Configuración agente flume - Local a HDFS - Proceso de extracción



Adrián Yared Armas de la Nuez

**Contenido**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

[**1. Objetivo 2**](#_heading=h.3znysh7)

[**2. Configuración del entorno Flume 2**](#_heading=h.2et92p0)

[**2.1 Inicia el Servicio de Flume 3**](#_heading=h.tyjcwt)

[**2.1.1 Comando 3**](#_heading=h.3dy6vkm)

[**2.1.2 Ejecución 3**](#_heading=h.1t3h5sf)

[**2.2 Crear una Configuración para Flume 3**](#_heading=h.4d34og8)

[**2.2.1 Comando 3**](#_heading=h.2s8eyo1)

[**2.2.2 Ejecución 4**](#_heading=h.17dp8vu)

[**2.3 Preparar el Directorio Local de Entrada 4**](#_heading=h.3rdcrjn)

[**2.3.1 Comando 4**](#_heading=h.26in1rg)

[**2.3.2 Ejecución 5**](#_heading=h.35nkun2)

[**2.3.3 Comando 5**](#_heading=h.1ksv4uv)

[**2.3.4 Ejecución 5**](#_heading=h.44sinio)

[**2.4 Ejecutar Apache Flume 5**](#_heading=h.2jxsxqh)

[**2.4.1 Comando 5**](#_heading=h.z337ya)

[**2.4.2 Ejecución 5**](#_heading=h.3j2qqm3)

[**2.4.3 Comando 6**](#_heading=h.1y810tw)

[**2.4.4 Ejecución 6**](#_heading=h.4i7ojhp)

[**2.4.5 Comando 6**](#_heading=h.2xcytpi)

[**2.4.6 Ejecución 6**](#_heading=h.1ci93xb)

## 

## 

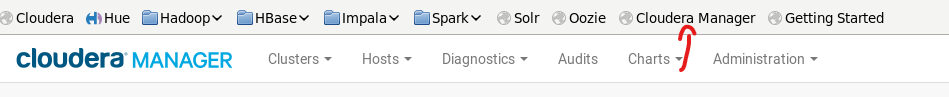
## 

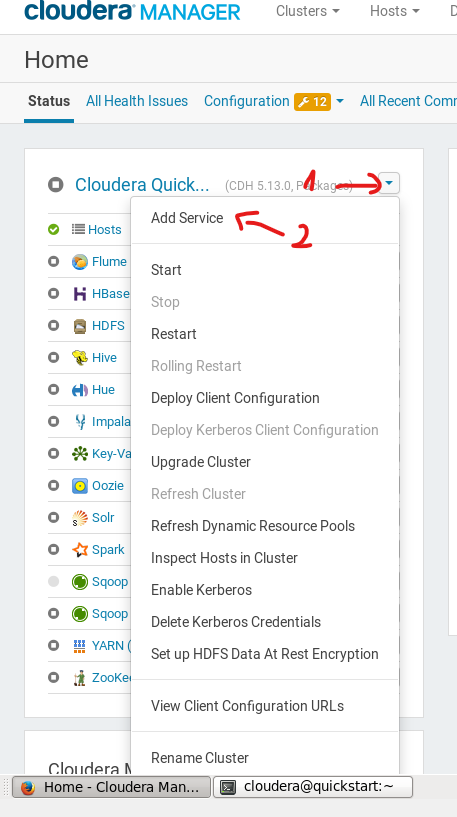
## **1. Objetivo**

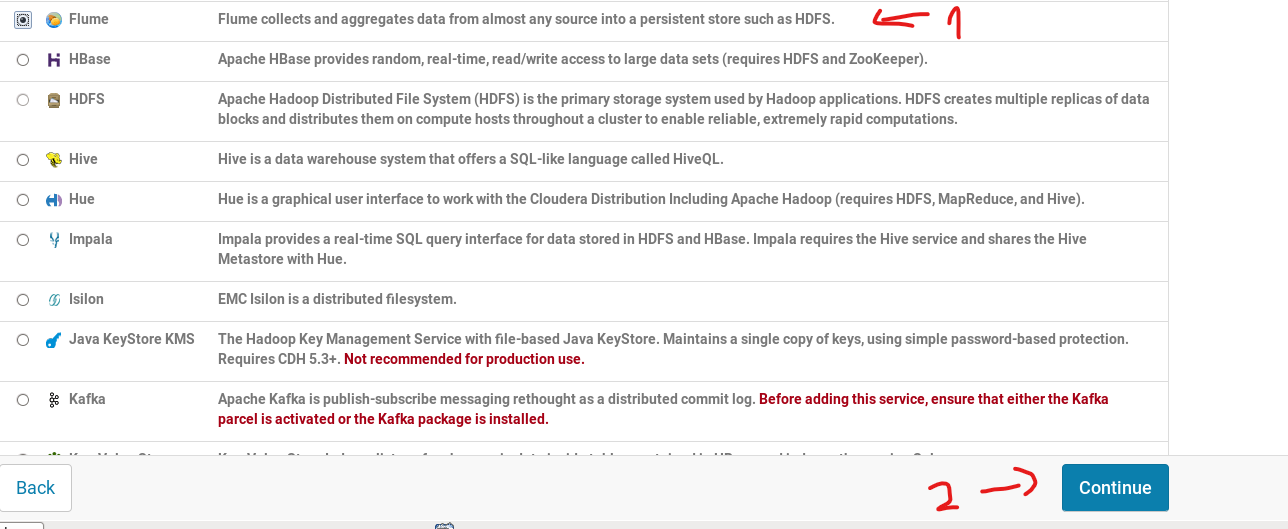
Automatizar la transferencia de archivos de texto desde un directorio local al sistema HDFS utilizando Flume y así analizar los datos con MapReduce (por ejemplo, el wordcount del ejercicio anterior).

## **2. Configuración del entorno Flume**

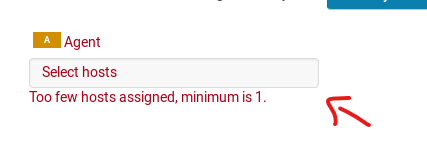
Verifica que Apache Flume esté instalado y configurado. En Cloudera, puedes instalar Flume desde Cloudera Manager si aún no está disponible.

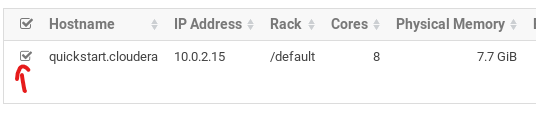














### **2.1 Inicia el Servicio de Flume**

#### **2.1.1 Comando**

sudo service flume-ng start

#### **2.1.2 Ejecución**

El servicio está iniciado automáticamente debido a cloudera.

### **2.2 Crear una Configuración para Flume**

#### **2.2.1 Comando**

**Archivo de Configuración (agente-flume.conf):**

# Nombre del agente

agent.sources = local-source

agent.sinks = hdfs-sink

agent.channels = memory-channel

# Configuración de la fuente (local-source)

agent.sources.local-source.type = spooldir

agent.sources.local-source.spoolDir = /path/to/local/input

agent.sources.local-source.fileHeader = true

# Configuración del canal (memory-channel)

agent.channels.memory-channel.type = memory

agent.channels.memory-channel.capacity = 1000

agent.channels.memory-channel.transactionCapacity = 100

# Configuración del sink (hdfs-sink)

agent.sinks.hdfs-sink.type = hdfs

agent.sinks.hdfs-sink.hdfs.path =/user/cloudera/flume-output

agent.sinks.hdfs-sink.hdfs.fileType = DataStream

agent.sinks.hdfs-sink.hdfs.writeFormat = Text

agent.sinks.hdfs-sink.hdfs.rollSize = 0

agent.sinks.hdfs-sink.hdfs.rollCount = 10

agent.sinks.hdfs-sink.hdfs.batchSize = 100

# Conectar componentes

agent.sources.local-source.channels = memory-channel

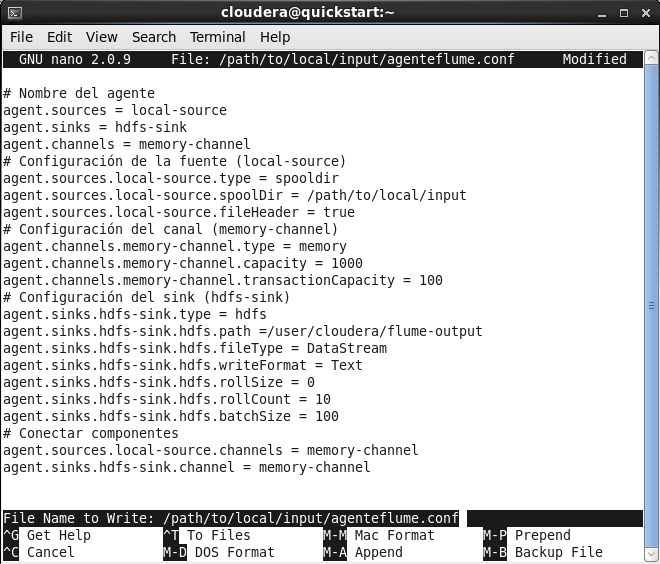
agent.sinks.hdfs-sink.channel = memory-channel

#### **2.2.2 Ejecución**

Creo la ruta y lo configuro en spoolDir:







### **2.3 Preparar el Directorio Local de Entrada**

#### **2.3.1 Comando**

**Crea el Directorio de Entrada Local:**

mkdir -p /path/to/local/input

#### 

#### **2.3.2 Ejecución**

Creo el directorio:



#### **2.3.3 Comando**

**Copia el Archivo de Texto al Directorio: Coloca elquijote.txt (u otros archivos de texto) en este directorio para que Flume lo procese:**

cp /path/to/elquijote.txt /path/to/local/input/

#### **2.3.4 Ejecución**

Compruebo si existe:



Lo copio en la ruta:



### **2.4 Ejecutar Apache Flume**

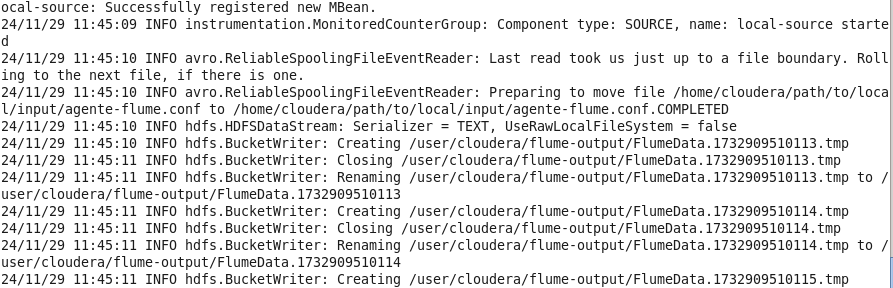
#### **2.4.1 Comando**

**Inicia el Agente Flume: Ejecuta Flume con el archivo de configuración:**

flume-ng agent --conf /path/to/local/input --conf-file path/to/local/input/agente-flume.conf --name agent -Dflume.root.logger=INFO,console

#### **2.4.2 Ejecución**



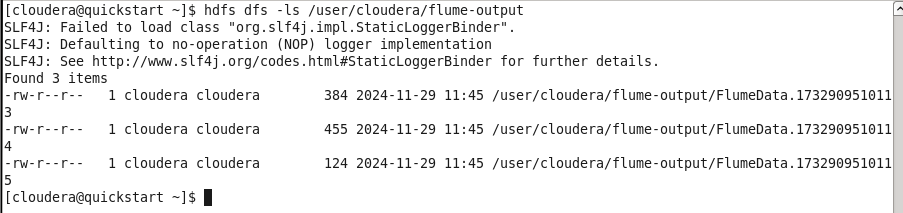


#### **2.4.3 Comando**

**Verifica que los Datos Llegaron a HDFS: Lista los archivos en la ruta configurada en el sink (/user/cloudera/flume-output):**

hdfs dfs -ls /user/cloudera/flume-output

#### **2.4.4 Ejecución**



#### **2.4.5 Comando**

**Muestra el Contenido Ingerido:**

hdfs dfs -cat /user/cloudera/flume-output/\*

#### **2.4.6 Ejecución**

