
Tarea 2.1 Pandas

Sistemas de Big Data
23/07/16 - I.E.S Fernando Wirtz
Alejandro Regueiro Ruiz

Fecha	Motivo del cambio
	Versión inicial

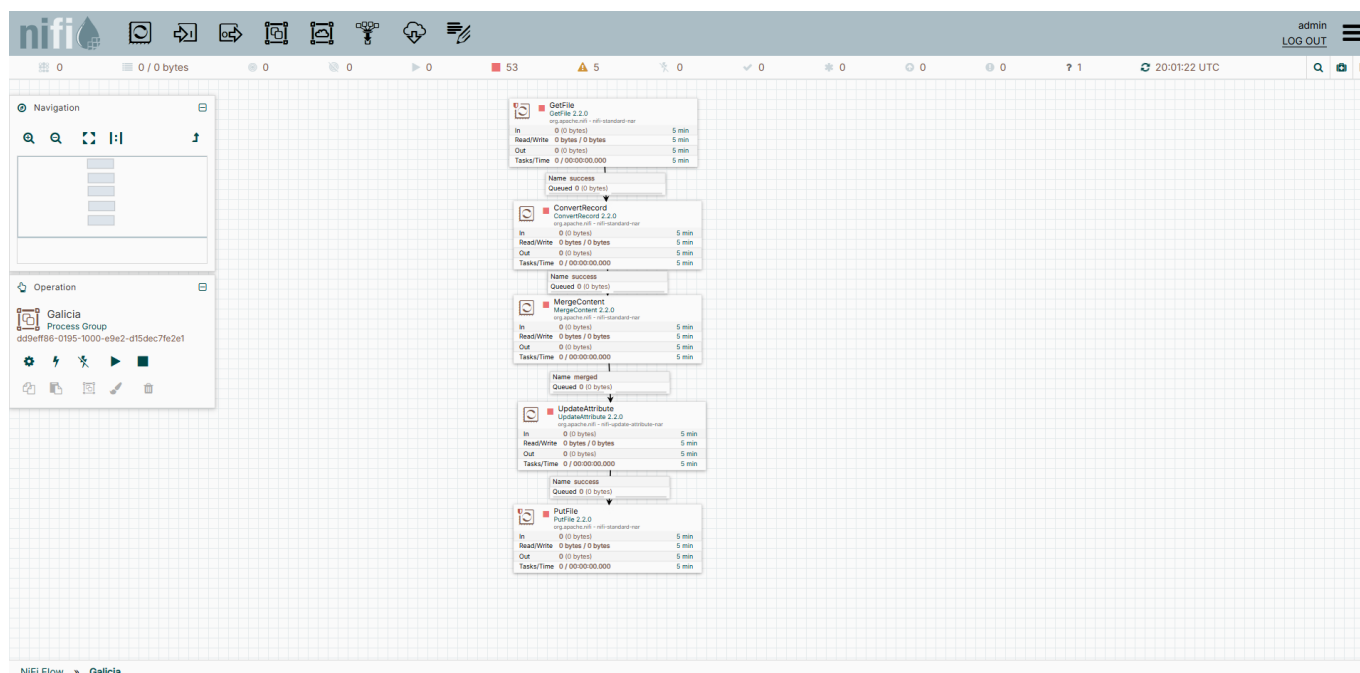
Índice

Título 1.....3

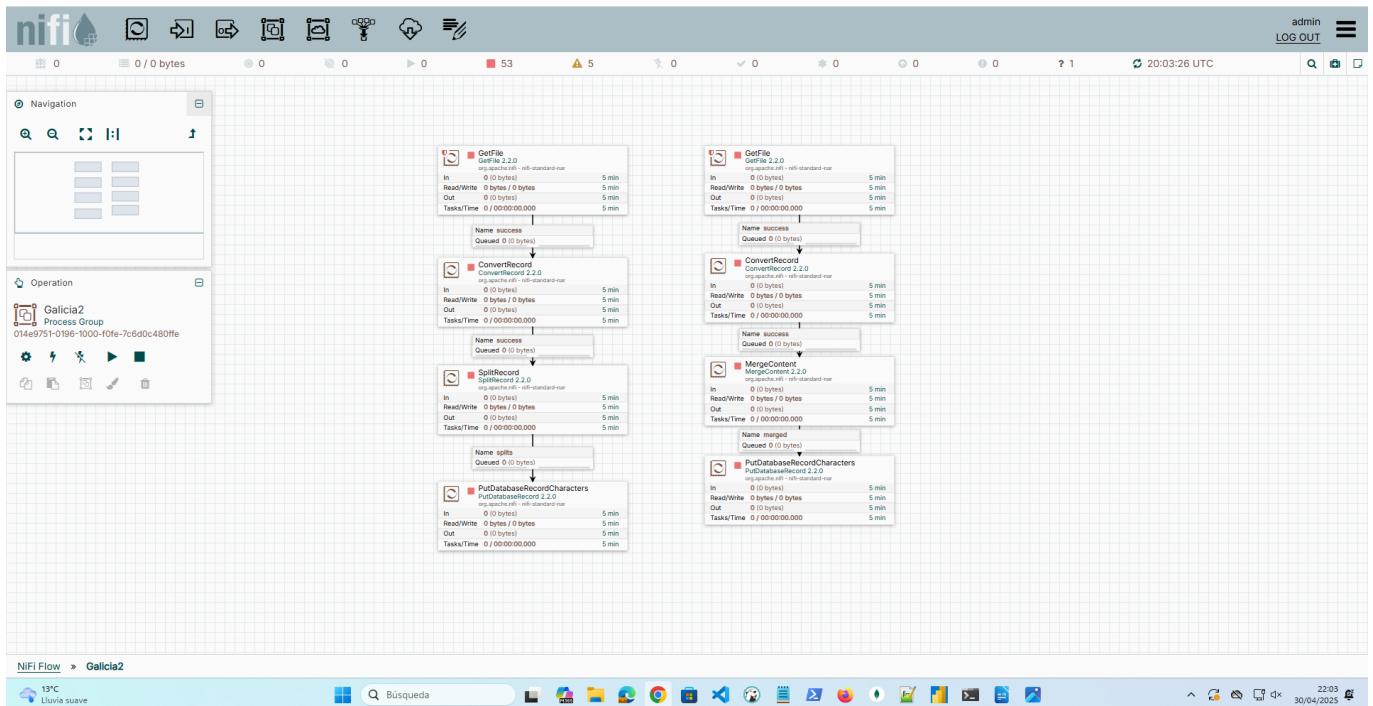
 Título 2.....3

 Titulo 3.....3

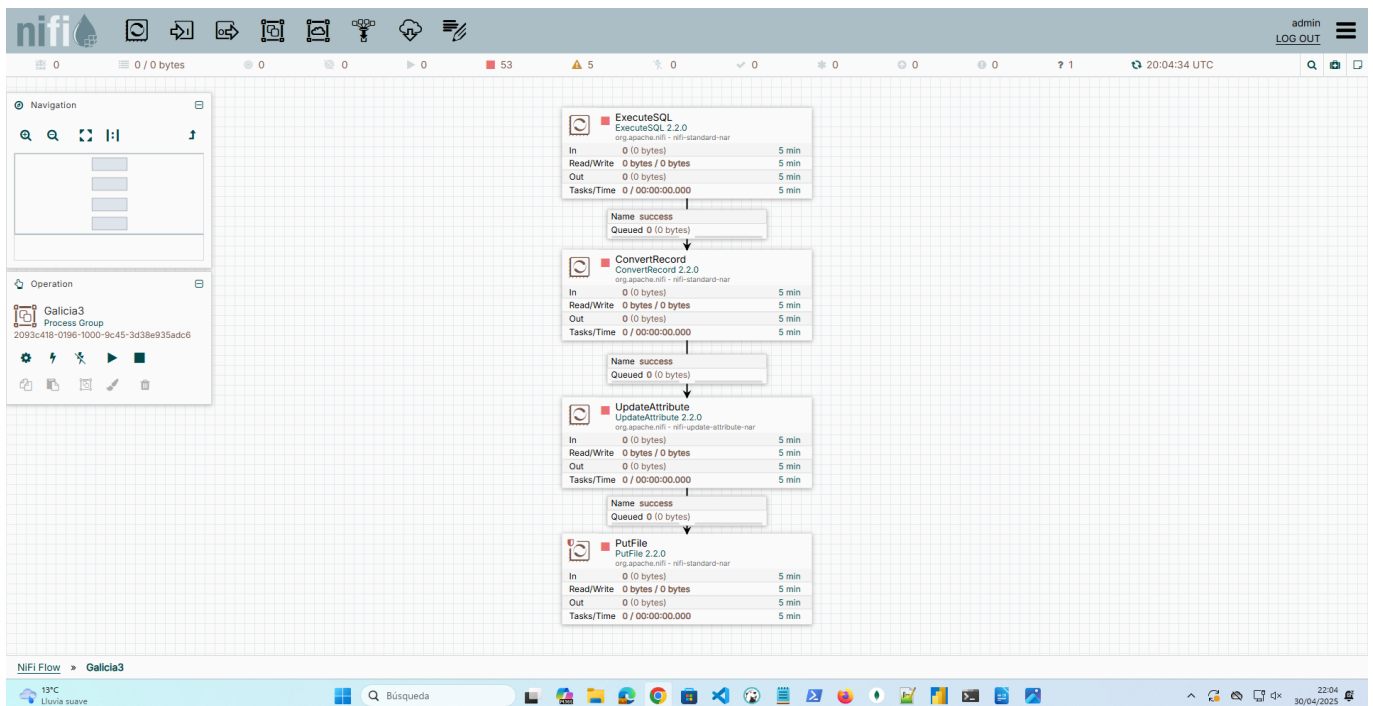


NIFI

Primer grupo de procesamiento donde recibimos los datos de las provincias(no los de meteogalicia) y los juntamos todos en un único csv



Segundo grupo donde recogemos el csv juntado previamente y también el csv descargado con el fax, portal_web, etc. de los concellos e importamos los csv a la base de datos para hacer posteriormente un join



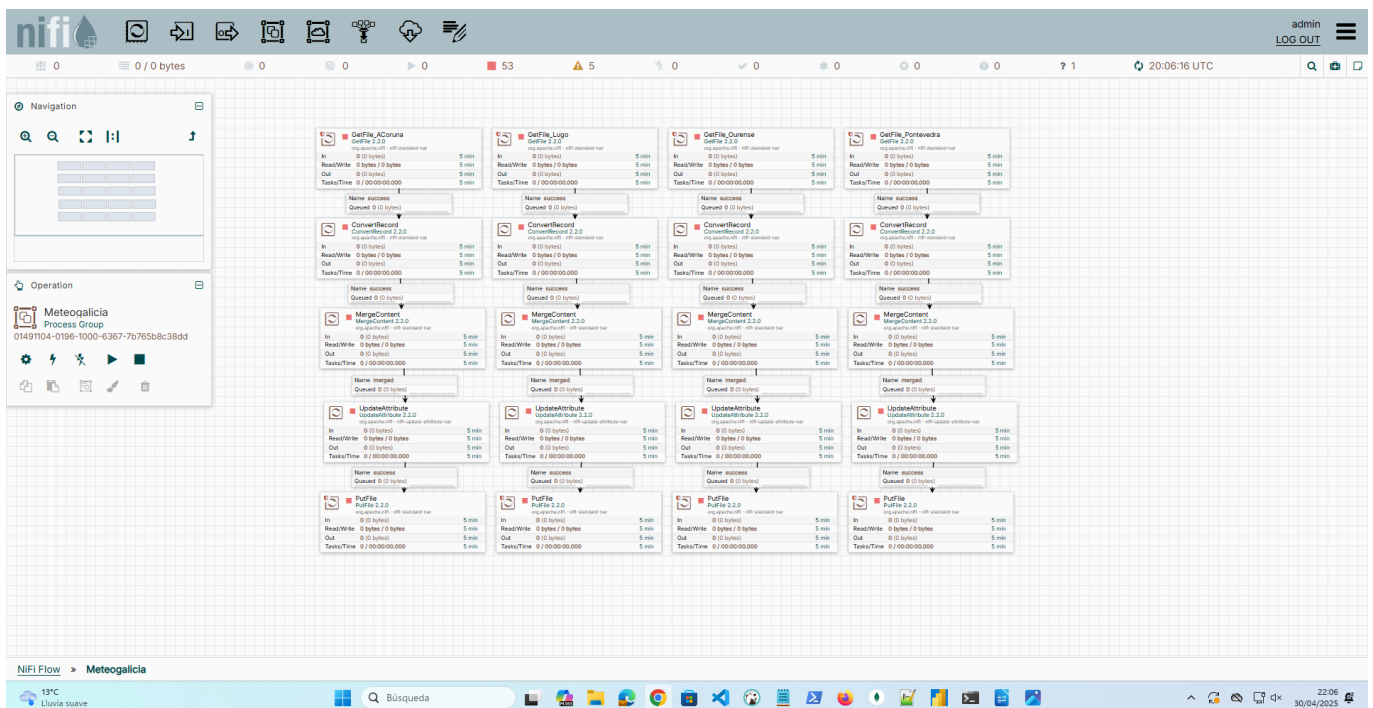
Tercer grupo y último de la primera parte donde ejecutamos un script sql donde hacemos el join de las tablas para y conseguir el csv con los datos mergeados.

SCRIPT SQL

SELECT

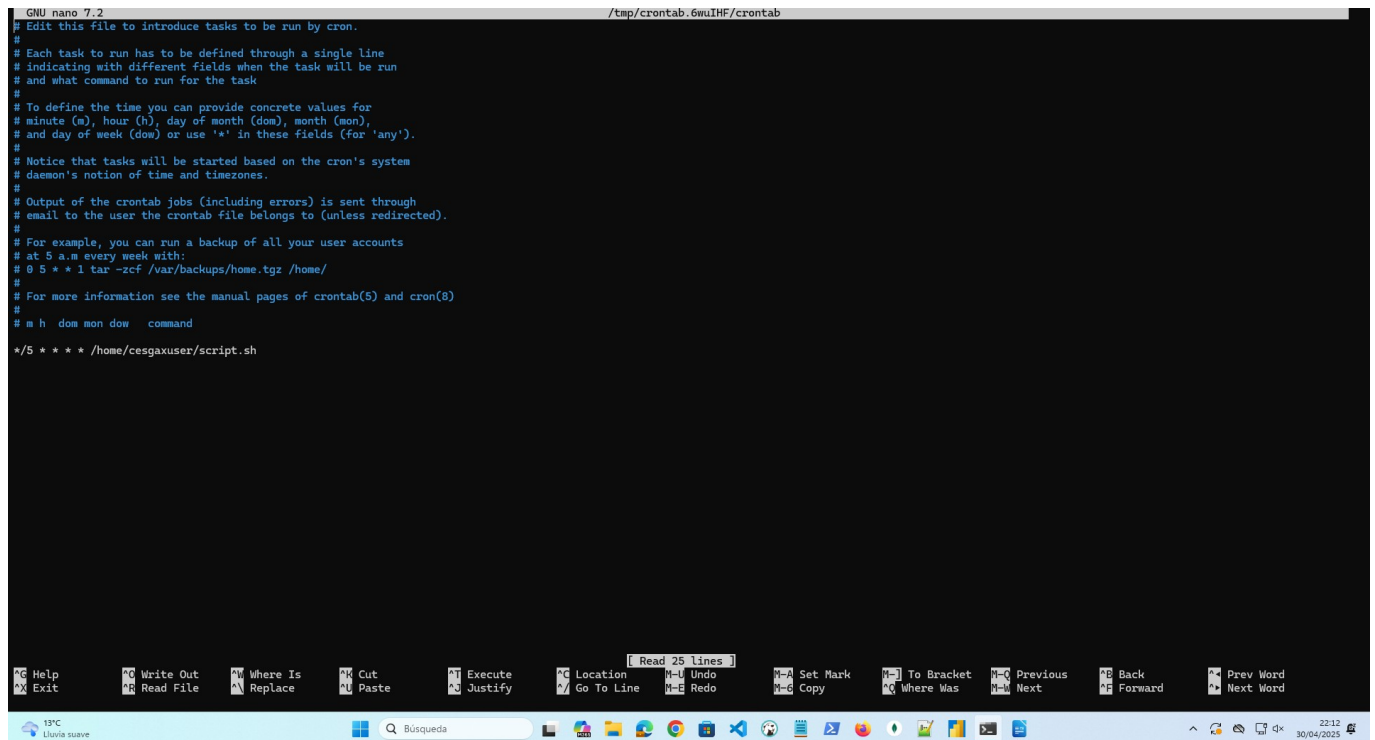
t1.Codigo,
t1.Nome,
t1.Enderezo,
t1.Concello,
t1.Provincia,
t1.Cod_postal,
t1.Telefono,
t1.Tipo_de_centro,
t1.COORDENADA_X,
t1.COORDENADA_Y,

t1.TITULARIDADE,
 t1.ENSINO_CONCERTADO,
 t1.DEPENDENTE,
 t2.telefono,
 t2.fax,
 t2.correo_electronico,
 t2.portal_web
 FROM
 todos t1
 JOIN
 ayuntas t2 ON t1.Concello = t2.concello



Único grupo de procesamiento donde juntamos todos los csv de cada provincia en uno solo, aquí hay el se aumento el número mínimo de ficheros recogidos por el getFile de 10 a 100 y se estableció en los mergeContent un mínimo de ficheros igual a la cantidad de csv que tiene cada directorio de cada provincia para asegurarse de que se junten todos los csv en 1

Después instalamos CRON en nuestra máquina del cesga y añadimos que cada 5 minutos se ejecutara “script.sh” enviado en la tarea, también para el correcto funcionamiento del script se creó una clave ssh que se asoció al Hadoop para que no se necesite introducir la contraseña de acceso del usuario al hadoop.



```
GNU nano 7.2 /tmp/crontab.6wuIHF/crontab
# Edit this file to introduce tasks to be run by cron.
#
# Each task to run has to be defined through a single line
# indicating with different fields when the task will be run
# and what command to run for the task
#
# To define the time you can provide concrete values for
# minute (m), hour (h), day of month (dom), month (mon),
# and day of week (dow) or use '*' in these fields (for 'any').
#
# Notice that tasks will be started based on the cron's system
# daemon's notion of time and timezones.
#
# Output of the crontab jobs (including errors) is sent through
# email to the user the crontab file belongs to (unless redirected).
#
# For example, you can run a backup of all your user accounts
# at 5 a.m every week with:
# 0 5 * * 1 tar -zcf /var/backups/home.tgz /home/
#
# For more information see the manual pages of crontab(5) and cron(8)
#
# m h dom mon dow   command
*/5 * * * * /home/cesgaxuser/script.sh
```

La foto de la clave ssh no la muestro por seguridad :)