

Flume. Práctica 1: Generación de datos secuenciales y almacenamiento en HDFS

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 va a generar una secuencia de números que se van a almacenar en el clúster de HDFS.

El agente de Flume utiliza los siguientes tipos de componentes:

- **Source:** secuencia (*seq*)
- **Channel:** memoria (*memory*)
- **Sink:** *hdfs*

Preparación del directorio en hdfs

Lo primero es crear en HDFS el directorio donde se va a guardar el resultado. Hay que tener en cuenta los permisos de este directorio para que el agente de Flume pueda guardar los archivos que se generan.

```
hdfs dfs -ls /user/alumno/flume
hdfs dfs -mkdir -p /user/alumno/flume/seqgen_data
```

Se comprueba que todo es correcto antes de lanzar la ejecución del agente:

```
[alumno@pasarela flume]$ hdfs dfs -ls /user/alumno/flume
Found 1 items
```

```
drwxr-xr-x  - alumno supergroup    0 2022-03-16 13:58 /user/alumno/flume/seqgen_data
```

Creación del fichero de configuración del agente

Al fichero de configuración se le pone el nombre **seqgen.conf** (nombre de ejemplo, significativo). Se crea en el host *pasarela* en un directorio que hayamos creado para practicar con Flume.

El fichero **seqgen.conf** tiene cinco partes:

- Nombres de los componentes del agente
- Configuración del Source
- Configuración del Sink
- Configuración del Channel
- Enlace de los componentes a través del canal

El contenido del fichero **seqgen.conf** sería el siguiente

Comprobar los datos de cada componente antes de guardar

```
# Nombramos a los componentes del agente
SeqGenAgent.sources = SeqSource
SeqGenAgent.channels = MemChannel
SeqGenAgent.sinks = HDFS

# Describimos el tipo de origen
SeqGenAgent.sources.SeqSource.type = seq

# Describimos el destino
SeqGenAgent.sinks.HDFS.type = hdfs
SeqGenAgent.sinks.HDFS.hdfs.path = hdfs://node1:8020/user/alumno/flume/seqgen_data/
SeqGenAgent.sinks.HDFS.hdfs.filePrefix = flume-caso1-seqgen
SeqGenAgent.sinks.HDFS.hdfs.rollInterval = 0
SeqGenAgent.sinks.HDFS.hdfs.rollCount = 1000
SeqGenAgent.sinks.HDFS.hdfs.fileType = DataStream
```

```
# Describimos la configuración del canal
SeqGenAgent.channels.MemChannel.type = memory
SeqGenAgent.channels.MemChannel.capacity = 1000
SeqGenAgent.channels.MemChannel.transactionCapacity = 100

# Unimos el origen y el destino a través del canal
SeqGenAgent.sources.SeqSource.channels = MemChannel
SeqGenAgent.sinks.HDFS.channel = MemChannel
```

Lanzamiento del agente de Flume

Antes de comenzar la ejecución del agente hay que comprobar que el usuario con el que lanzamos la orden tiene los permisos necesarios para escribir en el directorio HDFS donde va a escribir el agente.

La orden que lanza el agente es

```
flume-ng agent --conf /etc/flume-ng/conf --conf-file seqgen.conf --name SeqGenAgent \
-Dflume.root.logger=INFO,console
```

Los parámetros que se utilizan son:

- **--conf** /etc/flume-ng/conf : fichero de configuración de Flume
- **--conf-file** seqgen.conf : fichero de configuración del agente
- **--name** SeqGenAgent : nombre del agente, según se describe en el fichero
- **-Dflume.root.logger=INFO,console** : dónde se escriben los logs del agente

El agente será parado a los pocos segundos de iniciarse para que no llene de archivos el directorio. Estos archivos contienen simplemente una secuencia de números.

Comprobación del contenido del directorio en HDFS

Al acceder al directorio `/user/alumno/flume/seqgen_data` se puede ver que se han escrito muchos archivos

Home

/

user

/

alumno

/

flume

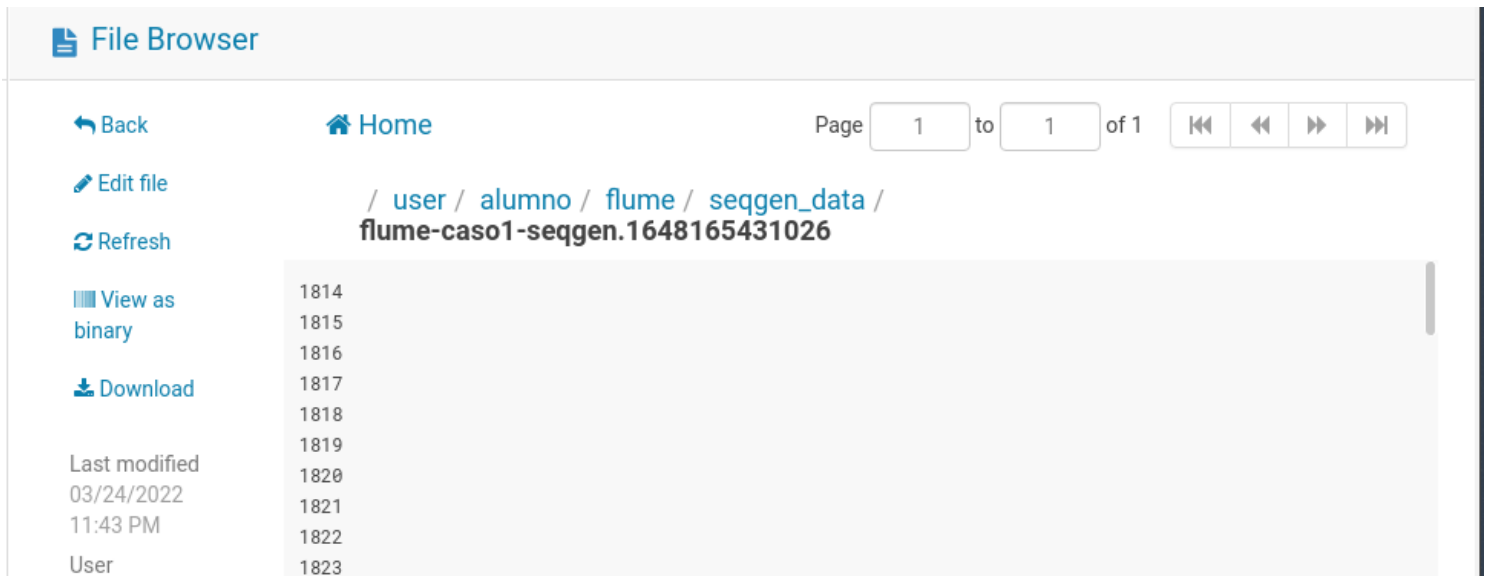
/

seqgen_data

Trash

Figure 1: Archivos de secuencia generados

El contenido de los archivos es simplemente una secuencia consecutiva numérica.



The screenshot shows a web-based File Browser interface. At the top, there's a header with a document icon and the text "File Browser". Below the header, there's a navigation bar with "Back" and "Home" links. To the right of the navigation bar, it says "Page 1 to 1 of 1" with navigation buttons. The main area shows a directory path: "/ user / alumno / flume / seqgen_data /" followed by the file name "flume-caso1-seqgen.1648165431026". On the left side, there's a sidebar with icons for "Back", "Edit file", "Refresh", "View as binary", and "Download". Below these icons, it shows "Last modified 03/24/2022 11:43 PM" and "User". The main content area displays a list of sequential numbers from 1814 to 1823.

File Name	Last modified	User
1814	03/24/2022 11:43 PM	1823
1815		
1816		
1817		
1818		
1819		
1820		
1821		
1822		
1823		

Figure 2: Contenido de los archivos seqgen

Para no llenar este directorio, una vez ejecutado, se pueden borrar los archivos creados

```
hdfs dfs -rm /user/alumno/flume/seqgen_data
```