Fuentes de Datos y Aprovisionamiento

Práctica 2: Procesos ETL. Apache Nifi

Daniel Pérez Efremova

ACLARACIONES

En esta práctica los dos flujos se han ejecutado en un mismo Dataflow de Apache Nifi porque el primer flujo propuesto es una subtarea del primero. Es decir, en el primero se pide escribir distintos ficheros en un mismo directorio y despues escribir los ficheros en distintos directorios segun su origen.

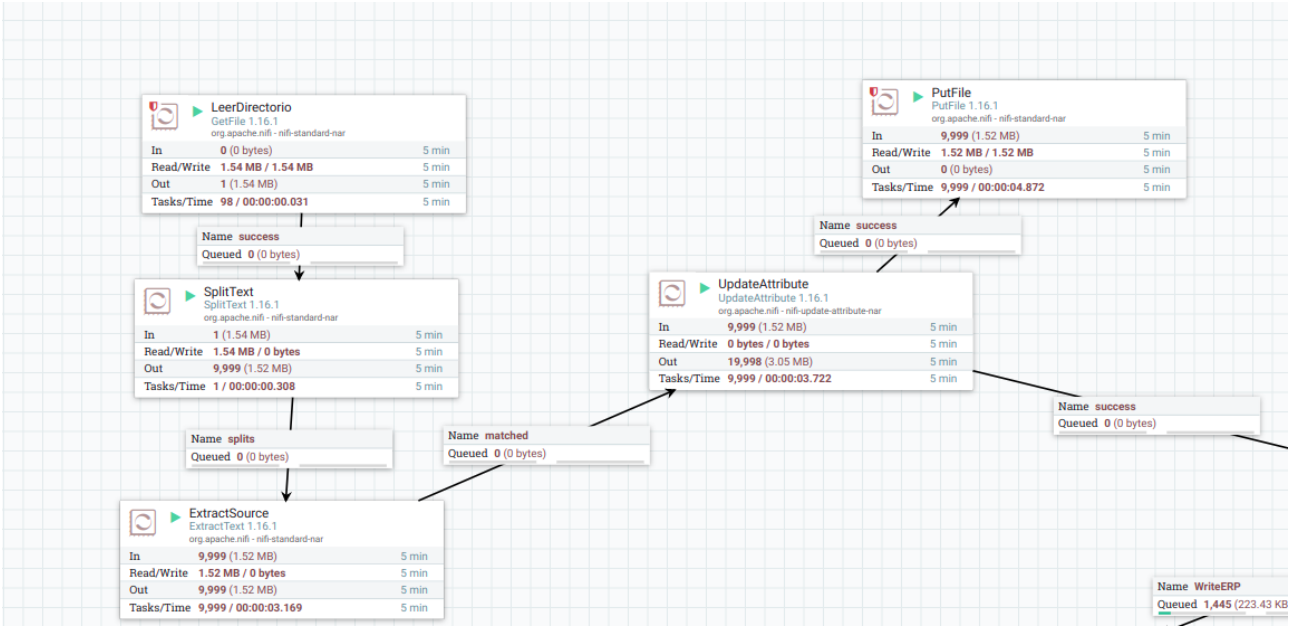
Por tanto, la única diferencia entre el primero y el segundo es la cantidad de Procesadores PutFile y el uso de una propiedad para redireccionar los ficheros.

Con la memoria se adjunta el fichero FLUJO2.xml del flujo de trabajo y el contenido de los directorios de salida (ejecutando **pwd && ls $(pwd) -R >> dirContents.txt** sobre el directorio usado para las pruebas).

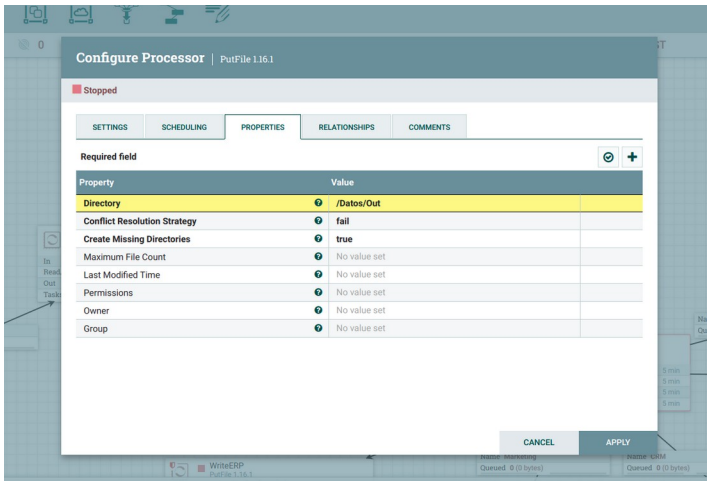
FLUJO 1.

En este flujo de trabajo se leen todos los ficheros de un directorio en “LeerDirectorio”. Después se fragmenta el fichero por cada línea en “SplitText”. Seguidamente con una regla Regex se extrae el orígen de la información de la primera columna de cada registro en “ExtractSource”. El siguiente paso es añadir el atributo “source” al fichero según el valor del paso anterior. Finalmente se escriben los ficheros en el nuevo directorio en “PutFile”.

Se añade captura de pantalla del flujo.



En la página siguiente se añaden capturas de pantalla de la configuración de cada procesador.



Configure Processor | GetFile 1.16.1

Stopped

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Input Directory	**Datos/inputETL**
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

CANCEL APPLY

Configure Processor | SplitText 1.16.1

Stopped

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Line Split Count	**1**
Maximum Fragment Size	No value set
Header Line Count	**0**
Header Line Marker Characters	No value set
Remove Trailing Newlines	**true**

CANCEL APPLY

Configure Processor | UpdateAttribute 1.16.1

Stopped

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	${filename}-${uuid}-${source}

ADVANCED CANCEL APPLY

Configure Processor | ExtractText 1.16.1

Stopped

SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Required field

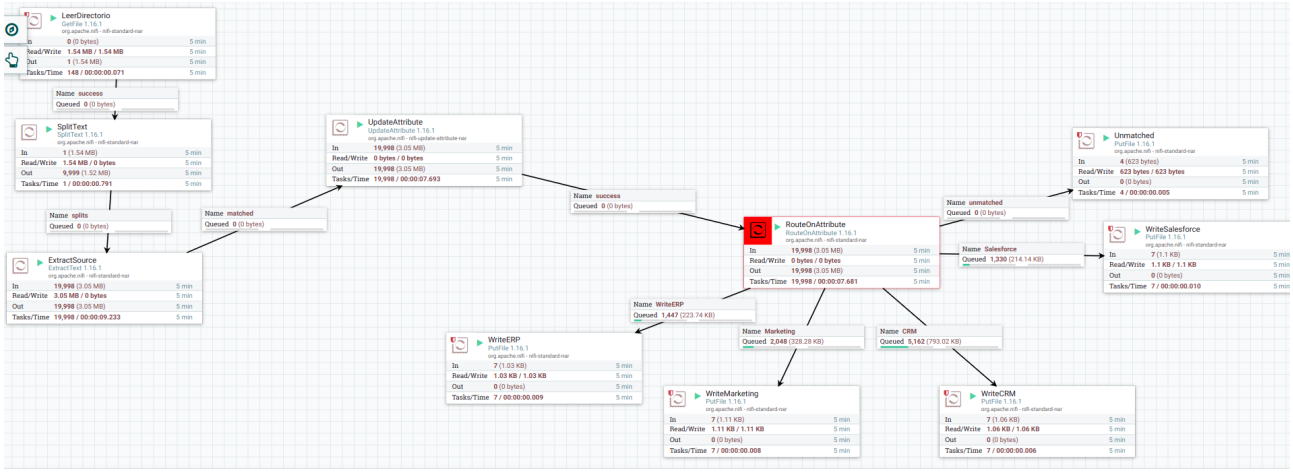
Property	Value
Enable Case-insensitive Matching	**false**
Permit Whitespace and Comments in Pattern	**false**
Enable DOTALL Mode	**false**
Enable Literal Parsing of the Pattern	**false**
Enable Multiline Mode	**false**
Enable Unicode-aware Case Folding	**false**
Enable Unicode Predefined Character Classes	**false**
Enable Unix Lines Mode	**false**
Include Capture Group 0	**true**
Enable repeating capture group	**false**
Enable named group support	false
source	(.*?),

CANCEL APPLY

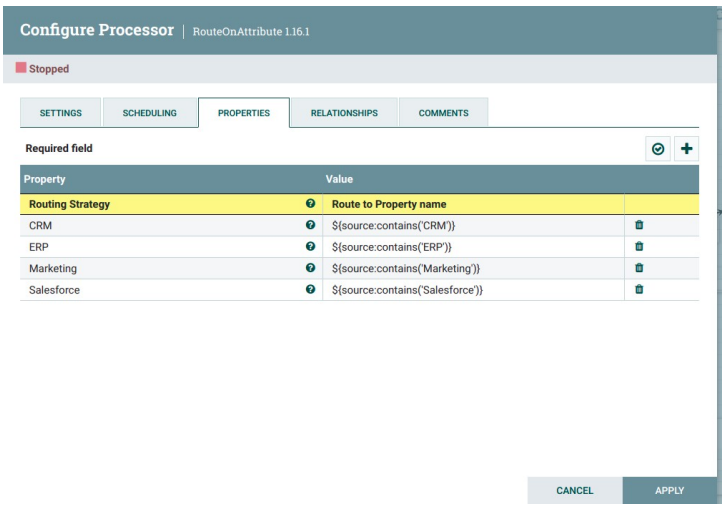
FLUJO 2

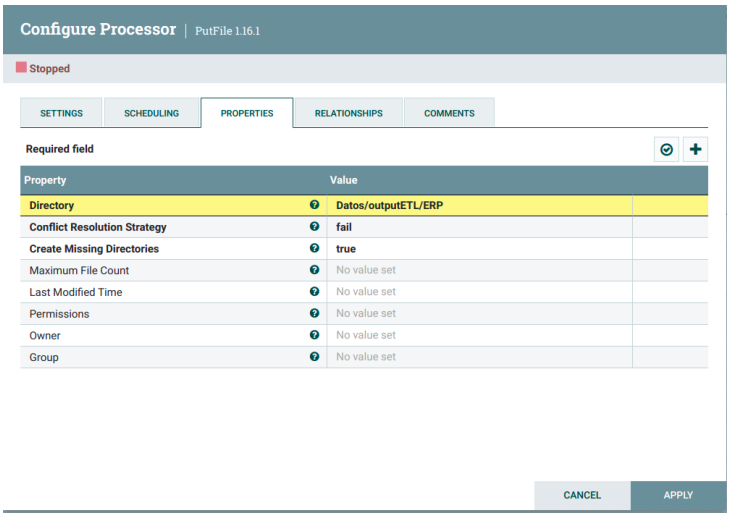
Este flujo es idéntico al anterior, salvo que se añaden nuevos procesadores para direccionar los ficheros a un directorio distinto dependiendo del valor de la primera columna del registro. Para esto, a cada fichero se le asocia la propiedad “source” que puede tomar los valores CRM, Salesforce, ERP o Marketing (segun el valor que tome el primer registro). El nodo “RouteOnAttribute” se encarga de comprobar el valor de esta propiedad y direccionar el fichero al nodo “PutFile” adecuado.

Se añade un nodo Unmatched para volcar todos los ficheros cuyo orígen no esté contemplado por el desarrollador, es decir, la propiedad source no es ninguna de las anteriores.



Se añaden capturas de los nuevos nodos, solo se adjunta un PutFile.





POR FALTA DE TIEMPO NO SE REALIZA LA PARTE OPCIONAL