



Flume Installation Guide

1. Starting flume only requires a configuration file which has the entire configuration for source, sink and channels.
2. We will take a spooling directory as source, HDFS as sink and memory as channel. That means flume will pick up files from a specified directory and put up in HDFS.
3. Configuration file: Create a configuration file named flume-conf.properties and add following configuration in that file.

```
Rashmis-MacBook-Pro:flume rashmi$ cat flume-  
conf.properties  
# example.conf: A single-node Flume configuration  
# Name the components on this agent  
agent.sources = src-1  
agent.sinks = hdfs-sink  
agent.channels = memory-channel  
#Source properties, its a spolling source which will  
take data from directory /Users/rashmi/players/runs  
agent.sources.src-1.type = spooldir  
agent.sources.src-1.spoolDir = /Users/rashmi/players/  
runs
```

```

agent.sources.src-1.fileHeader = true
# Use a channel which buffers events in memory
agent.channels.memory-channel.type = memory
agent.channels.memory-channel.capacity = 1000
agent.channels.memory-channel.transactionCapacity = 100
#Sink properties, hdfs source which will store data
here
agent.sinks.hdfs-sink.type = hdfs
agent.sinks.hdfs-sink.hdfs.path = /user/rashmi/players/
runs
agent.sinks.hdfs-sink.hdfs.fileType = DataStream
agent.sinks.hdfs-sink.hdfs.rollCount = 20
# Bind the source and sink to the channel
agent.sources.src-1.channels = memory-channel
agent.sinks.hdfs-sink.channel = memory-channel

```

4. Create a directory in local. Flume will pick up data from here.

```

Rashmis-MacBook-Pro:~$ mkdir -p /Users/rashmi/players/runs

```

5. Create a directory in HDFS. Flume will put data here.

```

Rashmis-MacBook-Pro:runs$ hadoop fs -mkdir /user/rashmi/players/runs

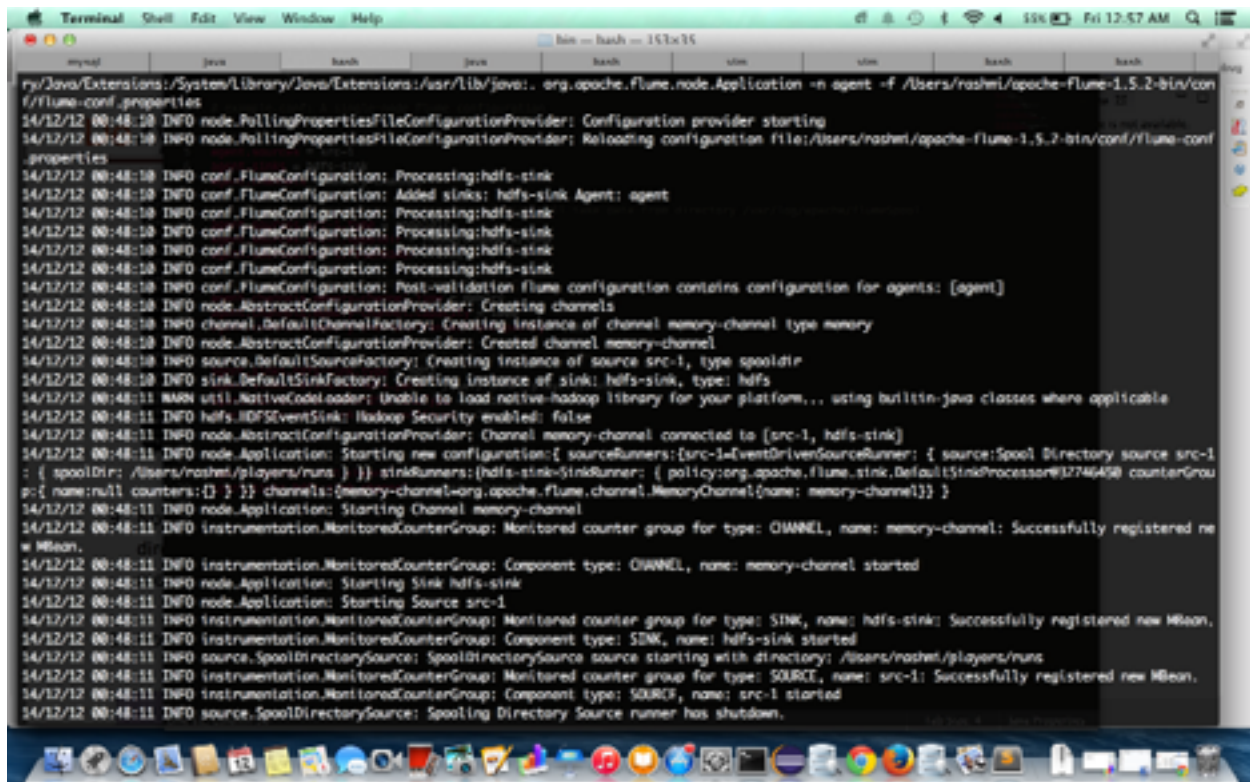
```

6. Now to run flume just run following command. This will start flume after some logging as shown in screenshot below.

```

Rashmis-MacBook-Pro:bin$ ./flume-ng agent -n agent -c conf -f /Users/rashmi/apache-flume-1.5.2-bin/conf/flume-conf.properties
Info: Including Hadoop libraries found via (/usr/local/bin/hadoop) for HDFS access
Info: Excluding /usr/local/Cellar/hadoop/2.6.0/libexec/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/Cellar/hadoop/2.6.0/libexec/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including HBASE libraries found via (/usr/local/bin/hbase) for HBASE access
Info: Excluding /usr/local/Cellar/hbase/0.98.8/libexec/bin/./lib/slf4j-api-1.6.4.jar from classpath
Info: Excluding /usr/local/Cellar/hbase/0.98.8/libexec/bin/./lib/slf4j-log4j12-1.6.4.jar from classpath
Info: Excluding /usr/local/Cellar/hadoop/2.6.0/libexec/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/Cellar/hadoop/2.6.0/libexec/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath

```



```
Terminal Shell Edit View Window Help
bin -- bash -- 153x35
~/Java/Extensions:/System/Library/Java/Extensions:/usr/lib/java: org.apache.flume.node.Application -n agent -f /Users/nashmi/apache-flume-1.5.2-bin/conf/flume-conf.properties
14/12/12 00:48:10 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting
14/12/12 00:48:10 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/Users/nashmi/apache-flume-1.5.2-bin/conf/flume-conf.properties
14/12/12 00:48:10 INFO conf.FlumeConfiguration: Processing:hdfs-sink
14/12/12 00:48:10 INFO conf.FlumeConfiguration: Added sinks: hdfs-sink Agent: agent
14/12/12 00:48:10 INFO conf.FlumeConfiguration: Processing:hdfs-sink
14/12/12 00:48:10 INFO conf.FlumeConfiguration: Processing:hdfs-sink
14/12/12 00:48:10 INFO conf.FlumeConfiguration: Processing:hdfs-sink
14/12/12 00:48:10 INFO conf.FlumeConfiguration: Processing:hdfs-sink
14/12/12 00:48:10 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [agent]
14/12/12 00:48:10 INFO node.AbstractConfigurationProvider: Creating channels
14/12/12 00:48:10 INFO channel.DefaultChannelFactory: Creating instance of channel memory-channel type memory
14/12/12 00:48:10 INFO node.AbstractConfigurationProvider: Created channel memory-channel
14/12/12 00:48:10 INFO source.DefaultSourceFactory: Creating instance of source src-1, type spooldir
14/12/12 00:48:10 INFO sink.DefaultSinkFactory: Creating instance of sink: hdfs-sink, type: hdfs
14/12/12 00:48:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
14/12/12 00:48:11 INFO hdfs.HDFSEventSink: Hadoop Security enabled: false
14/12/12 00:48:11 INFO node.AbstractConfigurationProvider: Channel memory-channel connected to [src-1, hdfs-sink]
14/12/12 00:48:11 INFO node.Application: Starting new configuration:[ sourceRunners:[src-1=EventDrivenSourceRunner: { source:Spool Directory source src-1: { spoolDir: /Users/nashmi/players/runs } } } sinkRunners:[hdfs-sink=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@12746400 counterGroup: { name:null counters:{} } } } channels:[memory-channel=org.apache.flume.channel.MemoryChannel{name: memory-channel}] ]
14/12/12 00:48:11 INFO node.Application: Starting Channel memory-channel
14/12/12 00:48:11 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: memory-channel: Successfully registered new MBean.
14/12/12 00:48:11 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: memory-channel started
14/12/12 00:48:11 INFO node.Application: Starting Sink hdfs-sink
14/12/12 00:48:11 INFO node.Application: Starting Source src-1
14/12/12 00:48:11 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfs-sink: Successfully registered new MBean.
14/12/12 00:48:11 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfs-sink started
14/12/12 00:48:11 INFO source.SpoolDirectorySource: SpoolDirectorySource source starting with directory: /Users/nashmi/players/runs
14/12/12 00:48:11 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: src-1: Successfully registered new MBean.
14/12/12 00:48:11 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: src-1 started
14/12/12 00:48:11 INFO source.SpoolDirectorySource: Spooling Directory Source runner has shutdown.
```

7. Now flume is ready to start its work. Whenever there will be any new file in spooling directory, flume source will pick it up and start putting it in memory channel. Once memory channel is full it will write it to HDFS.

8. Now start putting files in spooling directory.

9. Whenever a new file is copied into spooling directory flume picks it up. See screenshot below.

10. When you check HDFS directory, you will see some temp files and some normal files. Temp files are actually opened by flume and then once it finishes these are renamed to normal files. See screenshot below:

```
Terminal Shell Edit View Window Help
bin -- bash -- 151x35

players/runs/runs_opposition_2011.csv.COMPLETED
14/12/12 00:48:45 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Preparing to move file /Users/rashmi/players/runs/runs_opposition_2012.csv to /Users/rashmi/
players/runs/runs_opposition_2012.csv.COMPLETED
14/12/12 00:48:45 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Preparing to move file /Users/rashmi/players/runs/runs_opposition_2013.csv to /Users/rashmi/
players/runs/runs_opposition_2013.csv.COMPLETED
14/12/12 00:48:45 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Preparing to move file /Users/rashmi/players/runs/runs_opposition_2014.csv to /Users/rashmi/
players/runs/runs_opposition_2014.csv.COMPLETED
14/12/12 00:48:45 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:46 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:46 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:47 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:47 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:48 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:48 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:49 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:49 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:49 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseLocalFileSystem = false
14/12/12 00:48:49 INFO hdfs.BucketWriter: Creating /user/rashmi/players/runs/FlumeData.1418374129848.tmp
14/12/12 00:48:50 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:50 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.
14/12/12 00:48:51 INFO hdfs.BucketWriter: Closing /user/rashmi/players/runs/FlumeData.1418374129848.tmp
14/12/12 00:48:51 INFO hdfs.BucketWriter: Close tries incremented
14/12/12 00:48:51 INFO hdfs.BucketWriter: Renaming /user/rashmi/players/runs/FlumeData.1418374129848.tmp to /user/rashmi/players/runs/FlumeData.141837412
9848
14/12/12 00:48:51 INFO hdfs.BucketWriter: Creating /user/rashmi/players/runs/FlumeData.1418374129849.tmp
14/12/12 00:48:51 INFO hdfs.BucketWriter: Closing /user/rashmi/players/runs/FlumeData.1418374129849.tmp
14/12/12 00:48:51 INFO hdfs.BucketWriter: Close tries incremented
14/12/12 00:48:51 INFO hdfs.BucketWriter: Renaming /user/rashmi/players/runs/FlumeData.1418374129849.tmp to /user/rashmi/players/runs/FlumeData.141837412
9849
14/12/12 00:48:51 INFO hdfs.BucketWriter: Creating /user/rashmi/players/runs/FlumeData.1418374129850.tmp
14/12/12 00:48:51 INFO source.SpillDirectorySource: Spooling Directory Source runner has shutdown.577331 files.
14/12/12 00:48:51 INFO hdfs.BucketWriter: Closing /user/rashmi/players/runs/FlumeData.1418374129850.tmp
14/12/12 00:48:51 INFO hdfs.BucketWriter: Close tries incremented
14/12/12 00:48:51 INFO hdfs.BucketWriter: Renaming /user/rashmi/players/runs/FlumeData.1418374129850.tmp to /user/rashmi/players/runs/FlumeData.141837412
9850
```

```
Rashmi-MacBook-Pro:runs rashmi$ hadoop fs -ls /user/rashmi/players/runs
14/12/12 00:48:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where ap
Found 5 items
-rw-r--r-- 3 rashmi supergroup 741 2014-12-12 00:48 /user/rashmi/players/runs/FlumeData.1418374129848
-rw-r--r-- 3 rashmi supergroup 755 2014-12-12 00:48 /user/rashmi/players/runs/FlumeData.1418374129849
-rw-r--r-- 3 rashmi supergroup 740 2014-12-12 00:48 /user/rashmi/players/runs/FlumeData.1418374129850
-rw-r--r-- 3 rashmi supergroup 754 2014-12-12 00:48 /user/rashmi/players/runs/FlumeData.1418374129851
-rw-r--r-- 3 rashmi supergroup 489 2014-12-12 00:48 /user/rashmi/players/runs/FlumeData.1418374129852.tmp
```