

## Flume. Práctica 2: Encadenamiento de flujos de agentes

Si no se tiene instalado, previamente se instala el agente de Flume en Cloudera en el host node1.

Este agente tiene dependencias con Kafka y HDFS que son los servicios con los que vamos a interactuar en esta y posteriores prácticas.

En esta práctica se encadenan dos agentes de forma que el sumidero del primero es la fuente del segundo. Para unir ambos agentes se utiliza como nexo el tipo avro como sumidero-fuente de ambos agentes.

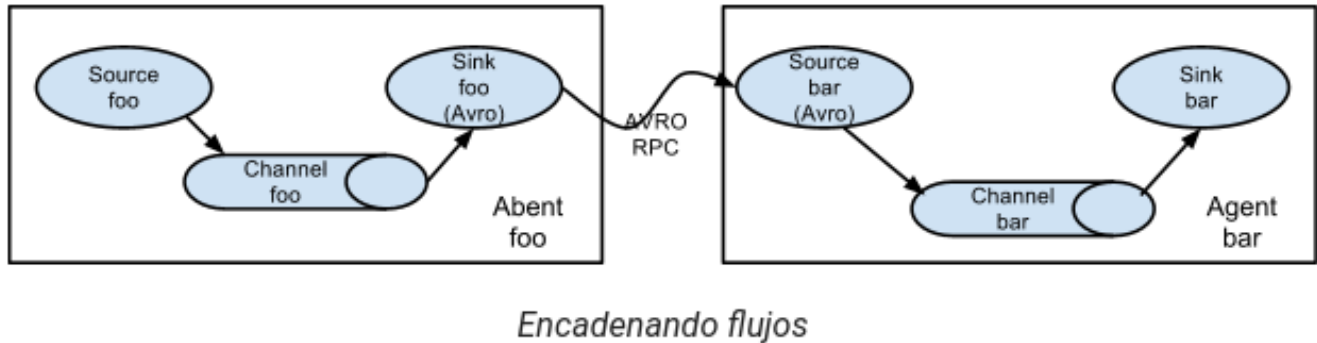


Figure 1: Esquema de funcionamiento de la práctica

Los dos agentes se configuran y arrancan en el mismo host, en este caso *pasarela*

### Primer agente: de netcat a Avro

Configuramos el primer agente en el fichero **netcat-avro.conf** con el código siguiente

```
# Nombramos a los componentes del agente
NetcatAvroAgent.sources = Netcat
NetcatAvroAgent.channels = FileChannel
NetcatAvroAgent.sinks = AvroSink

# Describimos el origen netcat en localhost:44444
NetcatAvroAgent.sources.Netcat.type = netcat
NetcatAvroAgent.sources.Netcat.bind = localhost
NetcatAvroAgent.sources.Netcat.port = 44444

# Describimos el destino como Avro en localhost:10003
NetcatAvroAgent.sinks.AvroSink.type = avro
NetcatAvroAgent.sinks.AvroSink.hostname = localhost
NetcatAvroAgent.sinks.AvroSink.port = 10003

# Unimos el origen y el destino a través del canal de fichero
NetcatAvroAgent.sources.Netcat.channels = FileChannel
NetcatAvroAgent.sinks.AvroSink.channel = FileChannel
NetcatAvroAgent.channels.FileChannel.type = file
NetcatAvroAgent.channels.FileChannel.dataDir = /home/alumno/practicas_curso/flume/data
NetcatAvroAgent.channels.FileChannel.checkpointDir = /home/alumno/practicas_curso/flume/checkpoint
```

A destacar como parámetros de configuración,

- tipo del sumidero: **avro**
- Puerto de escucha de source: 44444
- Puerto de entrega del sink: 10003
- **Canal utilizado:** **file**, que guarda en disco los mensajes que se transmiten. Necesita de un directorio del sistema de archivos (`/home/alumno/practicas_curso/flume/data`) y de un directorio para puntos de control `checkpointDir` (`/home/alumno/practicas_curso/flume/checkpoint`)

## Segundo agente: de Avro a HDFS

En este segundo agente, la fuente es el sumidero del NetcatAvroAgent en el puerto 10003, tipo Avro, y el sumidero es HDFS.

El fichero de configuración **avro-hdfs.conf** es el siguiente:

```
# Nombramos a los componentes del agente
AvroHdfsAgent.sources = AvroSource
AvroHdfsAgent.channels = MemChannel
AvroHdfsAgent.sinks = HdfsSink

# Describimos el origen como Avro en localhost:10003
AvroHdfsAgent.sources.AvroSource.type = avro
AvroHdfsAgent.sources.AvroSource.bind = localhost
AvroHdfsAgent.sources.AvroSource.port = 10003

# Describimos el destino HDFS
AvroHdfsAgent.sinks.HdfsSink.type = hdfs
AvroHdfsAgent.sinks.HdfsSink.hdfs.path = /user/alumno/flume/avro_data/
AvroHdfsAgent.sinks.HdfsSink.hdfs.fileType = DataStream
AvroHdfsAgent.sinks.HdfsSink.hdfs.writeFormat = Text

# Unimos el origen y el destino
AvroHdfsAgent.sources.AvroSource.channels = MemChannel
AvroHdfsAgent.sinks.HdfsSink.channel = MemChannel
AvroHdfsAgent.channels.MemChannel.type = memory
```

El directorio donde se van a guardar los mensajes en HDFS es */user/alumno/flume/avro\_data/*

## Lanzamiento del segundo agente

Se arranca en primer lugar el segundo agente, para que esté a la escucha de los mensajes para cuando se arranque el primer agente.

La orden que lo arranca es

```
flume-ng agent --conf /etc/flume-ng/conf --conf-file avro-hdfs.conf --name AvroHdfsAgent \
-Dflume.root.logger=INFO,console
```

Aparece al final como esperando a la escucha

```
[INFO - org.apache.flume.source.AvroSource.start(AvroSource.java:223)] Avro source AvroSource started.
```

Seguidamente, en otro terminal, se arranca el segundo agente

```
flume-ng agent --conf /etc/flume-ng/conf --conf-file netcat-avro.conf --name NetcatAvroAgent \
-Dflume.root.logger=INFO,console
```

```
(lifecycleSupervisor-1-0) [INFO - org.apache.flume.source.NetcatSource.start(NetcatSource.java:166)] Created serverSocket:sun.nio.c
h[127.0.0.1:44444]
(lifecycleSupervisor-1-1) [INFO - org.apache.flume.sink.AbstractRpcSink.start(AbstractRpcSink.java:311)] Rpc sink AvroSink started.
```

Figure 2: Agente NetcatAvroAgent a la escucha

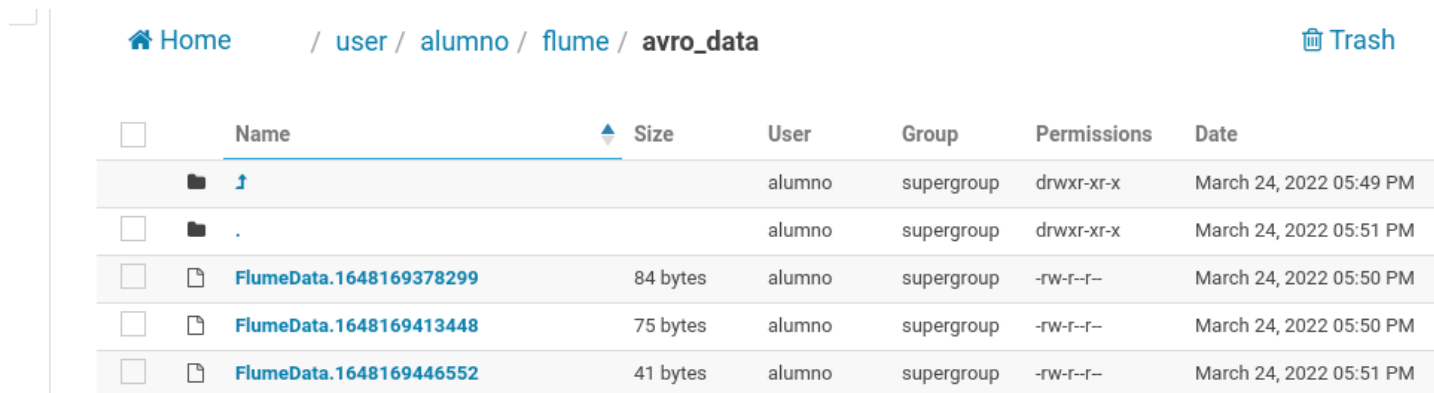
## Creación de mensajes de prueba

En un tercer terminal probamos el funcionamiento creando mensajes con la orden

```
nc localhost 44444
```

## Comprobación de los resultados

En primer lugar, accedemos a HDFS, en el directorio `/user/alumno/flume/avro_data/` donde deben estar creados los ficheros con los mensajes.










	Home	/	user	/	alumno	/	flume	/	avro_data		Trash
<input type="checkbox"/>	Name		Size		User		Group		Permissions		Date
<input type="checkbox"/>		f			alumno		supergroup		drwxr-xr-x		March 24, 2022 05:49 PM
<input type="checkbox"/>		.			alumno		supergroup		drwxr-xr-x		March 24, 2022 05:51 PM
<input type="checkbox"/>		FlumeData.1648169378299	84 bytes		alumno		supergroup		-rw-r--r--		March 24, 2022 05:50 PM
<input type="checkbox"/>		FlumeData.1648169413448	75 bytes		alumno		supergroup		-rw-r--r--		March 24, 2022 05:50 PM
<input type="checkbox"/>		FlumeData.1648169446552	41 bytes		alumno		supergroup		-rw-r--r--		March 24, 2022 05:51 PM

Figure 3: Mensajes recibidos en HDFS

/ user / alumno / flume / avro\_data / **FlumeData.1648169378299**

```
primera linea de prueba
de dos agentes encadenados
primer agente con source netcat
```

Es interesante acceder al sistema de archivos de Linux, en el directorio `/home/alumno/practicas_curso/flume/checkpoint` que se ha creado como control del flujo de mensajes.

```
[alumno@pasarela checkpoint]$ pwd
/home/alumno/practicas_curso/flume/checkpoint
[alumno@pasarela checkpoint]$ ls -lh
total 7,7M
-rw-rw-r--. 1 alumno alumno 7,7M mar 25 01:50 checkpoint
-rw-rw-r--. 1 alumno alumno 25 mar 25 01:50 checkpoint.meta
-rw-rw-r--. 1 alumno alumno 32 mar 25 01:50 inflightputs
-rw-rw-r--. 1 alumno alumno 32 mar 25 01:50 inflighttakes
-rw-rw-r--. 1 alumno alumno 0 mar 25 01:45 in_use.lock
drwxrwxr-x. 2 alumno alumno 6 mar 25 01:45 queueset
[alumno@pasarela checkpoint]$
```

Figure 4: Directorio de control intermedio