

Configuración agente flume - Servidor a HDFS - Proceso de extracción



Adrián Yared Armas de la Nuez



Contenido

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

The screenshot shows the Cloudera Manager interface. At the top, there's a navigation bar with links to Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. Below this is the Cloudera Manager logo and a navigation menu with Clusters, Hosts, Diagnostics, Audits, Charts, and Administration. The main content area shows the 'Home' page with tabs for Status, All Health Issues, Configuration (12), and All Recent Commands. A dropdown menu is open for 'Cloudera Quickstart' (CDH 5.13.0, Paragon), showing options like 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, and Rename Cluster. Red arrows point to the 'Add Service' dropdown and the 'Continue' button. Below the dropdown, there's a table listing services and their descriptions:


Service	Description
Flume	Flume collects and aggregates data from almost any source into a persistent store such as HDFS.
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+. Not recommended for production use.
Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log. Before adding this service, ensure that either the Kafka parcel is activated or the Kafka package is installed.

At the bottom of the table, there's a 'Back' button and a 'Continue' button. Red arrows point to the 'Continue' button.


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-exec.conf):

```
# Nombre del agente
agent.sources = exec-source
agent.sinks = hdfs-sink
agent.channels = memory-channel
agent.sources.exec-source.type = exec
agent.sources.exec-source.command = curl -o /tmp/quijote.txt
https://babel.upm.es/~angel/teaching/pps/quijote.txt
agent.sources.exec-source.command = cat /tmp/quijote.txt
agent.sources.exec-source.channels = memory-channel
# Configuración del canal (MemoryChannel)
agent.channels.memory-channel.type = memory
agent.channels.memory-channel.capacity = 1000
agent.channels.memory-channel.transactionCapacity = 100
# Configuración del sink (HDFSSink)
agent.sinks.hdfs-sink.type = hdfs
```



Configuración agente flume - Servidor a HDFS - Proceso de extracción

```
agent.sinks.hdfs-sink.hdfs.path = /user/cloudera/flume-output-server
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
agent.sinks.hdfs-sink.channel = memory-channel
```

2.2.2 Ejecución

```
agentes [cloudera@quickstart ~]$ nano agenteflumeexec.conf

# Nombre del agente
agent.sources = exec-source
agent.sinks = hdfs-sink
agent.channels = memory-channel
agent.sources.exec-source.type = exec
agent.sources.exec-source.command = curl -o /tmp/quijote.txt https://babel.upm.$
agent.sources.exec-source.command = cat /tmp/ElQuixotedelaMancha.txt
agent.sources.exec-source.channels = memory-channel
# Configuración del canal (MemoryChannel)
agent.channels.memory-channel.type = memory
agent.channels.memory-channel.capacity = 1000
agent.channels.memory-channel.transactionCapacity = 100
# Configuración del sink (HDFSSink) 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
[ Read 20 lines ]
^G Get Help  ^O WriteOut  ^R Read File ^V Prev Page ^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is  ^N Next Page ^U UnCut Text ^T To Spell
```

Archivo subido:

```
clipboardcache-3
clipboardcache-4
clipboardcache-5
clipboardcache-6
clipboardcache-7
clipboardcache-8
clipboardcache-9
cmflistener-stderr---agent-11886-1732739845-0B_n9e.log
cmflistener-stderr---agent-5513-1732904979-uEkbJf.log
cmflistener-stdout---agent-11886-1732739845-C7WKwM.log
cmflistener-stdout---agent-5513-1732904979-uEP49x.log
DonQuixotedelaMancha.txt
edc2aaaa-6ffd-4a09-9314-741284873281_resources
ff2679c4-80b3-43af-9a16-e4c2e6e8df49_resources
hadoop
hadoop-unjar1779017763898444030
hadoop-unjar30537379567640770
hadoop-unjar3693216385418239264
hadoop-unjar3765740324049213922
hadoop-unjar4203334404121345879
```



2.3 Preparar el Directorio Local de Entrada

2.3.1 Comando

Crea el Directorio de Entrada Local:

```
mkdir -p tmp
```

2.3.2 Ejecución

Creo el directorio:

```
[cloudera@quickstart ~]$ mkdir -p tmp
```

2.3.3 Comando

```
mv DonQuixotedelaMancha.txt /tmp/
```

2.3.4 Ejecución

Muevo el quijote a temp:

```
[cloudera@quickstart ~]$ mv DonQuixotedelaMancha.txt /tmp/
```

2.4 Ejecutar Apache Flume

2.4.1 Comando

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

```
flume-ng agent --conf /cloudera/home --conf-file /cloudera/home/agente-flume-exec.conf  
--name agent -Dflume.root.logger=INFO,console
```

2.4.2 Ejecución

Ejecución:

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

Resultado:

```
cloudera/ut2/flume-exec/flume-output/FlumeData.1733164057124  
24/12/02 10:29:35 INFO hdfs.BucketWriter: Creating /user/cloudera/ut2/flume-exec/flume-output/FlumeData.1733164057125.tmp  
24/12/02 10:29:35 INFO hdfs.BucketWriter: Closing /user/cloudera/ut2/flume-exec/flume-output/FlumeData.1733164057125.tmp  
24/12/02 10:29:35 INFO hdfs.BucketWriter: Renaming /user/cloudera/ut2/flume-exec/flume-output/FlumeData.1733164057125.tmp to /user/  
cloudera/ut2/flume-exec/flume-output/FlumeData.1733164057125
```



2.4.3 Comando

Quitar modo seguro (con el activo no funciona):

```
hdfs hdfs dfsadmin -safemode get
```

Muestro el contenido:

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

2.4.4 Ejecución

Muestra el Contenido Ingerido:

```
[cloudera@quickstart ~]$ hdfs dfs -cat /user/cloudera/ut2/flume-exec/flume-output/*
```

```
-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 ~]$
```