

## Configuración agente flume - Local a HDFS - Proceso de extracción

The Cloudera logo consists of a solid orange square. Centered within this square is the word "CLOUDERA" in a bold, white, sans-serif typeface. The letters are evenly spaced and have a slightly distressed or hand-drawn appearance.

**CLOUDERA**

Adrián Yared Armas de la Nuez



## Contenido

---

<b>1. Objetivo.....</b>	<b>2</b>
<b>2. Configuración del entorno Flume.....</b>	<b>2</b>
<b>2.1 Inicia el Servicio de Flume.....</b>	<b>3</b>
2.1.1 Comando.....	3
2.1.2 Ejecución.....	3
<b>2.2 Crear una Configuración para Flume.....</b>	<b>3</b>
2.2.1 Comando.....	3
2.2.2 Ejecución.....	4
<b>2.3 Preparar el Directorio Local de Entrada.....</b>	<b>4</b>
2.3.1 Comando.....	4
2.3.2 Ejecución.....	5
2.3.3 Comando.....	5
2.3.4 Ejecución.....	5
<b>2.4 Ejecutar Apache Flume.....</b>	<b>5</b>
2.4.1 Comando.....	5
2.4.2 Ejecución.....	5
2.4.3 Comando.....	6
2.4.4 Ejecución.....	6
2.4.5 Comando.....	6
2.4.6 Ejecución.....	6



### 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.

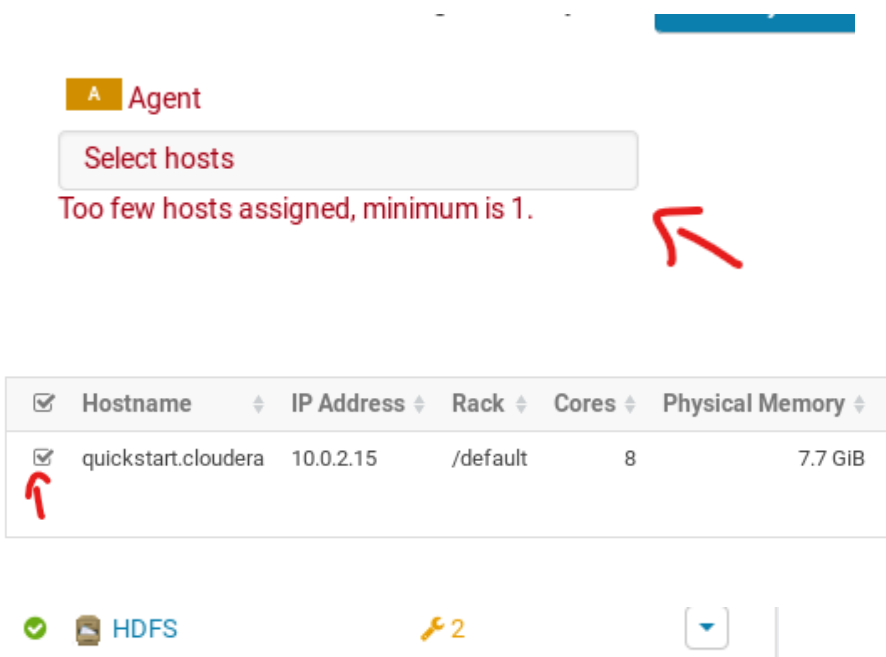
Cloudera Quickstart (CDH 5.13.0, Paragon) →

- Start
- Stop
- Restart
- Rolling Restart
- Deploy Client Configuration
- Deploy Kerberos Client Configuration
- Upgrade Cluster
- Refresh Cluster
- Refresh Dynamic Resource Pools
- Inspect Hosts in Cluster
- Enable Kerberos
- Delete Kerberos Credentials
- Set up HDFS Data At Rest Encryption
- View Client Configuration URLs
- Rename Cluster

Flume	Flume collects and aggregates data from almost any source into a persistent store such as HDFS. ← 1
HBase	Apache HBase provides random, real-time, read/write access to large data sets (requires HDFS and ZooKeeper).
HDFS	Apache Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute hosts throughout a cluster to enable reliable, extremely rapid computations.
Hive	Hive is a data warehouse system that offers a SQL-like language called HiveQL.
Hue	Hue is a graphical user interface to work with the Cloudera Distribution Including Apache Hadoop (requires HDFS, MapReduce, and Hive).
Impala	Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires the Hive service and shares the Hive Metastore with Hue.
Isilon	EMC Isilon is a distributed filesystem.
Java KeyStore KMS	The Hadoop Key Management Service with file-based Java KeyStore. Maintains a single copy of keys, using simple password-based protection. Requires CDH 5.3+. <b>Not recommended for production use.</b>
Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log. <b>Before adding this service, ensure that either the Kafka parcel is activated or the Kafka package is installed.</b>

Back → Continue

HBase HDFS Solr ZooKeeper



**A Agent**

Select hosts

Too few hosts assigned, minimum is 1.

<input checked="" type="checkbox"/>	Hostname	IP Address	Rack	Cores	Physical Memory
<input checked="" type="checkbox"/>	quickstart.cloudera	10.0.2.15	/default	8	7.7 GiB

✓ HDFS 2

## 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
```



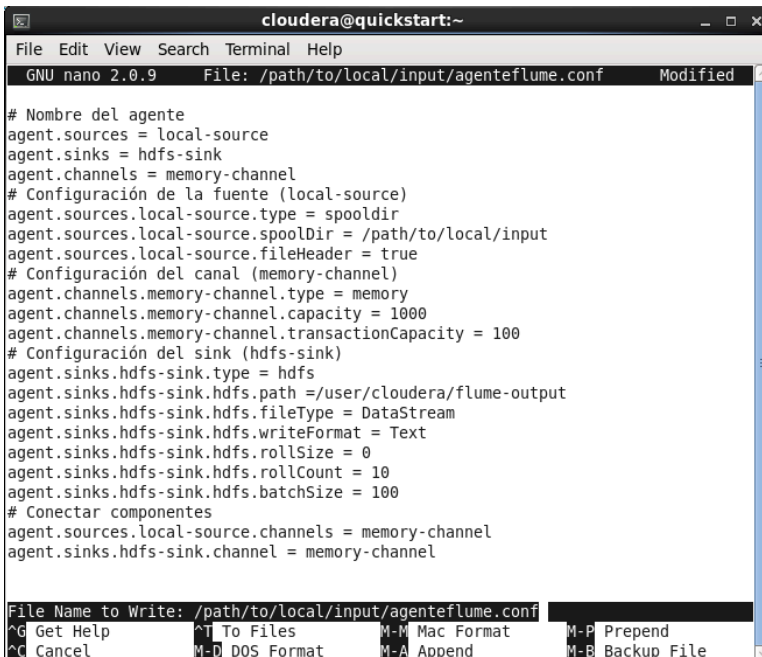
## Configuración agente flume - Local a HDFS - Proceso de extracción

```
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:

```
[cloudera@quickstart ~]$ mkdir -p ~/path/to/local/input
[cloudera@quickstart ~]$ sudo mkdir -p /path/to/local/input
[cloudera@quickstart ~]$ sudo chmod 755 /path
[cloudera@quickstart ~]$ sudo nano /path/to/local/input/agenteflume.conf
```



```
cloudera@quickstart:~
File Edit View Search Terminal Help
GNU nano 2.0.9 File: /path/to/local/input/agenteflume.conf Modified

# 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

File Name to Write: /path/to/local/input/agenteflume.conf
^G Get Help      ^T To Files      M-M Mac Format   M-P Prepend
^C Cancel        M-D DOS Format   M-A Append       M-B Backup File
```

## 2.3 Preparar el Directorio Local de Entrada

### 2.3.1 Comando

**Creo el Directorio de Entrada Local:**

```
mkdir -p /path/to/local/input
```



### 2.3.2 Ejecución

Creo el directorio:

```
[cloudera@quickstart ~]$ mkdir -p /path/to/local/input
```

### 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:

```
[cloudera@quickstart ~]$ ls -l path/to/local/input
total 4
-rw-r--r-- 1 root root 964 Nov 29 11:06 agente-flume.conf
```

Lo copio en la ruta:

```
[cloudera@quickstart ~]$ sudo cp DonQuixotedelaMancha.txt /path/to/local/input/
[cloudera@quickstart ~]$ █
```

## 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

```
[cloudera@quickstart ~]$ flume-ng agent --conf /path/to/flume/conf --conf-file path/to/local/input/agente-flume.conf --name agent -Dflume.root.logger=INFO,console
```

```
ocal-source: Successfully registered new MBean.
24/11/29 11:45:09 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: local-source started
24/11/29 11:45:10 INFO avro.ReliableSpoolingFileEventReader: Last read took us just up to a file boundary. Rolling to the next file, if there is one.
24/11/29 11:45:10 INFO avro.ReliableSpoolingFileEventReader: Preparing to move file /home/cloudera/path/to/local/input/agente-flume.conf to /home/cloudera/path/to/local/input/agente-flume.conf.COMPLETED
24/11/29 11:45:10 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
24/11/29 11:45:10 INFO hdfs.BucketWriter: Creating /user/cloudera/flume-output/FlumeData.1732909510113.tmp
24/11/29 11:45:11 INFO hdfs.BucketWriter: Closing /user/cloudera/flume-output/FlumeData.1732909510113.tmp
24/11/29 11:45:11 INFO hdfs.BucketWriter: Renaming /user/cloudera/flume-output/FlumeData.1732909510113.tmp to /user/cloudera/flume-output/FlumeData.1732909510113
24/11/29 11:45:11 INFO hdfs.BucketWriter: Creating /user/cloudera/flume-output/FlumeData.1732909510114.tmp
24/11/29 11:45:11 INFO hdfs.BucketWriter: Closing /user/cloudera/flume-output/FlumeData.1732909510114.tmp
24/11/29 11:45:11 INFO hdfs.BucketWriter: Renaming /user/cloudera/flume-output/FlumeData.1732909510114.tmp to /user/cloudera/flume-output/FlumeData.1732909510114
24/11/29 11:45:11 INFO hdfs.BucketWriter: Creating /user/cloudera/flume-output/FlumeData.1732909510115.tmp
```



### 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

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/flume-output
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
Found 3 items
-rw-r--r--  1 cloudera cloudera      384 2024-11-29 11:45 /user/cloudera/flume-output/FlumeData.1732909510113
-rw-r--r--  1 cloudera cloudera      455 2024-11-29 11:45 /user/cloudera/flume-output/FlumeData.1732909510114
-rw-r--r--  1 cloudera cloudera      124 2024-11-29 11:45 /user/cloudera/flume-output/FlumeData.1732909510115
[cloudera@quickstart ~]$
```

### 2.4.5 Comando

**Muestra el Contenido Ingerido:**

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

### 2.4.6 Ejecución

```
-Dulce esperanza mía,
que, rompiendo imposibles y malezas,
sigues firme la vía
que tú mesma te finges y aderezas:
no te desmaye el verte
a cada paso junto al de tu muerte.
No alcanzan perezosos
honrados triunfos ni vitoria alguna,
ni pueden ser dichosos
los que, no contrastando a la fortuna,
entregan, desvalidos,
al ocio blando todos los sentidos.
Que amor sus glorias venda
caras, es gran razón, y es trato justo,
pues no hay más rica prenda
que la que se quilata por su gusto;
y es cosa manifiesta
que no es de estima lo que poco cuesta.
Amorosas porfías
tal vez alcanzan imposibles cosas;
y así, aunque con las mías
sigo de amor las más dificultosas,
[cloudera@quickstart ~]$
```