

# PRÁCTICA N° 4.6

## Agente para detectar cambios en directorios y volcarlos en HDFS



- Nombre y apellidos: Alvaro Lucio-Villegas de Cea

Vamos a preparar el Fichero de Configuración.

- Utiliza el directorio flumeconf en la carpeta /home/cloudera, que ya utilizamos en la actividad 4.4
- Dentro de ese directorio, crea el fichero agentespooldir.conf.
- Utiliza como nombre del agente agentespooldir.
- Ese fichero debe contener todos los elementos de configuración que hemos visto en teoría. Vamos a concretarlos:

Vamos a ver cómo cambia la fuente, que ahora tiene un tipo distinto, denominado spooldir.

Por otro lado, hemos simplificado el sumidero, el sink, aunque veremos que se puede volver a complicar lo que necesitamos.

Fichero configuración:

Nombre:agentspooldir.conf

```
# Definición de componentes del agente
agentspooldir.sources = source1
agentspooldir.sinks = sink1
agentspooldir.channels = ch1
# Configuración de propiedades del source
agentspooldir.sources.source1.type = spooldir
agentspooldir.sources.source1.spoolDir = /tmp/datos
# Configuración de propiedades del sink
agentspooldir.sinks.sink1.type = hdfs
agentspooldir.sinks.sink1.hdfs.path = /flume/events1
# Configuración de un canal de tipo memoria
agentspooldir.channels.ch1.type = memory
agentspooldir.channels.ch1.capacity = 1000
agentspooldir.channels.ch1.transactionCapacity = 100
# Vincular source y sink al canal creado
agentspooldir.sources.source1.channels = ch1
agentspooldir.sinks.sink1.channel = ch1
```

```

agentespooldir.conf x generaficheros.sh x
# Definición de componentes del agente
agentespooldir.sources = source1
agentespooldir.sinks = sink1
agentespooldir.channels = ch1

# Configuración de propiedades del source
agentespooldir.sources.source1.type = spooldir
agentespooldir.sources.source1.spoolDir = /tmp/datos

# Configuración de propiedades del sink
agentespooldir.sinks.sink1.type = hdfs
agentespooldir.sinks.sink1.hdfs.path = /flume/events1

# Configuración de un canal de tipo memoria
agentespooldir.channels.ch1.type = memory
agentespooldir.channels.ch1.capacity = 1000
agentespooldir.channels.ch1.transactionCapacity = 100

# Vincular source y sink al canal creado
agentespooldir.sources.source1.channels = ch1
agentespooldir.sinks.sink1.channel = ch1
    
```

Propiedad	Valor	Significado
<b>FUENTE - SOURCE</b>		
<b>.sources.source1.type</b>	<b>Spooldir</b>	Tipo de fuente spooldir.
<b>.sources.source1.spoolDir</b>	<b>/tmp/datos</b>	Directorio de datos local que será vigilado por el agente flume para detectar nuevos ficheros y cargarlos
<b>SUMIDERO – SINK</b>		
<b>.sinks.sink1.type</b>	<b>hdfs</b>	Tipo de destino, en este caso hdfs.
<b>.sinks.sink1.hdfs.path</b>	<b>/flume1/events1</b>	Ruta del cluster hdfs en la que depositará los eventos leídos de la fuente (source)

Puede verse cómo hemos simplificado las propiedades y hemos cambiado el directorio de destino de HDFS, para que no se mezclen con los resultados de otras prácticas.

Vamos a desarrollar los distintos elementos de la práctica:

Prepara el programa que genera los ficheros:

- Crea el fichero generaficheros, que irá generando ficheros y moviéndolos al directorio de spool, es decir, a /tmp/datos.

- Utiliza el editor de texto y copia el siguiente código:

```
#!/bin/bash
i=1
while true
do
for ((j = 1; j <= 10; j++))
do
date >> /home/cloudera/log$i.txt
sleep 10
done
mv /home/cloudera/log$i.txt /tmp/datos/log$i.txt
let i=i+1
done
```



```
agentespooldir.conf x generaficheros.sh x
#!/bin/bash
i=1
while true
do
    for ((j = 1; j <= 10; j++))
    do
        date >> /home/cloudera/log$i.txt
        sleep 10
    done
    mv /home/cloudera/log$i.txt /tmp/datos/log$i.txt
    let i=i+1
done
```

Crea el directorio de Volcado en HDFS:

- Crea el directorio /flume1/events1 en HDFS, utilizando un terminal y las líneas de comando, mediante el comando “hdfs dfs.....”.

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /flume/events1
[cloudera@quickstart ~]$ hdfs dfs -ls /flume
Found 2 items
drwxr-xr-x - cloudera supergroup          0 2023-04-17 08:40 /flume/events
drwxr-xr-x - cloudera supergroup          0 2023-04-18 09:34 /flume/events1
[cloudera@quickstart ~]$ █
```

## Crea el Agente Flume

Siguiendo las indicaciones que aparecen en teoría para crear el agente Flume, construye el comando correspondiente, que tenga en cuenta lo siguiente:

- Nombre del Agente: `agentespooldir`
- Ruta del Directorio de Configuración: `/home/cloudera/flumeconf`
- Nombre del Fichero de configuración: `agentespooldir.conf`

```
[cloudera@quickstart ~]$ flume-ng agent -n agentespooldir -f /home/cloudera/flumeconf/agentespooldir.conf -Dflume.root.logger=INFO,console--Xmx512m
```

```
23/04/19 02:59:58 INFO conf.FlumeConfiguration: Processing:sink1
23/04/19 02:59:58 INFO conf.FlumeConfiguration: Added sinks: sink1 Agent: agentespooldir
23/04/19 02:59:58 INFO conf.FlumeConfiguration: Processing:sink1
23/04/19 02:59:58 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [agentespooldir]
23/04/19 02:59:58 INFO node.AbstractConfigurationProvider: Creating channels
23/04/19 02:59:58 INFO channel.DefaultChannelFactory: Creating instance of channel chl type memory
23/04/19 02:59:58 INFO node.AbstractConfigurationProvider: Created channel chl
23/04/19 02:59:58 INFO source.DefaultSourceFactory: Creating instance of source source1, type spooldir
23/04/19 02:59:58 INFO sink.DefaultSinkFactory: Creating instance of sink: sink1, type: hdfs
23/04/19 02:59:58 INFO node.AbstractConfigurationProvider: Channel chl connected to [source1, sink1]
23/04/19 02:59:58 INFO node.Application: Starting new configuration: { sourceRunners: {source1=EventDrivenSourceRunner: { source:Spool Directory source source1: { spoolDir: /tmp/datos } }} sinkRunners: {sink1=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@6ae0b6d8 counterGroup:{ name:null counters:{}} }} } channels: {chl=org.apache.flume.channel.MemoryChannel{name: chl}}
23/04/19 02:59:58 INFO node.Application: Starting Channel chl
23/04/19 02:59:58 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: chl: Successfully registered new MBean.
23/04/19 02:59:58 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: chl started
23/04/19 02:59:58 INFO node.Application: Starting Sink sink1
23/04/19 02:59:58 INFO node.Application: Starting Source source1
23/04/19 02:59:58 INFO source.SpooledDirectorySource: SpooledDirectorySource source starting with directory: /tmp/datos
23/04/19 02:59:58 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: sink1: Successfully registered new MBean.
23/04/19 02:59:58 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: sink1 started
23/04/19 02:59:58 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: source1: Successfully registered new MBean.
23/04/19 02:59:58 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: source1 started
```

## Ejecuta el script que genera los ficheros

- Una vez que el agente flume se está ejecutando, lanza el script “generaficheros”, para que comiencen a generarse los ficheros en el directorio de spool. Para ello, abre una consola y ejecuta el siguiente comando, en el directorio donde se encuentre el script

`./generaficheros`

```
[cloudera@quickstart ~]$ ./generaficheros.sh
```

Responde a las siguientes cuestiones:

- Observa la terminal donde has lanzado el agente flume. ¿Qué ocurre?

Se generan los registros de las acciones generadas del agente.

```
23/04/19 03:02:12 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681898532507.tmp
23/04/19 03:02:15 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681898532507.tmp
23/04/19 03:02:16 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681898532507.tmp to /flume/events1/FlumeData.1681898532508
23/04/19 03:02:16 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681898532508.tmp
```

- ¿Qué ocurre en el directorio de spool /temp/datos cuando se procesa un fichero por el agente flume?

Se procesan los ficheros de esta carpeta y una vez tratados por el agente se le agrega al nombre "Completed"

```
log1.txt.COMPLETED
[cloudera@quickstart datos]$ ls -l
total 4
-rw-rw-r-- 1 cloudera cloudera 319 Apr 19 03:02 log1.txt.COMPLETED
[cloudera@quickstart datos]$
```

- Observa el directorio HDFS desde el navegador, mediante la herramienta HUE. Consulta el contenido de los directorios y de los ficheros.

¿Cómo se organizan los ficheros generados por flume?

-Se organizan mediante la carpeta indicada en el fichero de configuración "events1"

¿Cómo se denominan los ficheros generados por flume?

-Por defecto tiene el prefijo el archivo creado de "FlumeData"

- Revisa el contenido de los ficheros generados e indica si son legibles. Investiga qué formato tienen esos ficheros en la documentación sobre los sinks de flume (<https://flume.apache.org/FlumeUserGuide.html#flume-sinks> ).

```

Home / flume / events1 / FlumeData.1681898532508

0000000: 53 45 51 06 21 6f 72 67 2e 61 70 61 63 68 65 2e SEQ.!org.apache.
0000010: 68 61 64 6f 6f 70 2e 69 6f 2e 4c 6f 6e 67 57 72 hadoop.io.LongWr
0000020: 69 74 61 62 6c 65 22 6f 72 67 2e 61 70 61 63 68 itable"org.apach
0000030: 65 2e 68 61 64 6f 6f 70 2e 69 6f 2e 42 79 74 65 e.hadoop.io.Byte
0000040: 73 57 72 69 74 61 62 6c 65 00 00 00 00 00 00 dd sWritable.....
0000050: fd 34 34 9b 9b f0 dc 87 f6 da 2f 9c c4 61 d2 00 .44...../..a..
0000060: 00 00 28 00 00 00 08 00 00 01 87 98 f6 fd e3 00 ..(.....
0000070: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 30 33 ...Wed Apr 19 03
0000080: 3a 30 32 3a 30 31 20 50 44 54 20 32 30 32 33 ff :02:01 PDT 2023.
0000090: ff ff ff dd fd 34 34 9b 9b f0 dc 87 f6 da 2f 9c .....44...../.
00000a0: c4 61 d2 .a.

```

- Modifica el formato de los ficheros generados para que sea legible (Recuerda que puedes hacerlo sin parar el agente). Revisa los nuevos ficheros flume para comprobar que son legibles.

```

generaficheros.sh x agentespooldir.conf x

# Definición de componentes del agente
agentespooldir.sources = source1
agentespooldir.sinks = sink1
agentespooldir.channels = ch1

# Configuración de propiedades del source
agentespooldir.sources.source1.type = spooldir
agentespooldir.sources.source1.spoolDir = /tmp/datos

# Configuración de propiedades del sink
agentespooldir.sinks.sink1.type = hdfs
agentespooldir.sinks.sink1.hdfs.path = /flume/events1
agentespooldir.sinks.sink1.hdfs.writeFormat = Text

# Configuración de un canal de tipo memoria
agentespooldir.channels.ch1.type = memory
agentespooldir.channels.ch1.capacity = 1000
agentespooldir.channels.ch1.transactionCapacity = 100

# Vincular source y sink al canal creado
agentespooldir.sources.source1.channels = ch1
agentespooldir.sinks.sink1.channel = ch1

```



Vemos que se han aplicado los cambios.

```
23/04/19 03:12:58 INFO node.Application: Starting new configuration: { sourceRunners: {source1=EventDrivenSourceRunner: { source:Pool Directory source source1: { spoolDir: /tmp/datos } }} sinkRunners:
sink1=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@34593f0a counterGroup: { name:null counters: {} }} channels: {ch1=org.apache.flume.channel.MemoryChannel{name: ch1}} }
23/04/19 03:12:58 INFO node.Application: Starting Channel ch1
23/04/19 03:12:58 INFO node.Application: Waiting for channel: ch1 to start. Sleeping for 500 ms
23/04/19 03:12:58 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: ch1 started
23/04/19 03:12:58 INFO node.Application: Starting Sink sink1
23/04/19 03:12:58 INFO node.Application: Starting Source source1
23/04/19 03:12:58 INFO source.SpoolDirectorySource: SpoolDirectorySource source starting with directory: /tmp/datos
23/04/19 03:12:58 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: sink1: Successfully registered new MBean.
23/04/19 03:12:58 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: sink1 started
23/04/19 03:12:58 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: source1: Successfully registered new MBean.
23/04/19 03:12:58 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: source1 started
```

En este caso ya se puede leer sin necesidad de cambiar el fichero de configuración.

Home / flume / events1 / 19 / 1034 / events.1681925717651

```
0000000: 53 45 51 06 21 6f 72 67 2e 61 70 61 63 68 65 2e SEQ.!org.apache.
0000010: 68 61 64 6f 6f 70 2e 69 6f 2e 4c 6f 6e 67 57 72 hadoop.io.LongWr
0000020: 69 74 61 62 6c 65 22 6f 72 67 2e 61 70 61 63 68 itable"org.apach
0000030: 65 2e 68 61 64 6f 6f 70 2e 69 6f 2e 42 79 74 65 e.hadoop.io.Byte
0000040: 73 57 72 69 74 61 62 6c 65 00 00 00 00 00 69 sWritable.....1
0000050: 86 eb 01 c1 56 5b 73 bc a3 bb 8e ff f5 d9 8f 00 ...V[s.....
0000060: 00 00 28 00 00 00 00 00 01 87 9a 95 be db 00 ..(.....
0000070: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 31 30 ...Wed Apr 19 10
0000080: 3a 33 33 3a 33 32 20 50 44 54 20 32 30 32 33 00 :33:32 PDT 2023.
0000090: 00 00 28 00 00 00 00 00 01 87 9a 95 be dc 00 ..(.....
00000a0: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 31 30 ...Wed Apr 19 10
00000b0: 3a 33 33 3a 34 32 20 50 44 54 20 32 30 32 33 00 :33:42 PDT 2023.
00000c0: 00 00 28 00 00 00 00 00 01 87 9a 95 be dc 00 ..(.....
00000d0: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 31 30 ...Wed Apr 19 10
00000e0: 3a 33 33 3a 35 32 20 50 44 54 20 32 30 32 33 00 :33:52 PDT 2023.
00000f0: 00 00 28 00 00 00 00 00 01 87 9a 95 be dc 00 ..(.....
0000100: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 31 30 ...Wed Apr 19 10
0000110: 3a 33 34 3a 30 32 20 50 44 54 20 32 30 32 33 00 :34:02 PDT 2023.
0000120: 00 00 28 00 00 00 00 00 01 87 9a 95 be dc 00 ..(.....
0000130: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 31 30 ...Wed Apr 19 10
0000140: 3a 33 34 3a 31 32 20 50 44 54 20 32 30 32 33 00 :34:12 PDT 2023.
0000150: 00 00 28 00 00 00 00 00 01 87 9a 95 be dd 00 ..(.....
0000160: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 31 30 ...Wed Apr 19 10
0000170: 3a 33 34 3a 32 32 20 50 44 54 20 32 30 32 33 00 :34:22 PDT 2023.
0000180: 00 00 28 00 00 00 00 00 01 87 9a 95 be dd 00 ..(.....
0000190: 00 00 1c 57 65 64 20 41 70 72 20 31 39 20 31 30 ...Wed Apr 19 10
```

- Para la ejecución de generaficheros. Utiliza el fichero “nba.csv” que se adjunta en la actividad y pásalo dentro de la máquina de Cloudera. Si tienes activado en Virtual Box la opción de Dispositivos/Arrastrar y Soltar, debería bastar con dejarlo en el Escritorio de la máquina Cloudera. Una vez que dispones del fichero dentro de la máquina, observa cuántos ficheros tienes actualmente en /flume1/events1. Prueba a mover el fichero a “nba.csv” a /tmp/datos, y observa lo que ocurre.

¿Cuántos ficheros nuevos se han creado en /flume1/events1 con el contenido del fichero nuevo?

-Se han generado 927 ficheros



## Visualización de los ficheros desde la HUE:

File Name	Size	Owner	Group	Permissions	Date
FlumeData.1681902432510	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432511	1.4 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432512	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432513	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432514	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432515	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432516	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432517	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432518	1.4 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432519	1.4 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432520	1.4 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432521	1.3 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432522	1.4 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432523	1.4 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM
FlumeData.1681902432524	1.4 KB	cloudera	supergroup	-rw-r--r--	April 19, 2023 04:07 AM

¿Cuánto espacio tiene cada uno aproximadamente?

-Tiene un tamaño aproximado de de 1.2 -1.4KB

Agregamos el CSV:

```
[c@cloudera@quickstart ~]$ ls /tmp/datos/
log1.txt.COMPLETED  log4.txt.COMPLETED  log7.txt.COMPLETED
log2.txt.COMPLETED  log5.txt.COMPLETED  log8.txt
log3.txt.COMPLETED  log6.txt.COMPLETED  nba.csv
[c@cloudera@quickstart ~]$
```

Ejecutamos el agente:

```
23/04/19 04:02:55 INFO hdfs.HDFSSequenceFile: writeFormat = Writable, UseRawLocalFileSystem = false
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175231.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175231.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175231.tmp to /flume/events1/FlumeData.1681902175231
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175232.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175232.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175232.tmp to /flume/events1/FlumeData.1681902175232
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175233.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175233.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175233.tmp to /flume/events1/FlumeData.1681902175233
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175234.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175234.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175234.tmp to /flume/events1/FlumeData.1681902175234
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175235.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175235.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175235.tmp to /flume/events1/FlumeData.1681902175235
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175236.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175236.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175236.tmp to /flume/events1/FlumeData.1681902175236
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175237.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175237.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175237.tmp to /flume/events1/FlumeData.1681902175237
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175238.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175238.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175238.tmp to /flume/events1/FlumeData.1681902175238
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175239.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Closing /flume/events1/FlumeData.1681902175239.tmp
23/04/19 04:02:55 INFO hdfs.BucketWriter: Renaming /flume/events1/FlumeData.1681902175239.tmp to /flume/events1/FlumeData.1681902175239
23/04/19 04:02:55 INFO hdfs.BucketWriter: Creating /flume/events1/FlumeData.1681902175240.tmp
```

Vemos desde la HUE el tamaño del fichero:

- ¿Cómo puedes hacer que cada fichero tenga 5 kb aproximadamente? Realiza el ajuste en el fichero de configuración, y vuelve a cargar el fichero “nba.csv”. Tendrás que cambiarle el nombre antes de arrastrarlo a /tmp/datos. Comprueba que los nuevos ficheros tienen aproximadamente 5kb cada uno e indica cuántos se han generado.

Añadimos de las columnas al fichero de configuración:

hdfs.rollSize = 4500 Cambiamos el tamaño del fichero.

hdfs.rollIntervall = 0 (Quitamos los valores predefinidos)

hdfs.rollCount = 0 (Quitamos los valores predefinidos)

## Fichero de Configuración:

```

# Definición de componentes del agente
agentespool.dir.sources = source1
agentespool.dir.sinks = sink1
agentespool.dir.channels = ch1

# Configuración de propiedades del source
agentespool.dir.sources.source1.type = spooldir
agentespool.dir.sources.source1.spoolDir = /tmp/datos

# Configuración de propiedades del sink
agentespool.dir.sinks.sink1.type = hdfs
agentespool.dir.sinks.sink1.hdfs.path = /flume/events1
agentespool.dir.sinks.sink1.hdfs.writeFormat = Text
agentespool.dir.sinks.sink1.hdfs.rollSize = 4500
agentespool.dir.sinks.sink1.hdfs.rollInterval = 0
agentespool.dir.sinks.sink1.hdfs.rollCount = 0

# Configuración de un canal de tipo memoria
agentespool.dir.channels.ch1.type = memory
agentespool.dir.channels.ch1.capacity = 1000
agentespool.dir.channels.ch1.transactionCapacity = 100

# Vincular source y sink al canal creado
agentespool.dir.sources.source1.channels = ch1
agentespool.dir.sinks.sink1.channel = ch1
    
```

## Comprobación:

- Activa nuevamente el script que generaficheros de forma automática. Modifica el fichero de configuración del agente, para que los nuevos ficheros se organicen dentro de /flume1/events1 por días y cada 2 minutos. Ajusta de nuevo el tamaño de los ficheros generados en HDFS a 1024 bytes.

## Fichero de Configuración:

```
# Definición de componentes del agente
agentespooldir.sources = source1
agentespooldir.sinks = sink1
agentespooldir.channels = ch1

# Configuración de propiedades del source
agentespooldir.sources.source1.type = spooldir
agentespooldir.sources.source1.spooldir = /tmp/datos

# Configuración de propiedades del sink
agentespooldir.sinks.sink1.type = hdfs

#agentespooldir.sinks.sink1.hdfs.path = /flume/events1
#agentespooldir.sinks.sink1.hdfs.writeFormat = Text
#agentespooldir.sinks.sink1.hdfs.rollSize = 4500
#agentespooldir.sinks.sink1.hdfs.rollIntervall = 0
#agentespooldir.sinks.sink1.hdfs.rollCount = 0

agentespooldir.sinks.sink1.hdfs.path = /flume/events1/%d/%H%M
agentespooldir.sinks.sink1.hdfs.filePrefix = events
agentespooldir.sinks.sink1.hdfs.round = true
agentespooldir.sinks.sink1.hdfs.roundValue = 2
agentespooldir.sinks.sink1.hdfs.roundUnit = minute
agentespooldir.sinks.sink1.hdfs.useLocalTimeStamp = true

# Configuración de un canal de tipo memoria
agentespooldir.channels.ch1.type = memory
agentespooldir.channels.ch1.capacity = 1000
agentespooldir.channels.ch1.transactionCapacity = 100

# Vincular source y sink al canal creado
agentespooldir.sources.source1.channels = ch1
agentespooldir.sinks.sink1.channel = ch1
```

## Comprobación:

1022

events.1681924946687

events.1681924946688

events.1681924946689

events.1681924946690

events.1681924946691

events.1681924946692

events.1681924946693

events.1681924946694

events.1681924946695

events.1681924946696

events.1681924946697

events.1681924946698

events.1681924946699

events.1681924946700

events.1681924946701

events.1681924946702

events.1681924946703

events.1681924946704

View as text

Edit file

Download

View file location

Refresh

Last modified

04/19/2023 5:22 PM

User

cloudera

Group

supergroup

Size

1.4 KB

Mode

100644

Home

/ flume / events1 / 19 / 1022 / events.1681924946687

```
000000: 53 45 51 06 21 6f 72 67 2e 61 70 61 63 68 65 2e SEQ!org.apache.
000010: 68 61 64 6f 6f 70 2e 69 6f 2e 4c 6f 6e 67 57 72 hadoop.io.LongWr
000020: 69 74 61 62 6c 65 22 6f 72 67 2e 61 70 61 63 68 itable"org.apach
000030: 65 2e 68 61 64 6f 6f 70 2e 69 6f 2e 42 79 74 65 e.hadoop.io.Byte
000040: 73 57 72 69 74 61 62 6c 65 00 00 00 00 00 4e sWritable.....N
000050: 83 54 c6 c0 e6 6c bb ff fd c7 c0 d4 26 a0 71 00 .T...1.....&q.
000060: 00 00 9c 00 00 00 00 00 00 01 87 9a 89 fb 4d 00 .....M.
000070: 00 00 90 ef bb bf 70 6c 61 79 65 72 20 6e 61 6d .....player nam
000080: 65 2c 53 65 61 73 6f 6e 2c 41 67 65 2c 54 65 61 e,Season, Age, Tea
000090: 6d 2c 4c 67 2c 50 6f 73 2c 47 2c 47 53 2c 4d 50 m,Lg,Pos,G,GS,MP
0000a0: 2c 46 47 2c 46 47 41 2c 46 47 25 2c 33 50 2c 33 ,FG,FGA,FG%,3P,3
0000b0: 50 41 2c 33 50 25 2c 32 50 2c 32 50 41 2c 32 50 PA,3P%,2P,2PA,2P
0000c0: 25 2c 65 46 47 25 2c 46 54 2c 46 54 41 2c 46 54 %,eFG%,FT,FTA,FT
0000d0: 25 2c 4f 52 42 2c 44 52 42 2c 54 52 42 2c 41 53 %,ORB,DRB,TRB,AS
0000e0: 54 2c 53 54 4c 2c 42 4c 4b 2c 54 4f 56 2c 50 46 T,STL,BLK,TOV,PF
0000f0: 2c 50 54 53 2c 76 61 63 69 6f 2c 54 72 70 20 44 ,PTS,vacio,Trp D
000100: 62 6c 6d 00 00 00 00 00 00 00 00 00 00 01 87 00 h1
```