

## Práctica 3: Captura de datos con Flume

Para capturar datos con **Flume** es necesario configurar un agente que recibe datos desde una fuente (**source**), que volcará en un sumidero (**sink**).

### Configuración del agente Flume

Hay que crear un archivo que contiene las configuraciones necesarias de:

- Los componentes del **agente**: indican los nombres del source, sink y channel.
- La configuración del **source**: en este caso se abre una conexión con la aplicación netcat en el puerto 55555
- La configuración del **channel**: tipo (por memoria) y capacidades del canal para los eventos
- La configuración del **sink**: el destino son archivos almacenados en hdfs

Las líneas del archivo **agenteFlume.conf** son, en este caso

```
#declaracion de componentes
agente.sources = sr1
agente.channels = chn1
agente.sinks = snk1

#configuracion del source
agente.sources.sr1.type = netcat
agente.sources.sr1.bind = localhost
agente.sources.sr1.port = 55555
agente.sources.sr1.channels = chn1

#configuracion del channel
agente.channels.chn1.type = memory
#La cantidad maxima de eventos almacenados en el canal
agente.channels.chn1.capacity = 1000
#La cantidad maxima de eventos que el canal capturara de la fuente por transacción
agente.channels.chn1.transactionCapacity = 100

#definimos la configuracion del Sink
agente.sinks.snk1.type = hdfs
agente.sinks.snk1.hdfs.path = hdfs://node1:8020/user/alumno/flume-puerto
#DataStream no comprime el archivo de salida
agente.sinks.snk1.hdfs.fileType = DataStream
agente.sinks.snk1.channel = chn1
```

### Arranque del proceso de Flume

Ejecutar el siguiente comando para lanzar el proceso Flume para ejecutar el archivo creado. Se usa la opción *-conf-file* para indicar el archivo de configuración, *-conf* para la configuración de Flume, y *-name* para el nombre del agente (debe coincidir con el del fichero de configuración agenteFlume.conf)

```
sudo flume-ng agent --conf /etc/flume-ng/conf --conf-file agenteFlume.conf \
-name agente -Dflume.root.logger=INFO,console
```

Debe aparecer en la última línea este mensaje que indica que Flume se ha conectado al puerto

```
....Created serverSocket:sun.nio.ch.ServerSocketChannelImpl[/127.0.0.1:55555]
```

Y el terminal donde lanzamos el proceso Flume, debería aparecer en la última línea un mensaje parecido a este:

```
Creating hdfs://node1:8020/user/alumno/flume-puerto/FlumeData.163XXXXX444
```

### Arranque de la fuente de datos

Como en este caso el origen del canal de Flume es la aplicación **netcat**, lanzamos este comando para enviar datos al agente Flume y escribimos alguna frase de ejemplo:

```
### [alumno@pasarela ~]$ nc localhost 55555
Muestra de la escritura de este ordenador
OK
```

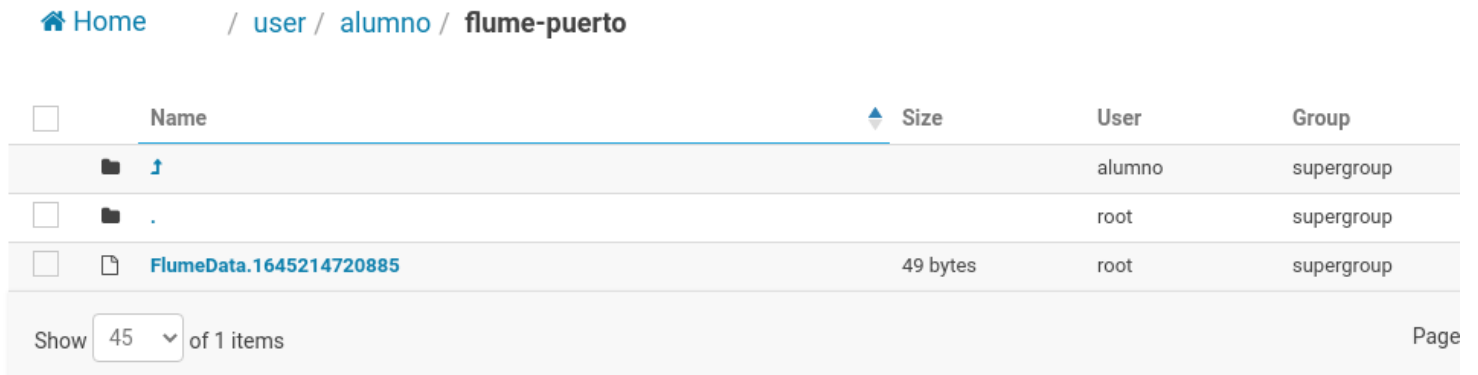
Para comprobar que el puerto está escuchando podemos usar el comando lsof

```
lsof -i -P -n
```

## Comprobación de los datos escritos en HDFS

Se puede verificar que se están escribiendo los datos en HDFS en la ruta configurada como salida `/user/alumno/flume-puerto/`

En HUE se puede mirar con el menú Files



<input type="checkbox"/>	Name	Size	User	Group
<input type="checkbox"/>	↑		alumno	supergroup
<input type="checkbox"/>	.		root	supergroup
<input type="checkbox"/>	FlumeData.1645214720885	49 bytes	root	supergroup

Show 45 of 1 items Page

Figure 1: Archivos creados como sink

O bien usando la orden por terminal hdfs adecuada

```
hdfs dfs -ls /user/alumno/flume-puerto
```

**Recoge en un pantallazo os datos escritos con Flume en HDFS** como muestra de que has hecho la práctica. Recuerda que se debe ver tu nombre en la imagen.

## Parada de los servicios

Para terminar de forma adecuada el agente Flume

- Pulsamos Ctrl+D en el terminal con netcat
- Pulsamos Ctrl+C en el agente Flume