

Configuration Concept

Flume 1.6.0 User Guide — Apache Flume documentation - Mozilla Firefox

Flume 1.6.0 User Guide — Apa...

file

Namesnode HDFS:/tmp/ashu Jobtracker Tasktracker HBase

A Flume event is defined as a unit of data flow having a byte payload and an optional set of string attributes. A Flume agent is a (JVM) process that hosts the components through which events flow from an external source to the next destination (hop).

```
graph LR
    WS((Web Server)) --> S((Source))
    subgraph Agent
        S --> C[(Channel)]
        C --> K((Sink))
    end
    K --> HDFS[(HDFS)]
```

A Flume source consumes events delivered to it by an external source like a web server. The external source sends events to Flume in a format that is recognized by the target Flume source. For example, an Avro Flume source can be used to receive Avro events from Avro clients or other Flume agents in the flow that send events from an Avro sink. A similar flow can be defined using a Thrift Flume Source to receive events from a Thrift Sink or a Flume Thrift Rpc Client or Thrift clients written in any language generated from the Flume thrift protocol. When a Flume source receives an event, it stores it into one or more channels. The channel is a passive store that keeps the event until it's consumed by a Flume sink. The file channel is one example - it is backed by the local filesystem. The sink removes the event from the channel and puts it into an external repository like HDFS (via Flume HDFS sink) or forwards it to the Flume source of the next Flume agent (next hop) in the flow. The source and sink within the given agent run asynchronously with the events staged in the channel.

Complex flows

Sample CDR Files

Text Editor

Ashu

Computer

- Home
- Desktop
- Documents
- Downloads
- Music
- Pictures
- Videos
- File System From HDFS

Ashu

- SpoolDir
- Cusotmer_002.txt
- Customer_001.txt

"Customer_001.txt" selected (7.6 kB)

Terminal

```
warmachine@roxxx:~$
warmachine@roxxx:~$
```

Customer_001.txt (~/.Ashu) - gedit

CDRConfigurator.config Customer_001.txt

```
CustomerIndex|CustomerCount|PatternCDRIndex|CustomerCDRCount|
CustomerPatternDuration|CustomerProfileDuration|Cust_IMSI|Cust_ISDN|
Cust_IMEI|CallType|DateOfCall|TimeOfCall|Duration|1stCorrespType|
1stCorrespISDN|2ndCorrespType|2ndCorrespISDN|PLMN|DataVolumeUp|
DataVolumeDown|CustOperator|CallService|ProfileMarker|PatternMarker
0|1|0|10|900|900|208100000000000|0600000000|350000000000000|MOC|18/07/2014|
00:00:00|90|FRAF2|0610000001|||FRAF2|||FRAF2|Voice|Profile tutorial #1|
Pattern #1 - 10 outgoing voice calls of 1'30" and toward the same corresp.
0|1|1|10|900|900|208100000000000|0600000000|350000000000000|MOC|18/07/2014|
00:23:56|90|FRAF2|0610000001|||FRAF2|||FRAF2|Voice|Profile tutorial #1|
Pattern #1 - 10 outgoing voice calls of 1'30" and toward the same corresp.
0|1|2|10|900|900|208100000000000|0600000000|350000000000000|MOC|18/07/2014|
00:50:41|90|FRAF2|0610000001|||FRAF2|||FRAF2|Voice|Profile tutorial #1|
Pattern #1 - 10 outgoing voice calls of 1'30" and toward the same corresp.
0|1|3|10|900|900|208100000000000|0600000000|350000000000000|MOC|18/07/2014|
01:21:23|90|FRAF2|0610000001|||FRAF2|||FRAF2|Voice|Profile tutorial #1|
Pattern #1 - 10 outgoing voice calls of 1'30" and toward the same corresp.
0|1|4|10|900|900|208100000000000|0600000000|350000000000000|MOC|18/07/2014|
01:42:43|90|FRAF2|0610000001|||FRAF2|||FRAF2|Voice|Profile tutorial #1|
Pattern #1 - 10 outgoing voice calls of 1'30" and toward the same corresp.
0|1|5|10|900|900|208100000000000|0600000000|350000000000000|MOC|18/07/2014|
02:07:59|90|FRAF2|0610000001|||FRAF2|||FRAF2|Voice|Profile tutorial #1|
Pattern #1 - 10 outgoing voice calls of 1'30" and toward the same corresp.
```

Plain Text Tab Width: 8 Ln 6, Col 39 INS

SpoolDir Configuration

The screenshot shows the Apache Flume 1.6.0 User Guide in a Mozilla Firefox browser. The page is titled "Flume 1.6.0 User Guide — Apache Flume documentation - Mozilla Firefox". The address bar shows "file:///". The page content includes a list of properties for the SpoolDir source, a table of properties, and a search bar at the bottom.

1. If a file is written to after being placed into the spooling directory, Flume will print an error to its log file and stop processing.

2. If a file name is reused at a later time, Flume will print an error to its log file and stop processing.

To avoid the above issues, it may be useful to add a unique identifier (such as a timestamp) to log file names when they are moved into the spooling directory.

Despite the reliability guarantees of this source, there are still cases in which events may be duplicated if certain downstream failures occur. This is consistent with the guarantees offered by other Flume components.

Property Name	Default	Description
channels	-	
type	-	The component type name, needs to be <code>spoolDir</code> .
spoolDir	-	The directory from which to read files from.
fileSuffix	<code>.COMPLETED</code>	Suffix to append to completely ingested files
deletePolicy	<code>never</code>	When to delete completed files: <code>never</code> or <code>immediate</code>
fileHeader	<code>false</code>	Whether to add a header storing the absolute path filename.
fileHeaderKey	<code>file</code>	Header key to use when appending absolute path filename to event header.
basenameHeader	<code>false</code>	Whether to add a header storing the basename of the file.
basenameHeaderKey	<code>basename</code>	Header Key to use when appending basename of file to event header.
ignorePattern	<code>^\$</code>	Regular expression specifying which files to ignore (skip)
trackerDir	<code>.flumespool</code>	Directory to store metadata related to processing of files. If this path is not an absolute path, then it is interpreted as relative to the <code>spoolDir</code> .
consumeOrder	<code>oldest</code>	In which order files in the spooling directory will be consumed <code>oldest</code> , <code>youngest</code> and <code>random</code> .

Find: Previous Next Highlight all Match case

Configuration Information

The screenshot shows a text editor window titled "CDRConfigurator.config (~/.workspace/Sample/Resources) - gedit". The file contains configuration for the CDR source and sink. The configuration is as follows:

```
cdr.sources = src
cdr.channels = ch1
cdr.sinks = snk

cdr.sources.src.type = spoolDir
cdr.sources.src.basenameHeader = true
cdr.sources.src.spoolDir = /SpoolDir

cdr.channels.ch1.type = memory

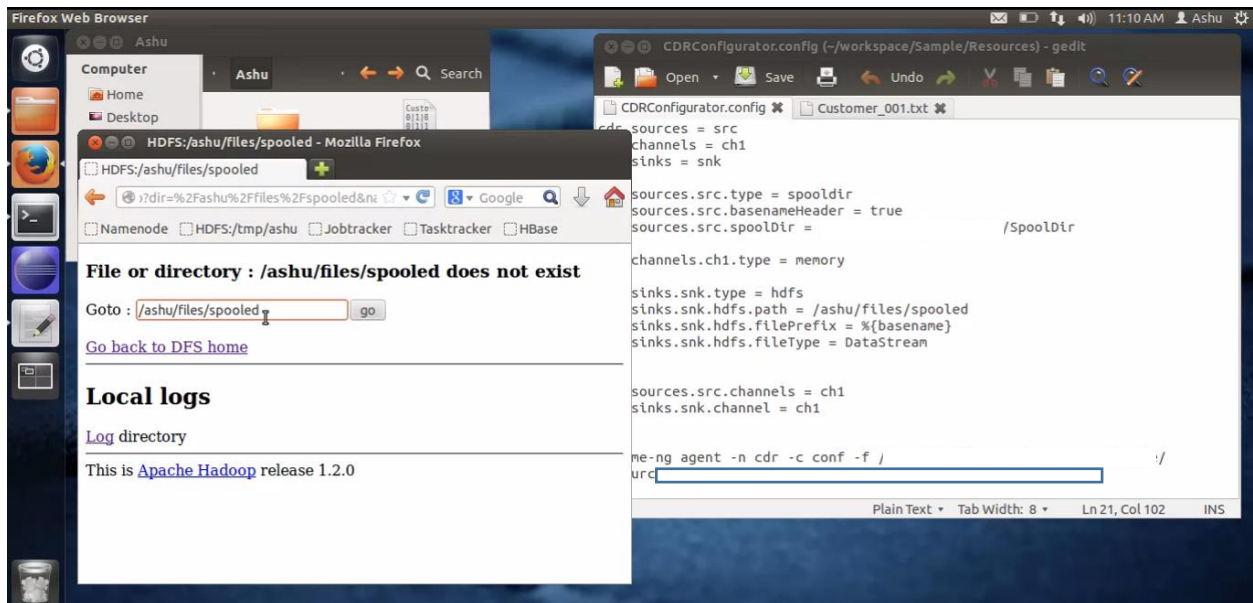
cdr.sinks.snk.type = hdfs
cdr.sinks.snk.hdfs.path = /ashu/files/spooled
cdr.sinks.snk.hdfs.filePrefix = ${basename}
cdr.sinks.snk.hdfs.fileType = DataStream

cdr.sources.src.channels = ch1
cdr.sinks.snk.channel = ch1

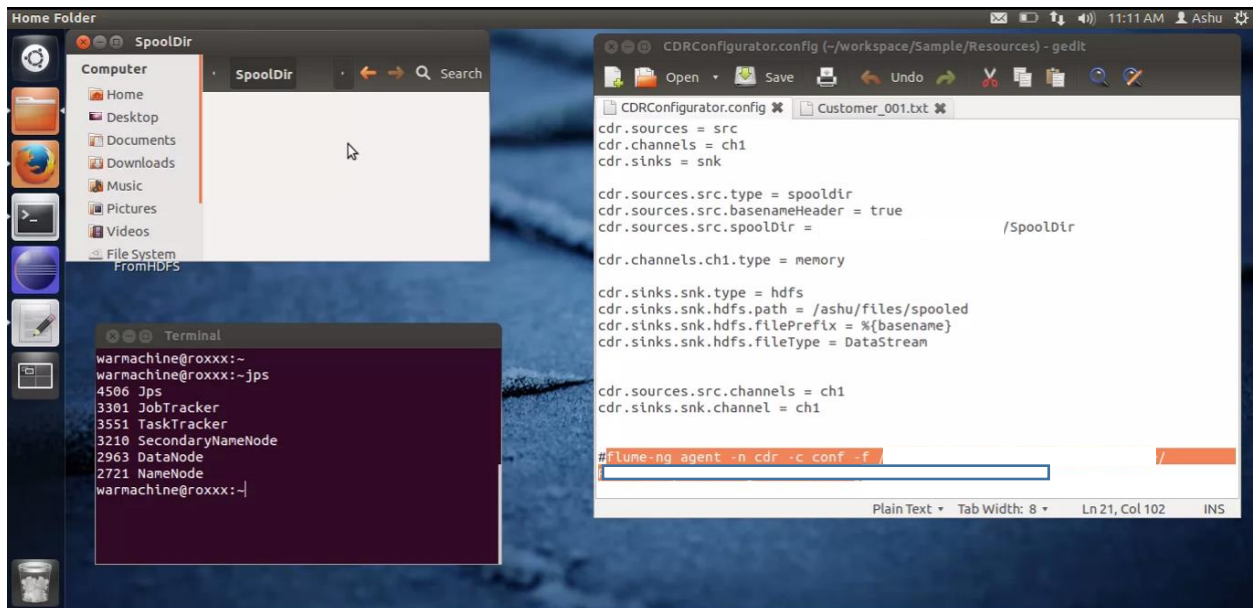
#flume-ng agent -n cdr -c conf -f
```

Find: Previous Next Highlight all Match case

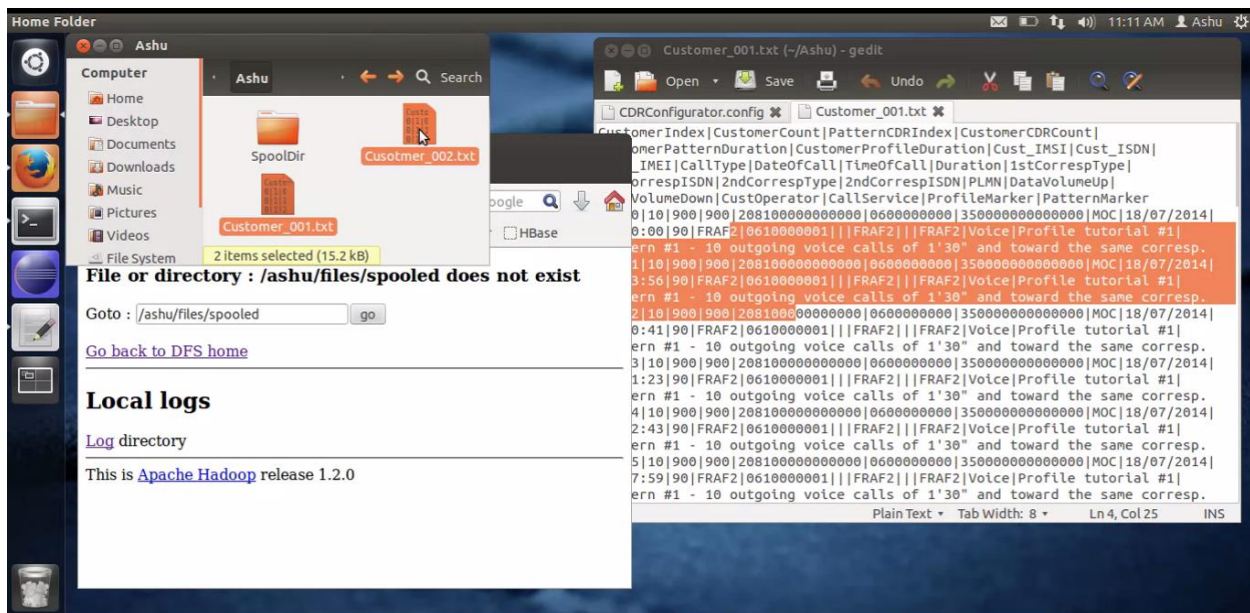
HDFS Directory Path



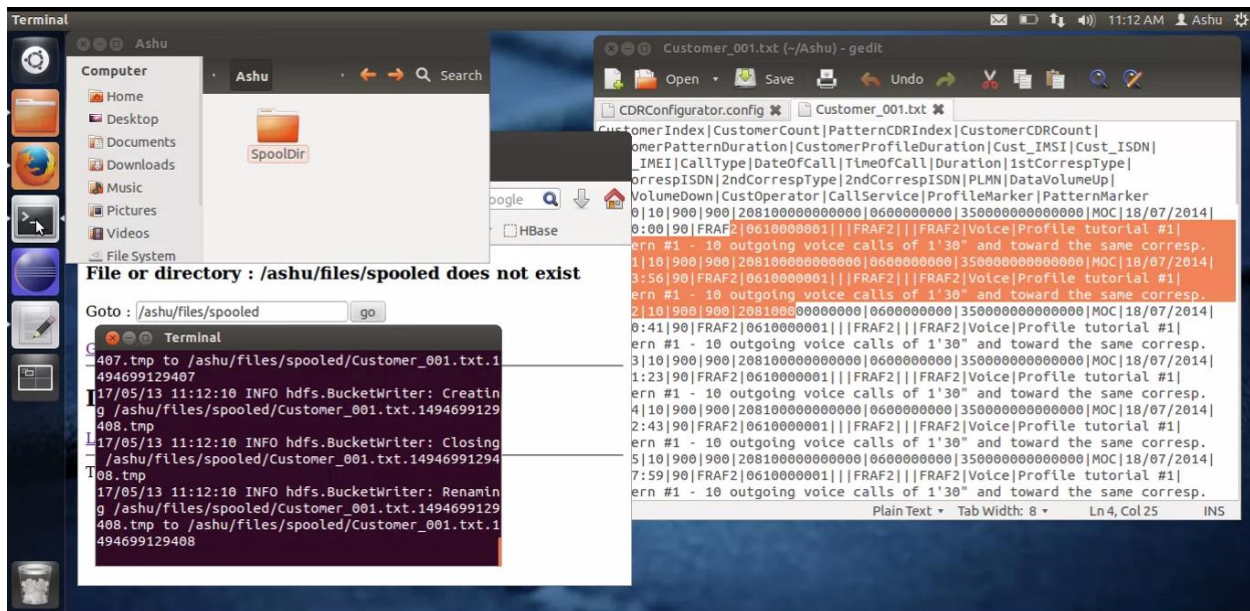
Empty SpoolDir Initially



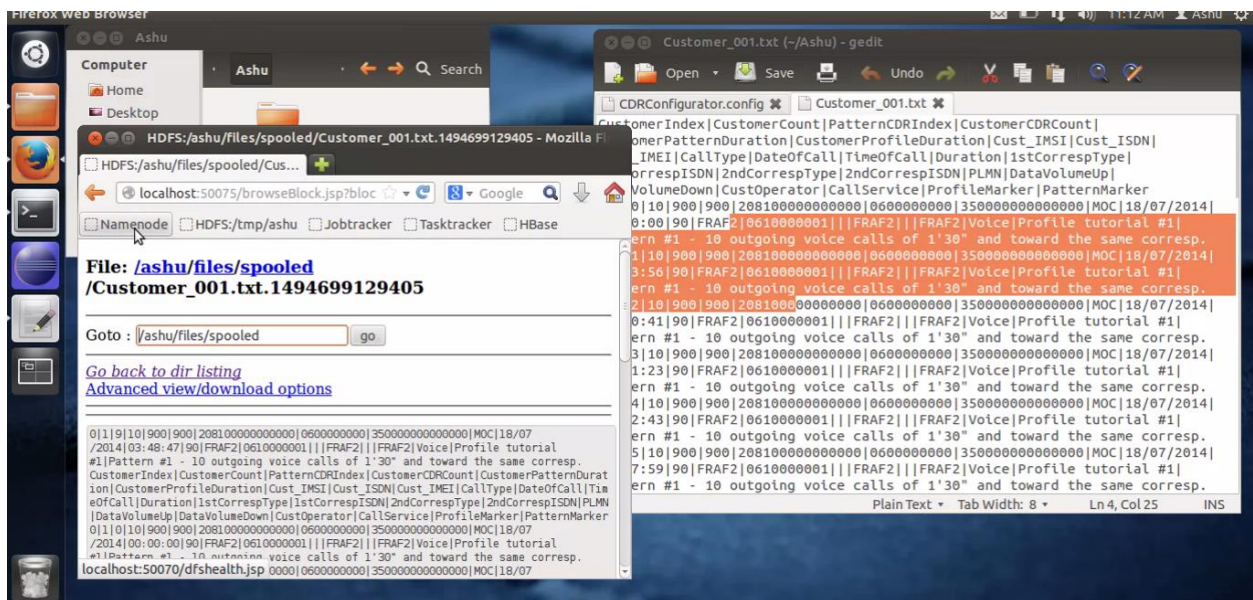
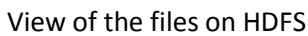
Pushing the CDR files to SpoolDir



Jobs Triggered Automatically



Files Exported to HDFS



Files Configured with .completed extension after the job is finished

The screenshot displays a Linux desktop environment with a dark theme. The top panel shows the system clock at 11:12 AM and the user name 'Ashu'. The left sidebar contains icons for the Home Folder, Computer, and File System. The main window is titled 'SpoolDir' and shows a directory view of '/ashu/files/spooled'. Two files are visible: 'Cusotmer_002.txt.COMPLETED' and 'Customer_001.txt.COMPLETED'. A status bar at the bottom of the window indicates 'Contents of directory /ashu/files/spooled'. Below this, a table lists the files with their names, types, sizes, and replication counts.

Name	Type	Size	Replication	B	S
Cusotmer_002.txt.1494699126669	file	1.18 KB	1	64	M
Cusotmer_002.txt.1494699126670	file	1.08 KB	1	64	M
Cusotmer_002.txt.1494699126671	file	1.18 KB	1	64	M
Cusotmer_002.txt.1494699126672	file	1.08 KB	1	64	M

The right side of the screen shows a text editor window titled 'Customer_001.txt'. The editor contains a large block of text, which appears to be a log or configuration file. The text is partially obscured by a redacted area, but it includes fields like 'CustomerIndex', 'CustomerCount', 'PatternCDRIndex', 'CustomerCDRCount', 'IMEI', 'CallType', 'DateOfCall', 'TimeOfCall', 'Duration', '1stCorrespType', '2ndCorrespType', '2ndCorrespISDN', 'PLMN', 'DataVolumeUp', 'VolumeDown', 'CustOperator', 'CallService', 'ProfileMarker', 'PatternMarker', and 'MOC'. The text is formatted with vertical bars separating the fields.