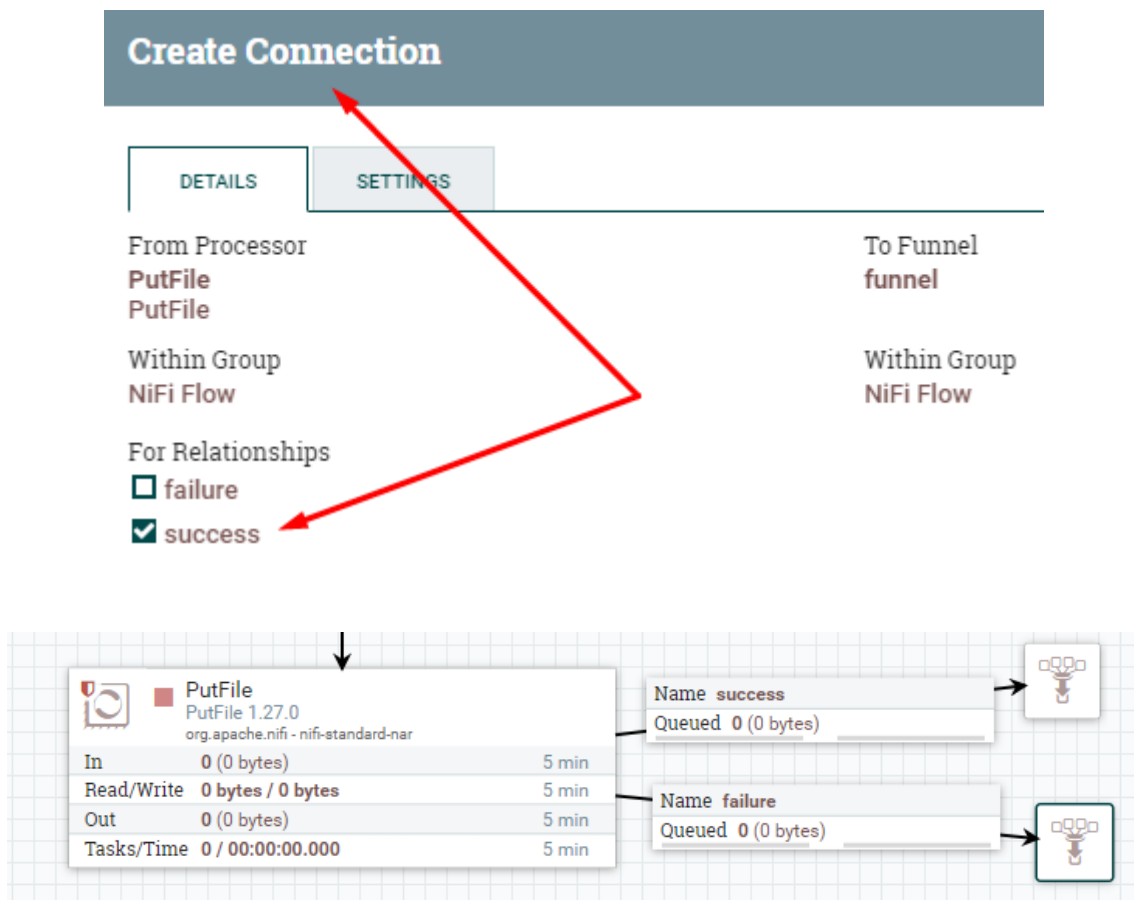
| Property | Value |
|------------------------------|--------------------------------------|
| Directory | |
| Conflict Resolution Strategy | 1 /home/hadoop/workspace/entradaNifi |
| Create Missing Directories | |

Required field

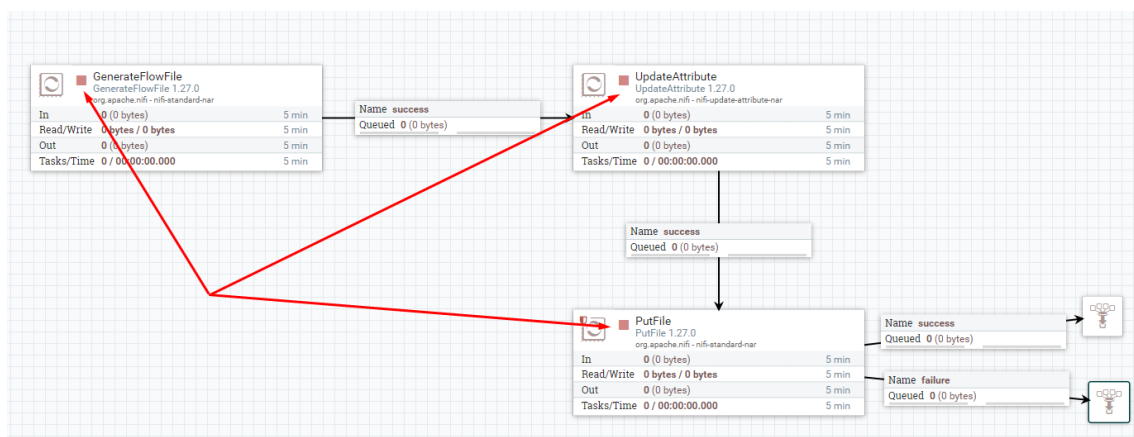
| Property | Value |
|------------------------------|------------------------|
| Directory | |
| Conflict Resolution Strategy | fail |
| Create Missing Directories | replace |
| Maximum File Count | ignore |
| Last Modified Time | fail |
| Permissions | Reference parameter... |
| Owner | |

Paso 4. Añadimos dos elementos “Funnel” para ver en sus colas si los ficheros se han colocado correctamente o si ha fallado.



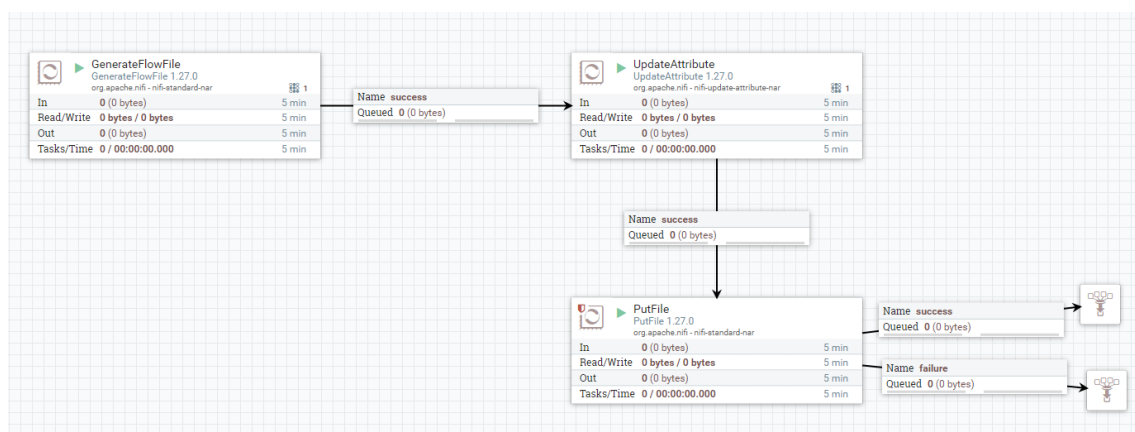
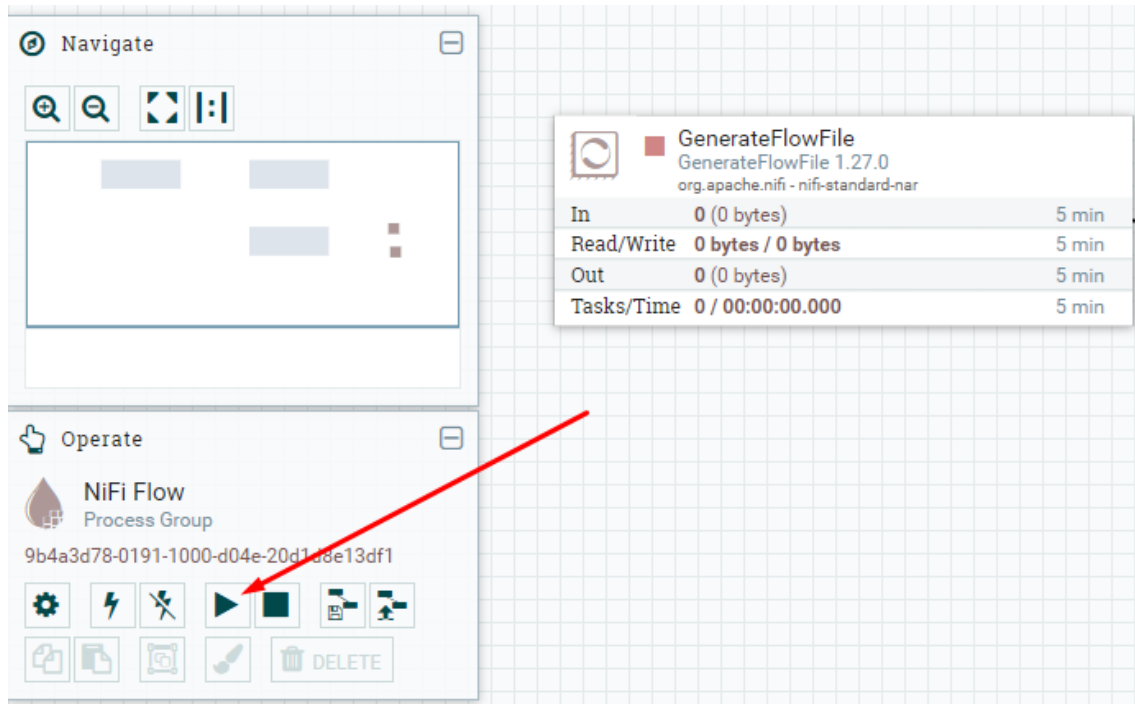


Creadas las conexiones, podemos comprobar que todos los procesos del flujo se pueden arrancar, ya que todos tienen el símbolo de parada.

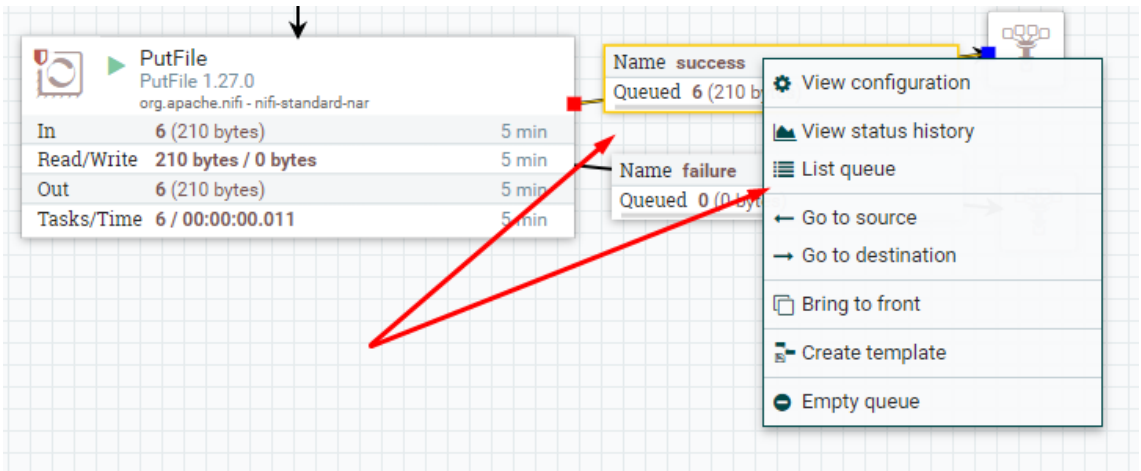


Paso 5. Arranco el flujo y lo compruebo.

Para arrancar los procesos puedo seleccionar todo con Shift o pulsar sobre el área de trabajo y arrancarlo desde el apartado de “operar”.



```
hadoop@hadoop-virtualbox:~$ ls -lart workspace/entradaNifi/
total 24
drwxrwxr-x 4 hadoop hadoop 4096 ago 29 02:38 ..
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:07 ficheros_20240829030723
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:07 ficheros_20240829030743
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:08 ficheros_20240829030803
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:08 ficheros_20240829030823
drwxrwxr-x 2 hadoop hadoop 4096 ago 29 03:08 .
hadoop@hadoop-virtualbox:~$
hadoop@hadoop-virtualbox:~$ cat workspace/entradaNifi/ficheros_20240829030723
Fichero creado: 2024-08-29 03:07:23hadoop@hadoop-virtualbox:~$
hadoop@hadoop-virtualbox:~$
```



success

Displaying 7 of 7 (245.00 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 |
|----------|--------------------------------------|-------------------------|-------------|-----------------|------------------|-----------|
| 1 | 4535c797-9328-4815-9268-031fecbcb4f | ficheros_20240829030723 | 35.00 bytes | 00:02:17.886 | 00:02:17.893 | No |
| 2 | db35bb88-fe26-4484-be1e-52a049ab5e9d | ficheros_20240829030743 | 35.00 bytes | 00:01:57.887 | 00:01:57.912 | No |
| 3 | 3665cc75-a940-4d7d-95ae-1d9a91a707e8 | ficheros_20240829030803 | 35.00 bytes | 00:01:37.886 | 00:01:37.905 | No |
| 4 | fa26b74c-892a-4471-9b98-1a2d027a92b3 | ficheros_20240829030823 | 35.00 bytes | 00:01:17.883 | 00:01:17.900 | No |
| 5 | 558360c1-2360-4cc3-acc3-7249a2541f01 | ficheros_20240829030843 | 35.00 bytes | 00:00:57.873 | 00:00:57.895 | No |
| 6 | 767f0f51-aada-437a-bca7-20481176413e | ficheros_20240829030903 | 35.00 bytes | 00:00:37.859 | 00:00:37.881 | No |
| 7 | b652c710-9a0e-4cbe-af55-2eeddc45efc7 | ficheros_20240829030923 | 35.00 bytes | 00:00:17.856 | 00:00:17.877 | No |

success

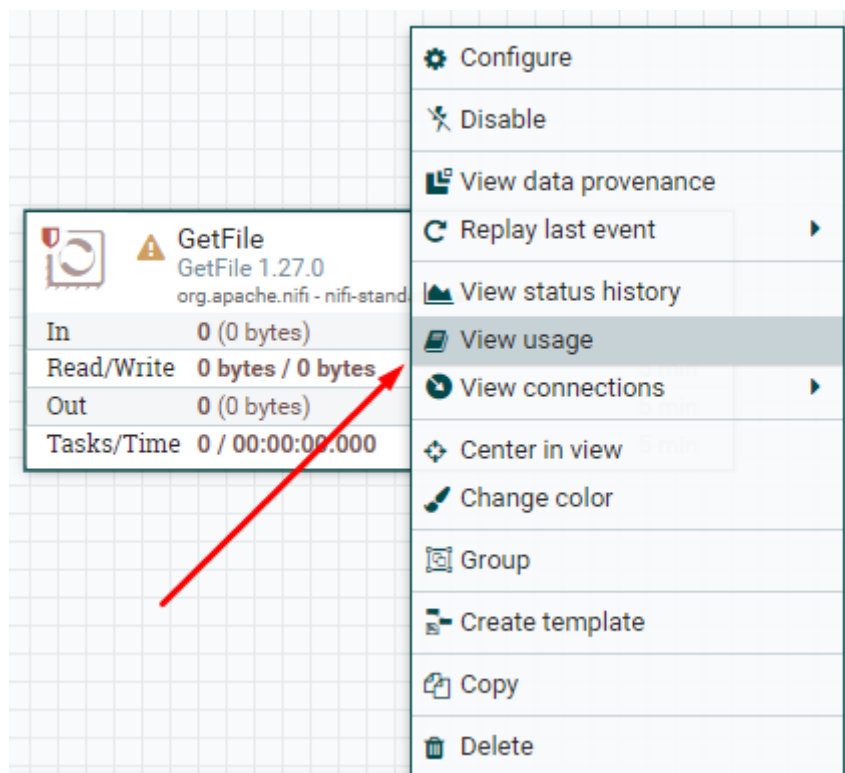
Displaying 7 of 7 (245.00 bytes)

| Position | UUID | Filename | File Size |
|----------|--------------------------------------|-------------------------|-------------|
| 1 | 4535c797-9328-4815-9268-031fecbcb4f | ficheros_20240829030723 | 35.00 bytes |
| 2 | db35bb88-fe26-4484-be1e-52a049ab5e9d | ficheros_20240829030743 | 35.00 bytes |
| 3 | 3665cc75-a940-4d7d-95ae-1d9a91a707e8 | ficheros_20240829030803 | 35.00 bytes |
| 4 | fa26b74c-892a-4471-9b98-1a2d027a92b3 | ficheros_20240829030823 | 35.00 bytes |
| 5 | 558360c1-2360-4cc3-acc3-7249a2541f01 | ficheros_20240829030843 | 35.00 bytes |
| 6 | 767f0f51-aada-437a-bca7-20481176413e | ficheros_20240829030903 | 35.00 bytes |
| 7 | b652c710-9a0e-4cbe-af55-2eeddc45efc7 | ficheros_20240829030923 | 35.00 bytes |

Paso 6. Creamos otro flujo que coja los ficheros del directorio de entrada y los ponga en el de salida.

| Displaying 1 of 360 | | | Getfil |
|---|---------|---|--------|
| Type ^ | Version | Tags | |
|  GetFile | 1.27.0 | ingress, input, restricted, get, fil... | |

Con la opción “View usage” del menú contextual del proceso podría acceder a la documentación de una forma sencilla.



- NiFi Documentation 1.27.0 GetFile 1.27.0

General

- Overview
- Getting Started
- User Guide
- Expression Language Guide
- RecordPath Guide
- Admin Guide
- Toolkit Guide
- Walkthroughs

Description:

Creates FlowFiles from files in a directory. NiFi will ignore files it doesn't have at least read permissions for.

Tags:

local, files, filesystem, ingest, ingress, get, source, input

Properties:

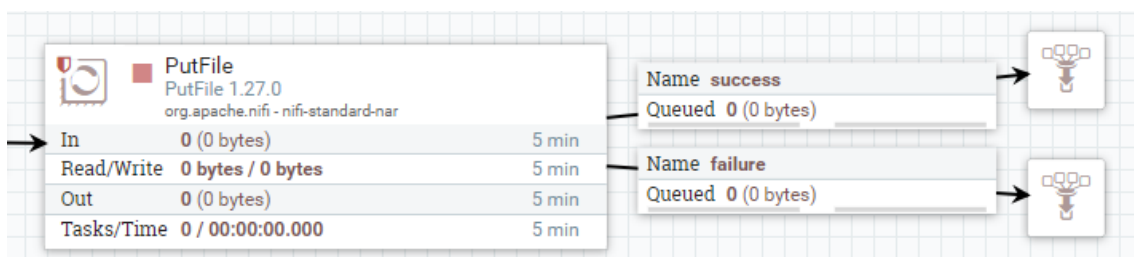
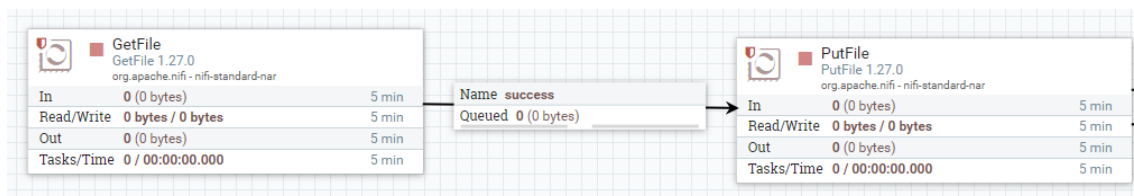
Configure Processor | GetFile 1.27.0

Invalid

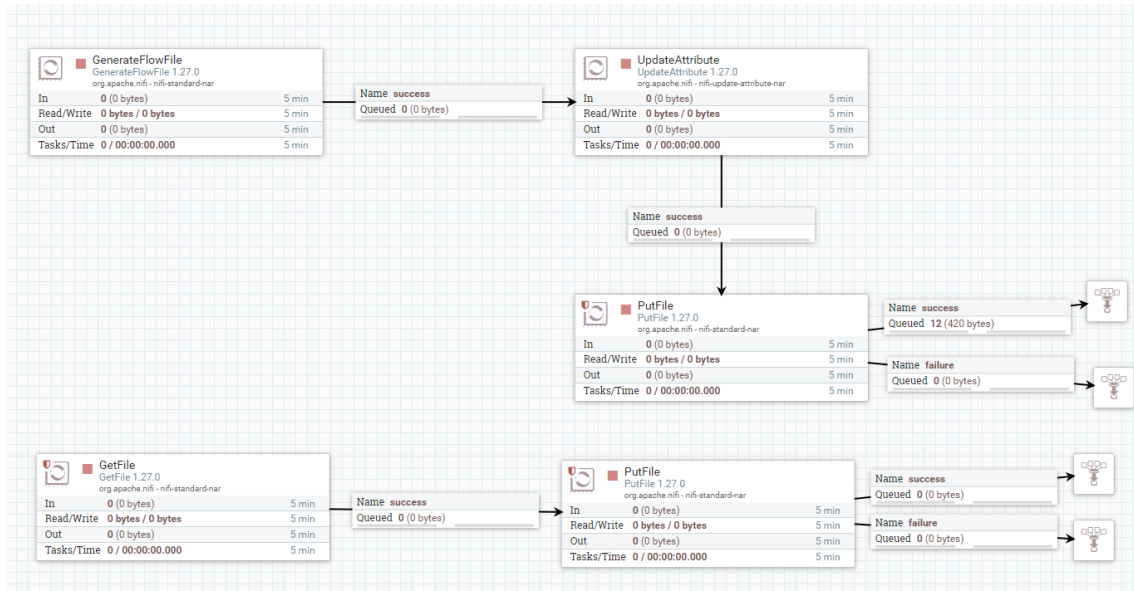
| SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS |
|--|------------------------------------|------------|---------------|----------|
| Required field ✓ + | | | | |
| Property | Value | | | |
| Input Directory | /home/hadoop/workspace/entradaNifi | | | |
| File Filter | [*\].* | | | |
| Path Filter | No value set | | | |
| Batch Size | 10 | | | |
| Keep Source File | false | | | |
| Recurse Subdirectories | true | | | |
| Polling Interval | 0 sec | | | |
| Ignore Hidden Files | true | | | |
| Minimum File Age | 0 sec | | | |
| Maximum File Age | No value set | | | |
| Minimum File Size | 0 B | | | |
| Maximum File Size | No value set | | | |

Required field

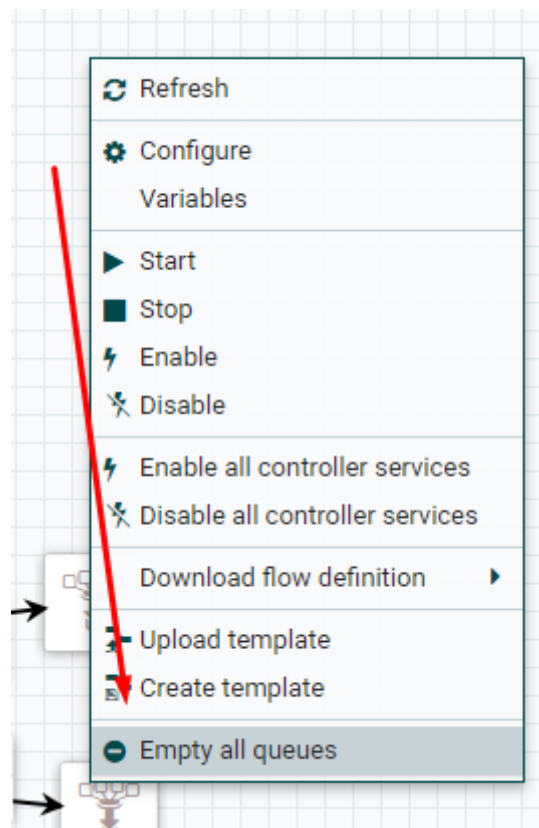
| Property | Value | |
|------------------|--|--|
| Input Directory | /home/hadoop/workspace/entradaNifi | |
| File Filter | Default value: false [*\].* | |
| Path Filter | Expression language scope: Not Supported No value set | |
| Batch Size | Sensitive property: false 10 | |
| Keep Source File | false | |

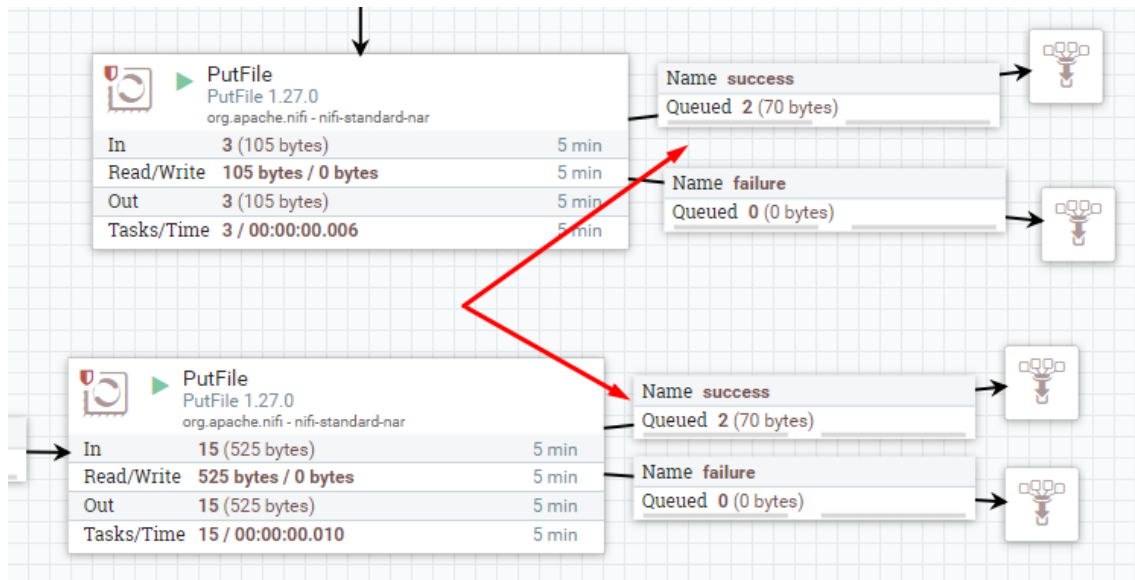


Paso 7. Arrancamos el flujo de NiFi y comprobamos el resultado.



Antes de arrancar limpiamos todas las colas.





```
hadoop@hadoop-virtualbox:~$ ls -lart workspace/entradaNifi/
total 8
drwxrwxr-x 4 hadoop hadoop 4096 ago 29 02:38 ..
drwxrwxr-x 2 hadoop hadoop 4096 ago 29 03:21 .
hadoop@hadoop-virtualbox:~$
hadoop@hadoop-virtualbox:~$ ls -lart workspace/salidaNifi/
total 80
drwxrwxr-x 4 hadoop hadoop 4096 ago 29 02:38 ..
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030743
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030943
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030723
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030843
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030923
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030903
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030823
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829031043
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829031023
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829030803
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829032012
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829031103
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829031003
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829032032
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:20 ficheros_20240829032052
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:21 ficheros_20240829032112
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:21 ficheros_20240829032132
-rw-rw-r-- 1 hadoop hadoop 35 ago 29 03:21 ficheros_20240829032152
drwxrwxr-x 2 hadoop hadoop 4096 ago 29 03:21 .
hadoop@hadoop-virtualbox:~$
```