PEC PROCESAMIENTO DE STREAMS E INGESTA DE DATOS

MIGUEL PÉREZ CARO

EJERCICIO NIFI

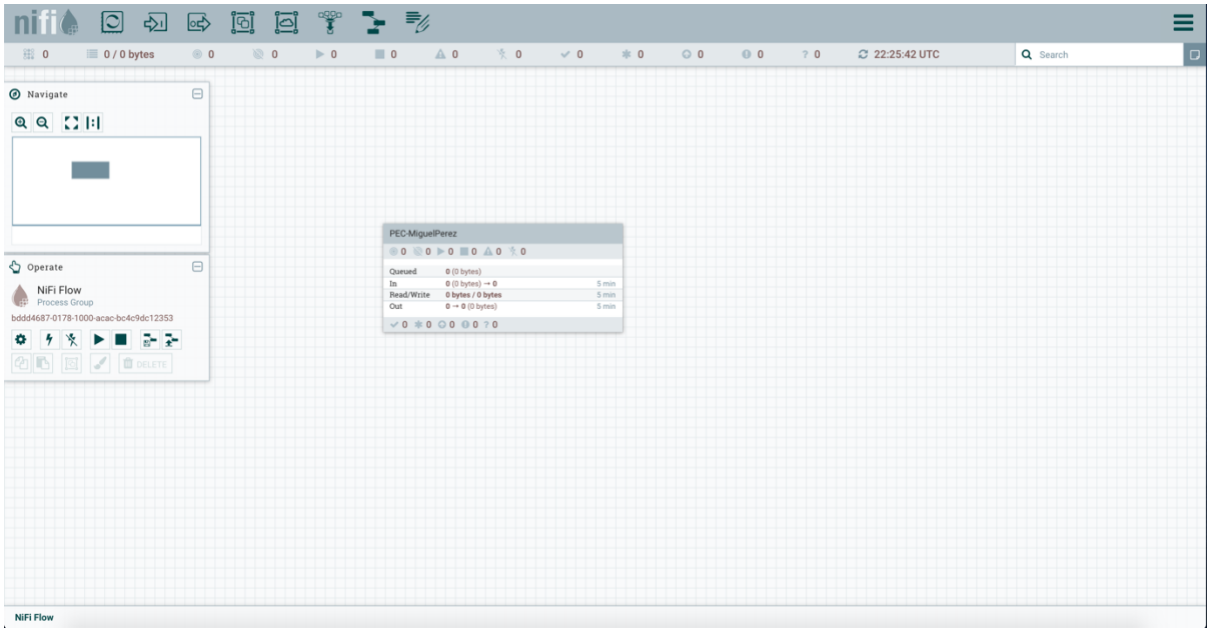
Para la realización de los apartados de este ejercicio se han de usar los archivos que se encuentran en la carpeta FicherosPEC.zip, por lo que se envía dicho fichero zip por ssh a la máquina virtual, se descomprime y se aloja la carpeta tweets en el path: /home/ubuntu/bigdata. La carpeta está compuesta por una multitud de jsons, un archivo desktop.ini, y dos carpetas, english y spanish. A continuación, se muestra la cantidad de archivos por pantalla, que servirá para comprobar posteriormente que los resultados mostrados por los procesos realizados en Nifi son correctos:

```
(base) ubuntu@master-1:~/bigdata/tweets$ ls | wc -l
512
(base) ubuntu@master-1:~/bigdata/tweets$ cd english/
(base) ubuntu@master-1:~/bigdata/tweets/english$ ls | wc -l
94
(base) ubuntu@master-1:~/bigdata/tweets/english$ cd ..
(base) ubuntu@master-1:~/bigdata/tweets$ cd spanish/
(base) ubuntu@master-1:~/bigdata/tweets/spanish$ ls | wc -l
33
(base) ubuntu@master-1:~/bigdata/tweets/spanish$
```

Una vez comentado esta introducción al ejercicio, se comienza con la resolución de los apartados:

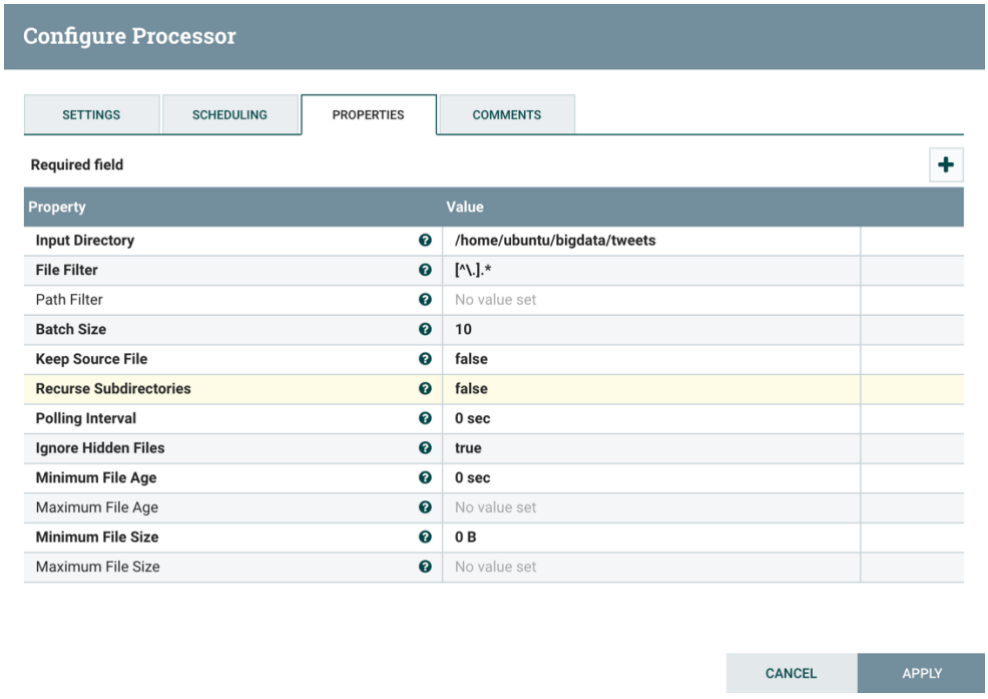
a) Creación de un grupo (Process Group) dentro de NiFi con el nombre PEC-<NombreAlumno>.

Simplemente hay que pinchar el botón para crear el Process Group y asignarle el nombre indicado.

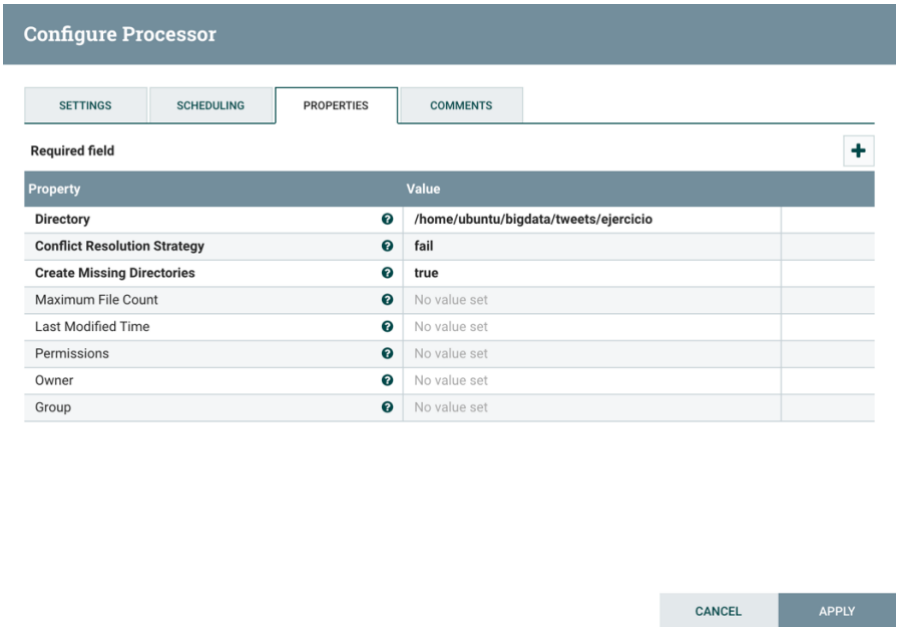


b) Crear un flujo en NiFi que coja solo los ficheros de la carpeta tweets (del zip proporcionado) y no de sus subcarpetas y los deje en el directorio tweets/ejercicio.

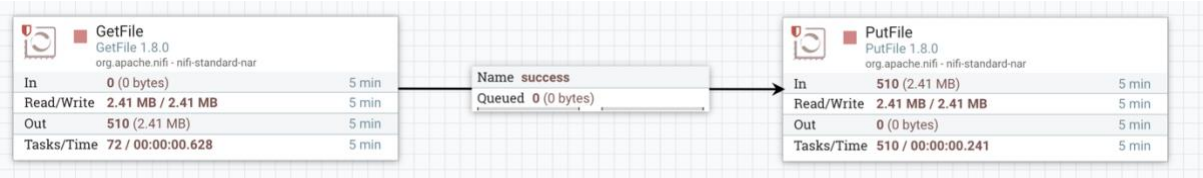
Para la resolución de este ejercicio hay que crear el proceso GetFile y unirlo con el proceso PutFile, de forma que se cogen los archivos de un directorio y se dejan en otro. Para el proceso GetFile, es necesario informar de donde se van a extraer los archivos, que en este caso es el path /home/ubuntu/bigdata/tweets, especificando además el campo Recurse Subdirectories como false para que no recorra los subdirectorios.



Para el proceso PutFile habrá que especificar donde se van a alojar dichos archivos, que será en /home/ubuntu/bigdata/tweets/ejercicio, estableciendo Create Missing Directories a true para que cree el directorio ejercicio en caso de que no exista



Finalmente, el flujo completo una vez finalizado quedaría tal que:

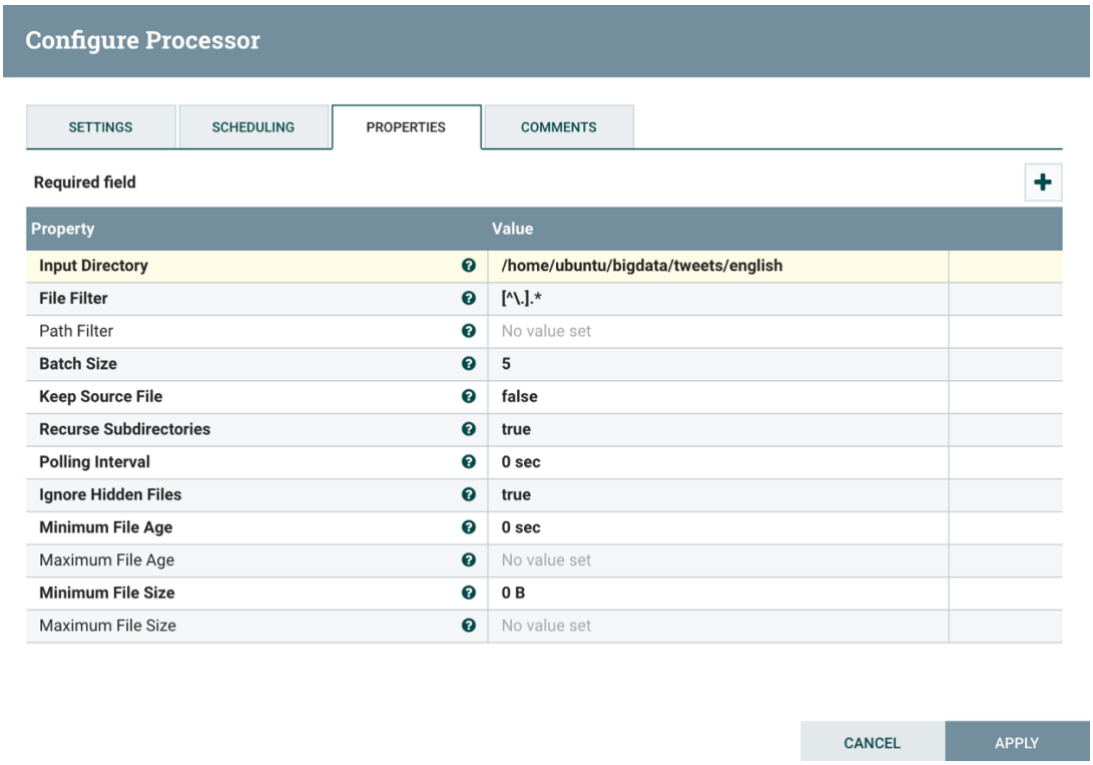


Se observa que se han transferido todos los documentos de la carpeta tweets a excepción de los dos directorios, english y spanish. También se muestra el contenido actual de la carpeta tweets:

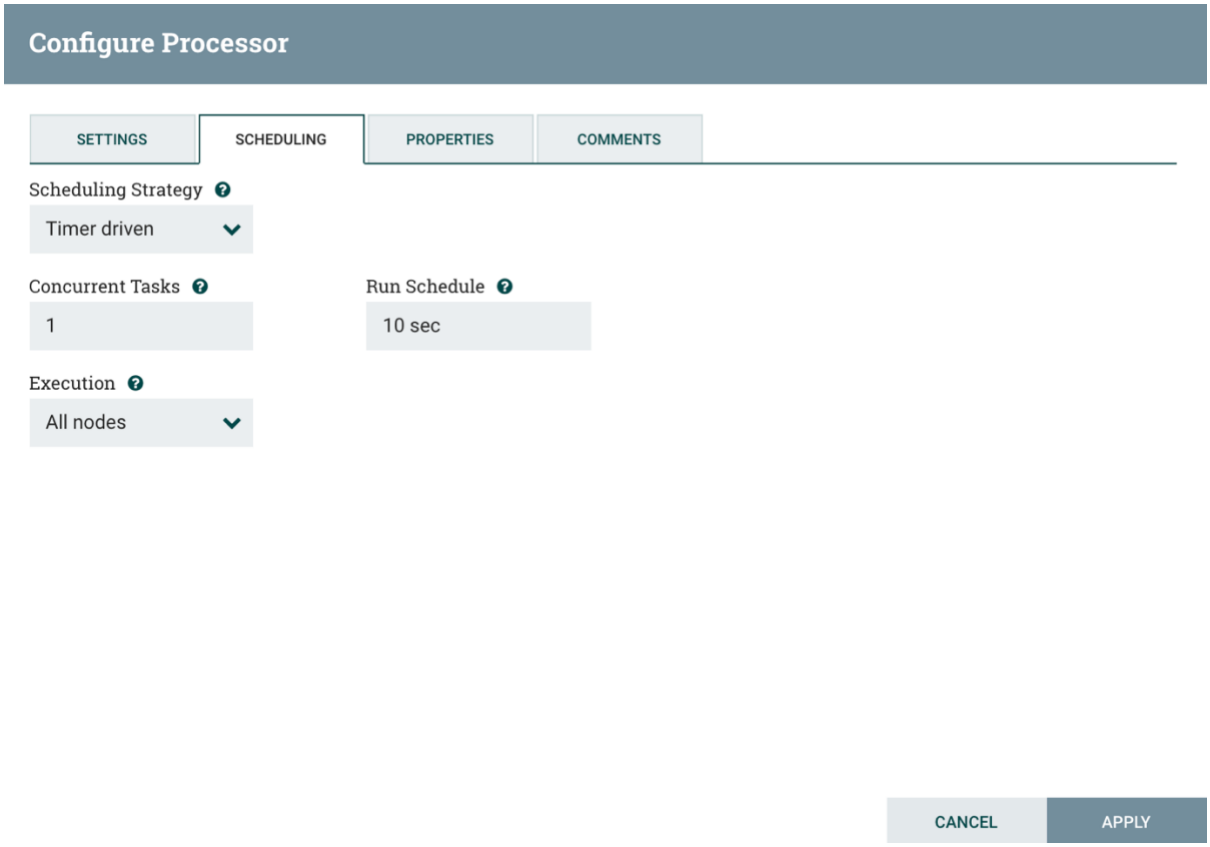
```
(base) ubuntu@master-1:~/bigdata/tweets$ ls
ejercicio  english  spanish
(base) ubuntu@master-1:~/bigdata/tweets$ cd ejercicio
(base) ubuntu@master-1:~/bigdata/tweets/ejercicio$ ls
00a0e5c9-749f-4209-b852-652b52fa4257.json  80112690-27d2-4244-86df-a6e4946cbefe.json
00bf1029-cbbd-4b19-a0bd-7d406a13488b.json  80887cea-eebb-45f0-90e4-c4247c8e4906.json
01b1a4bc-3595-4310-a170-1c63008db1ff.json  812479f8-4858-4a26-9ba4-0ef543eb767b.json
025e66c5-4824-4d81-accf-a3442766b39d.json  8151e570-ac6e-40f5-98bf-6503d5327243.json
029c7146-c227-4e07-8357-16262b052f5f.json  81c68a93-140e-4452-9612-a8a8814c3aef.json
02d3dd82-bae8-44c0-a35d-e85e5e587578.json  820aec44-66ac-48d4-8d97-2b9069049cc4.json
03ad5a01-ccdb-4c96-a46a-83328112d3c4.json  82f7233c-c270-46be-811a-506613cb6861.json
040c19ea-0fac-406f-a4fd-4ac6f913ab4a.json  83081e12-896d-4256-89e3-94551a243ffe.json
```

c) Crear un nuevo flujo que coja los tweets de 5 en 5 de la carpeta tweets/english cada 10 segundos y los deje de nuevo en el directorio tweets/ejercicio:

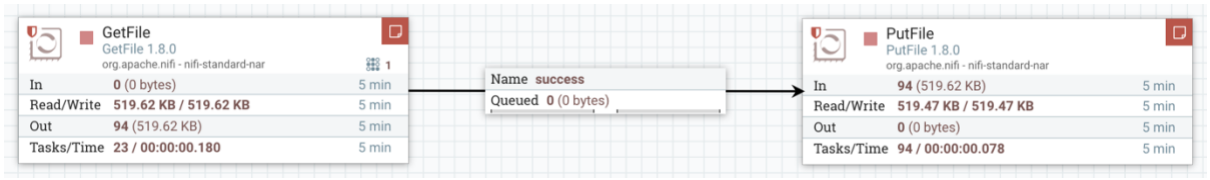
Es un proceso muy similar anterior con la diferencia de que hay que modificar el directorio del que se extraen los tweets, en este caso es la carpeta tweets/english, que el batch file es 5, para que se cojan los tweets de 5 en 5 y que hay que marcar un scheduling de 10 segundos. Para ello se hacen las modificaciones pertinentes que se demuestran con las siguientes capturas de pantalla:



En la imagen anterior se observa como el directorio se ha modificado, así como el batch size. En este caso el valor del campo Rescurse Subdirectories se ha dejado en su valor por defecto, que es true, ya que no afecta al resultado final. En la siguiente imagen se muestra como se ha asignado un scheduling de 10 segundos:



Finalmente, se muestra el flujo final:



Se observa que se han transferido los 94 documentos que contenía la carpeta english.