

Apache Flume User Guide

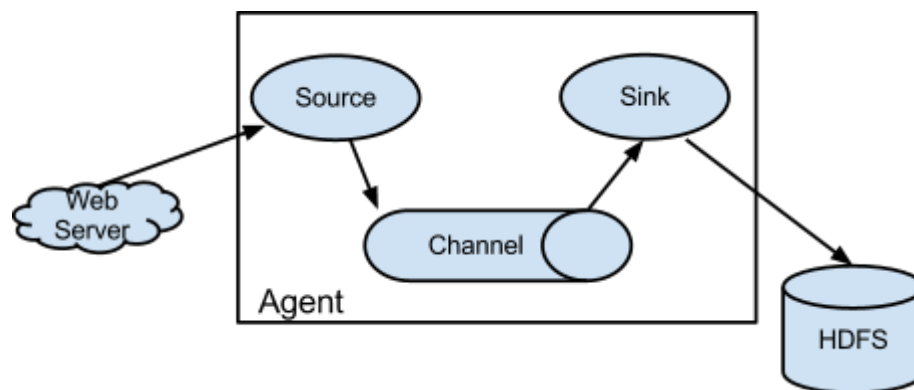
Table of Contents

What is Apache Flume?	2
Flume Architecture	2
What is Flume Agent/Agent?	2
What is Flume Source/Source?	2
What is Flume Channel/Channel?	2
What is Flume Sink/Sink?	3
What is Interceptor?	3
What is Channel Selector?	3
What is Sink Processors?	3
What is Event Serializers?	3
Simple Flume Agent Example	4
References:	8

What is Apache Flume?

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store.

Flume Architecture



What is Flume Agent/Agent?

A Flume agent is a process (JVM) that hosts the components that allow Events to flow from an external source to an external destination.

An Event is a unit of data that flows through a Flume agent.

The Event flows from Source to Channel to Sink, and is represented by an implementation of the Event interface.

An Event carries a payload (byte array) that is accompanied by an optional set of headers (string attributes).

What is Flume Source/Source?

A Source consumes Events having a specific format, and those Events are delivered to the Source by an external source like a web server.

When a Source receives an Event, it stores it into one or more Channels.

What is Flume Channel/Channel?

The Channel is a passive store that holds the Event until that Event is consumed by a Sink.

One type of Channel available in Flume is the FileChannel which uses the local filesystem as its backing store.

What is Flume Sink/Sink?

A Sink is responsible for removing an Event from the Channel and putting it into an external repository like HDFS (in the case of an HDFSEventSink) or forwarding it to the Source at the next hop of the flow.

The Source and Sink within the given agent run asynchronously with the Events staged in the Channel.

What is Interceptor?

An interceptor can modify or even drop events based on any criteria chosen by the developer.

or

Interceptors in Flume are those who have the capability to modify/drop events in-flight.

What is Channel Selector?

Channel selector is used to determine which channel we should select to transfer the data in case of multiple channels.

or

Channel selector is the component of Flume that determines which channel that particular Flume event should go into when a group of channels exists.

1. Replicating channel selector
2. Multiplexing channel selector

What is Sink Processors?

Sink processor is mechanism by which you can create a fail-over task and load balancing.

or

To invoke a particular sink from the selected group of sinks, we generally use sink processors.

Also, we use it to create failover paths for our sinks or load balance events across multiple sinks from a channel.

What is Event Serializers?

An Event Serializer is the mechanism by which a FlumeEvent is converted into another format for output. By default, the text serializer, which outputs just the Flume event body, is used.

There is another serializer, `header_and_text`, which outputs both the headers and the body.

Simple Flume Agent Example

Flume Agent Definition Configuration

```
# Define a source, a channel, and a sink
flume_agent1.sources = spooldir_source
flume_agent1.channels = mem_channel
flume_agent1.sinks = hdfs_sink

# Set the source type to Spooling Directory and set the directory
# location to /home/cloudera/flume_agent1_data

flume_agent1.sources.spooldir_source.type = spooldir
flume_agent1.sources.spooldir_source.spoolDir = /home/cloudera/flume_agent1_data
flume_agent1.sources.spooldir_source.basenameHeader = true

# Configure the channel as simple in-memory queue
flume_agent1.channels.mem_channel.type = memory
flume_agent1.channels.mem_channel.capacity = 1000

# Define the HDFS sink and set its path to your target HDFS directory
flume_agent1.sinks.hdfs_sink.type = hdfs
flume_agent1.sinks.hdfs_sink.hdfs.path =
hdfs://quickstart.cloudera:8020/user/cloudera/data/flume/flume_agent1_data
flume_agent1.sinks.hdfs_sink.hdfs.fileType = DataStream
flume_agent1.sinks.hdfs_sink.hdfs.writeFormat = Text

# Disable rollover functionality as we want to keep the original files
flume_agent1.sinks.hdfs_sink.rollCount = 0
flume_agent1.sinks.hdfs_sink.rollInterval = 0
flume_agent1.sinks.hdfs_sink.rollSize = 0
flume_agent1.sinks.hdfs_sink.idleTimeout = 0

# Set the files to their original name
flume_agent1.sinks.hdfs_sink.hdfs.filePrefix = %{basename}

# Connect source and sink
flume_agent1.sources.spooldir_source.channels = mem_channel
flume_agent1.sinks.hdfs_sink.channel = mem_channel
```

Run Flume Agent

```
flume-ng agent -n flume_agent1 -f /home/cloudera/flume_example/flume-agent-spooldir.conf
```



```
cloudera@quickstart:~  
./../search/lib/lucene-suggest.jar:/usr/lib/flume-ng/./search/lib/maximum-dh-1.0.0.jar:/usr/lib/flume-ng/./search/lib/metadata-extractor-2.6.2.jar:/usr/lib/  
flume-ng/./search/lib/metrics-core-3.0.2.jar:/usr/lib/flume-ng/./search/lib/metrics-healthchecks-3.0.2.jar:/usr/lib/flume-ng/./search/lib/metrics-json-3.  
0.2.jar:/usr/lib/flume-ng/./search/lib/metrics-jvm-3.0.2.jar:/usr/lib/flume-ng/./search/lib/metrics-servlets-3.0.2.jar:/usr/lib/flume-ng/./search/lib/netc  
df-4.2-min.jar:/usr/lib/flume-ng/./search/lib/netty-3.10.5.Final.jar:/usr/lib/flume-ng/./search/lib/netty-all-4.0.23.Final.jar:/usr/lib/flume-ng/./search/  
lib/noggit-0.5.jar:/usr/lib/flume-ng/./search/lib/org.restlet-2.1.4.jar:/usr/lib/flume-ng/./search/lib/org.restlet.ext.servlet-2.1.4.jar:/usr/lib/flume-ng/  
./search/lib/paranamer-2.3.jar:/usr/lib/flume-ng/./search/lib/parquet-avro.jar:/usr/lib/flume-ng/./search/lib/parquet-column.jar:/usr/lib/flume-ng/./sear  
ch/lib/parquet-common.jar:/usr/lib/flume-ng/./search/lib/parquet-encoding.jar:/usr/lib/flume-ng/./search/lib/parquet-format.jar:/usr/lib/flume-ng/./search  
lib/parquet-hadoop.jar:/usr/lib/flume-ng/./search/lib/parquet-jackson.jar:/usr/lib/flume-ng/./search/lib/pdfbox-1.8.4.jar:/usr/lib/flume-ng/./search/lib/  
poi-3.10-beta2.jar:/usr/lib/flume-ng/./search/lib/poi-ooxml-3.10-beta2.jar:/usr/lib/flume-ng/./search/lib/poi-ooxml-schemas-3.10.1.jar:/usr/lib/flume-ng/./  
search/lib/poi-scratchpad-3.10-beta2.jar:/usr/lib/flume-ng/./search/lib/protobuf-java-2.5.0.jar:/usr/lib/flume-ng/./search/lib/rome-0.9.jar:/usr/lib/flume  
-ng/./search/lib/Saxon-HE-9.5.1-5.jar:/usr/lib/flume-ng/./search/lib/snakeyaml-1.10.jar:/usr/lib/flume-ng/./search/lib/snappy-java-1.0.4.1.jar:/usr/lib/fl  
ume-ng/./search/lib/solr-cell.jar:/usr/lib/flume-ng/./search/lib/solr-core.jar:/usr/lib/flume-ng/./search/lib/solr-solrj.jar:/usr/lib/flume-ng/./search/l  
ib/spatial4j-0.4.1.jar:/usr/lib/flume-ng/./search/lib/tagsoup-1.2.1.jar:/usr/lib/flume-ng/./search/lib/tika-core-1.5.jar:/usr/lib/flume-ng/./search/lib/ti  
ka-parsers-1.5.jar:/usr/lib/flume-ng/./search/lib/tika-xmp-1.5.jar:/usr/lib/flume-ng/./search/lib/ua-parser-1.3.0.jar:/usr/lib/flume-ng/./search/lib/vorbi  
s-java-core-0.1.jar:/usr/lib/flume-ng/./search/lib/vorbis-java-core-0.1-tests.jar:/usr/lib/flume-ng/./search/lib/vorbis-java-tika-0.1.jar:/usr/lib/flume-ng  
./search/lib/watx-asl-3.2.7.jar:/usr/lib/flume-ng/./search/lib/worcelimpl-2.9.1.jar:/usr/lib/flume-ng/./search/lib/xml-apis-1.3.04.jar:/usr/lib/flume-ng/  
./search/lib/xmlbeans-2.6.0.jar:/usr/lib/flume-ng/./search/lib/xmlenc-0.52.jar:/usr/lib/flume-ng/./search/lib/xmpcore-5.1.2.jar:/usr/lib/flume-ng/./searc  
h/lib/xz-1.0.jar:/usr/lib/flume-ng/./search/lib/zookeeper.jar' -Djava.library.path=/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/native:/usr/lib/hbase/bin  
/./lib/native/Linux-amd64-64 org.apache.flume.node.Application -n flume_agent1 -f /home/cloudera/flume_example/flume-agent-spoolidir.conf  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
19/04/25 11:06:18 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting  
19/04/25 11:06:18 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/home/cloudera/flume_example/flume-agent-spoolidir.conf  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Added sinks: hdfs sink Agent: flume_agent1  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [flume_agent1]  
19/04/25 11:06:18 INFO node.AbstractConfigurationProvider: Creating channels  
19/04/25 11:06:18 INFO channel.DefaultChannelFactory: Creating instance of channel mem_channel type memory  
19/04/25 11:06:18 INFO node.AbstractConfigurationProvider: Created channel mem_channel  
19/04/25 11:06:18 INFO source.DefaultSourceFactory: Creating instance of source spoolidir_source, type spoolidir  
19/04/25 11:06:18 INFO sink.DefaultSinkFactory: Creating instance of sink: hdfs sink, type: hdfs  
19/04/25 11:06:18 INFO node.AbstractConfigurationProvider: Channel mem_channel connected to [spoolidir_source, hdfs_sink]
```

```
cloudera@quickstart:~  
19/04/25 11:06:18 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/home/cloudera/flume_example/flume-agent-spoolidir.conf  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Processing:hdfs_sink  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Added sinks: hdfs sink Agent: flume_agent1  
19/04/25 11:06:18 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [flume_agent1]  
19/04/25 11:06:18 INFO node.AbstractConfigurationProvider: Creating channels  
19/04/25 11:06:18 INFO channel.DefaultChannelFactory: Creating instance of channel mem_channel type memory  
19/04/25 11:06:18 INFO node.AbstractConfigurationProvider: Created channel mem_channel  
19/04/25 11:06:18 INFO source.DefaultSourceFactory: Creating instance of source spoolidir_source, type spoolidir  
19/04/25 11:06:18 INFO sink.DefaultSinkFactory: Creating instance of sink: hdfs sink, type: hdfs  
19/04/25 11:06:18 INFO node.AbstractConfigurationProvider: Channel mem_channel connected to [spoolidir_source, hdfs_sink]  
19/04/25 11:06:18 INFO node.Application: Starting new configuration: { sourceRunners:[spoolidir_source=EventDrivenSourceRunner: { source:Spool Directory source  
spoolidir_source: { spoolidir: /home/cloudera/flume_agent1_data } } sinkRunners:[hdfs_sink=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@16  
9d6a08 counterGroup: { name:null counters: { } } ] channels:[mem_channel=org.apache.flume.channel.MemoryChannel(name: mem_channel)] } }  
19/04/25 11:06:18 INFO node.Application: Starting Channel mem_channel  
19/04/25 11:06:18 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mem_channel: Successfully registered new MBean  
19/04/25 11:06:18 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mem_channel started  
19/04/25 11:06:18 INFO node.Application: Starting Sink hdfs_sink  
19/04/25 11:06:18 INFO node.Application: Starting Source spoolidir_source  
19/04/25 11:06:18 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfs_sink: Successfully registered new MBean.  
19/04/25 11:06:18 INFO source.SpoolDirectorySource: SpoolDirectorySource source starting with directory: /home/cloudera/flume_agent1_data  
19/04/25 11:06:18 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfs_sink started  
19/04/25 11:06:18 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: spoolidir_source: Successfully registered new MB  
ean.  
19/04/25 11:06:18 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: spoolidir_source started  
19/04/25 11:06:18 INFO avro.ReliableSpoolingFileEventReader: Last read took us just up to a file boundary. Rolling to the next file, if there is one.  
19/04/25 11:06:18 INFO avro.ReliableSpoolingFileEventReader: Preparing to move file /home/cloudera/flume_agent1_data/test_data.txt to /home/cloudera/flume_ag  
ent1_data/test_data.txt.COMPLETED  
19/04/25 11:06:18 INFO hdfs.HDFSDataStream: SerialIzser = TEXT, UseRawLocalFileSystem = false  
19/04/25 11:06:20 INFO hdfs.BucketWriter: Creating hdfs://quickstart.cloudera:8020/user/cloudera/data/flume/flume_agent1_data/test_data.txt.1556215579853.tmp  
19/04/25 11:06:57 INFO hdfs.BucketWriter: Closing hdfs://quickstart.cloudera:8020/user/cloudera/data/flume/flume_agent1_data/test_data.txt.1556215579853.tmp  
19/04/25 11:06:57 INFO hdfs.BucketWriter: Renaming hdfs://quickstart.cloudera:8020/user/cloudera/data/flume/flume_agent1_data/test_data.txt.1556215579853.tmp  
to hdfs://quickstart.cloudera:8020/user/cloudera/data/flume/flume_agent1_data/test_data.txt.1556215579853  
19/04/25 11:06:57 INFO hdfs.HDFSEventSink: Writer callback called.
```

```
cloudera@quickstart:~/flume_agent1_data
[cloudera@quickstart flume_agent1_data]$ pwd
/home/cloudera/flume_agent1_data
[cloudera@quickstart flume_agent1_data]$ ls -lrt
total 4
-rw-rw-r-- 1 cloudera cloudera 402 Apr 25 10:58 test_data.txt.COMPLETED
[cloudera@quickstart flume_agent1_data]$
```

default

Tables

orders_dtl

Query

Search data and saved documents...

Jobs

cloudera

File Browser

Search for file name

Actions




Move to trash

Upload

New

Home

/ user / cloudera / data / flume / flume_agent1_data

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 .		cloudera	cloudera	drwxr-xr-x	April 25, 2019 10:18 AM
<input type="checkbox"/>	 .		cloudera	cloudera	drwxr-xr-x	April 25, 2019 11:06 AM
<input type="checkbox"/>	 test_data.txt.1556215579853	403 bytes	cloudera	cloudera	-rw-r--r--	April 25, 2019 11:06 AM

Show 45 of 1 items

Page 1 of 1

The screenshot shows the Hue File Browser interface. The top navigation bar includes the Hue logo, a search bar, and a user profile. The left sidebar shows a file tree with 'default' and 'orders_dtl'. The main area displays file details for 'test_data.txt' in the path '/ user / cloudera / data / flume / flume_agent1_data /'. The file is 403 B in size and was last modified on 25/04/2019 at 18:06. The content of the file is a paragraph about Flume.

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

References:

<https://flume.apache.org/releases/content/1.9.0/FlumeDeveloperGuide.html>

<https://flume.apache.org/releases/content/1.9.0/FlumeUserGuide.html>