

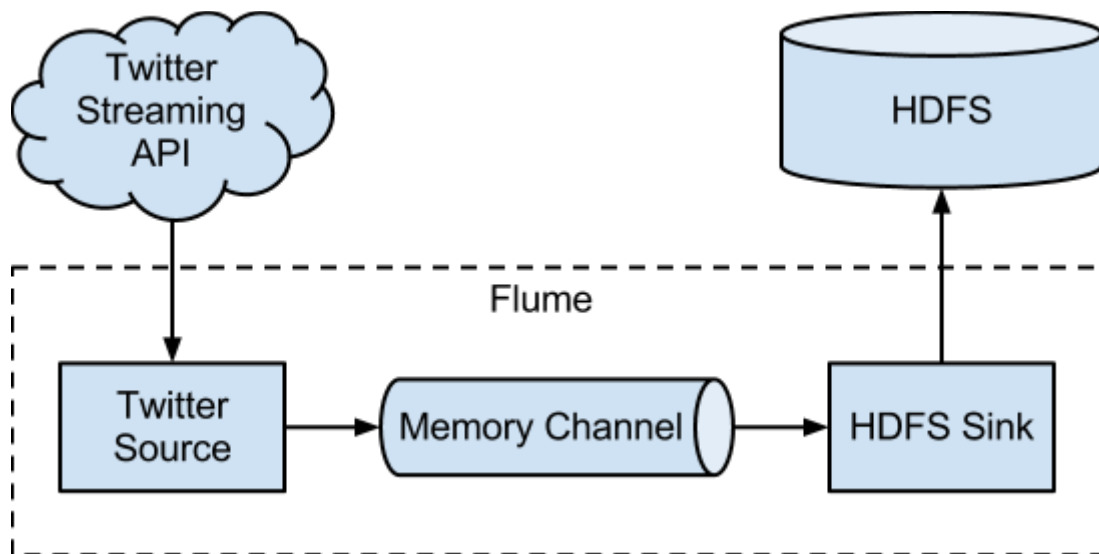
How to configure apache flume in Cloudlab to fetch Twitter Data

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data.

Our use case:

Collect data from the Twitter Streaming API, and forward it to HDFS.

Event	Accessing Twitter API
Sources	Twitter
Channel	Memory Channel
Sink	HDFS
Agent	Twitter Agent



Prerequisites:

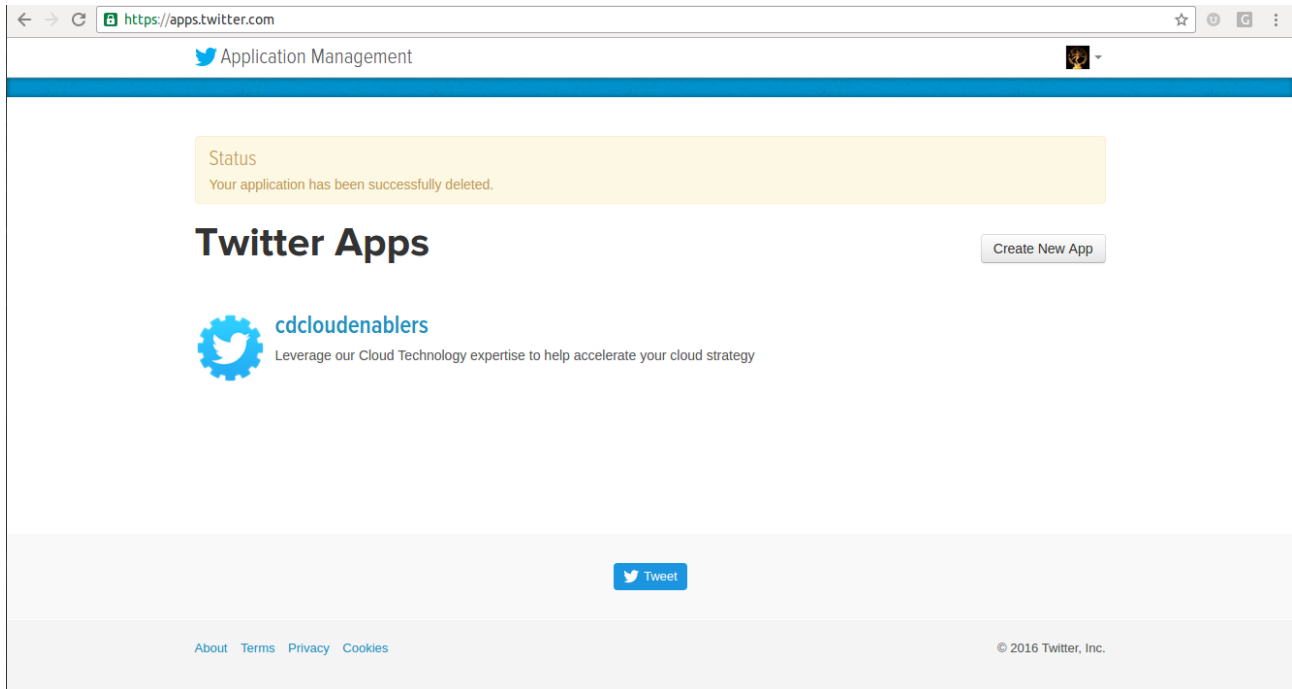
1. Cloudlab webconsole access.
2. Twitter account to create an application in twitter application.

Step1:

Log in to your twitter account and create an application.

<https://apps.twitter.com/>

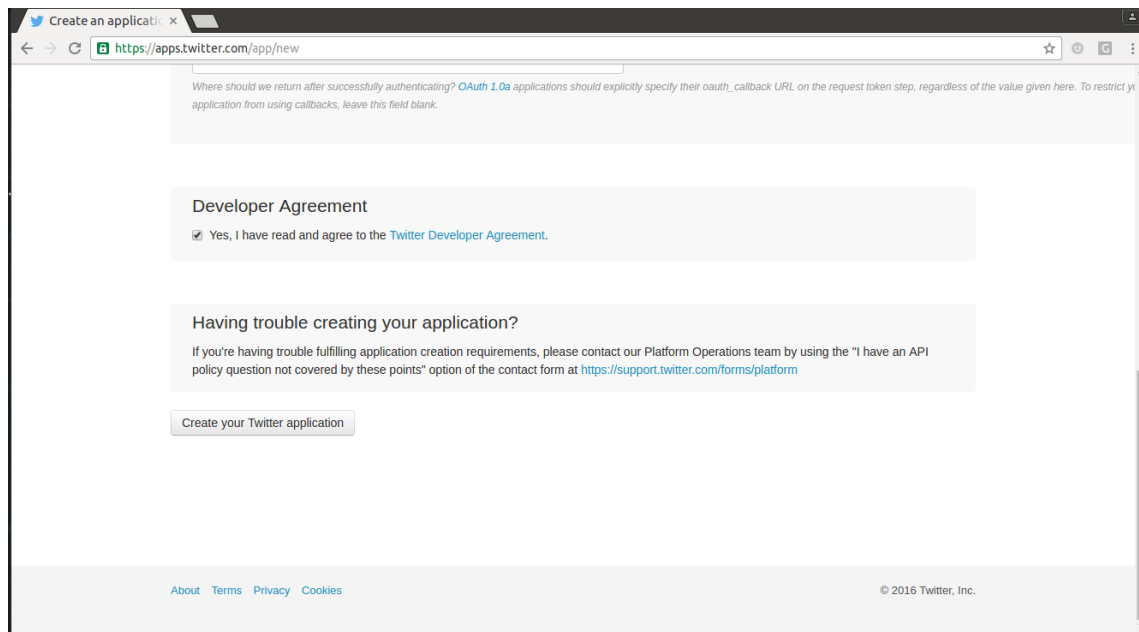
Select “Create New App”



Assign name to your application, Add description and your website details.

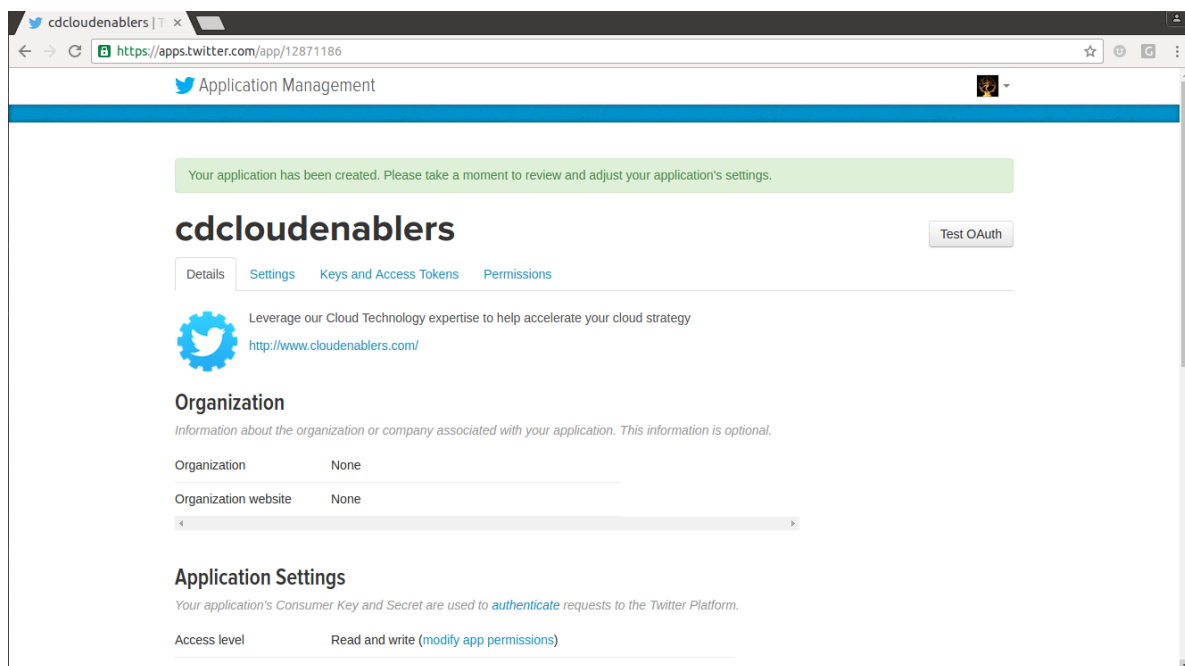
A screenshot of the 'Create an application' form on the Twitter Application Management page. The form is titled 'Create an application' and is located under the 'Application Management' header. It contains several input fields and labels: 'Name' with the value 'cloudenablers', 'Description' with the value 'Leverage our Cloud Technology expertise to help accelerate your cloud strategy', 'Website' with the value 'http://www.cloudenablers.com/', and 'Callback URL' which is currently empty. Each input field has a small asterisk indicating it is required. Below the 'Callback URL' field, there is a note: 'Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict y...'. The form is set against a light gray background with a white border.

Accept the agreement and proceed to create your application.



The screenshot shows the 'Create an application' page on the Twitter developer portal. The browser address bar displays 'https://apps.twitter.com/app/new'. A text box at the top explains the OAuth 1.0a callback URL requirement. Below this is the 'Developer Agreement' section with a checked checkbox 'Yes, I have read and agree to the Twitter Developer Agreement.' A help section titled 'Having trouble creating your application?' provides contact information for the Platform Operations team. A 'Create your Twitter application' button is located at the bottom of the form area. The footer contains links for 'About', 'Terms', 'Privacy', and 'Cookies', along with the copyright notice '© 2016 Twitter, Inc.'

Get consumer key, consumer secret, access token and access token secret for adding these details to get twitter access log from bigdata lab server.



The screenshot shows the 'Application Management' page for an application named 'cdcloudenablers'. The browser address bar displays 'https://apps.twitter.com/app/12871186'. A green notification banner states: 'Your application has been created. Please take a moment to review and adjust your application's settings.' The page has a blue header with the Twitter logo and the text 'Application Management'. Below the notification, the application name 'cdcloudenablers' is displayed, followed by a 'Test OAuth' button. A tabbed interface shows 'Details', 'Settings', 'Keys and Access Tokens', and 'Permissions', with 'Settings' currently selected. The 'Settings' section includes a Twitter logo, the text 'Leverage our Cloud Technology expertise to help accelerate your cloud strategy', and the URL 'http://www.cloudenablers.com/'. Under the 'Organization' heading, there are fields for 'Organization' (set to 'None') and 'Organization website' (set to 'None'). The 'Application Settings' section follows, with a note about Consumer Key and Secret usage. The 'Access level' is set to 'Read and write (modify app permissions)'. The bottom of the page shows the beginning of the 'Consumer Key (API Key)' and 'Consumer Secret (API Secret)' fields.

Step2:

Log in to cloudlab webconsole with access details provided by your course instructor.

<https://labs.simplilearn.com/big-data/webconsole>

Check the current directory

```
$pwd
```

Create a new folder to keep your flume configuration file.

```
$mkdir -p /home/sabapathy/flume/conf
```

```
$cd /home/sabapathy/flume/conf
```

```
$touch flume.conf
```

Copy the configuration file available in the following link to flume.conf file.

<https://s3.amazonaws.com/simplilearnlab/flume.conf>

Modify the configuration files according to your user name and credentials.

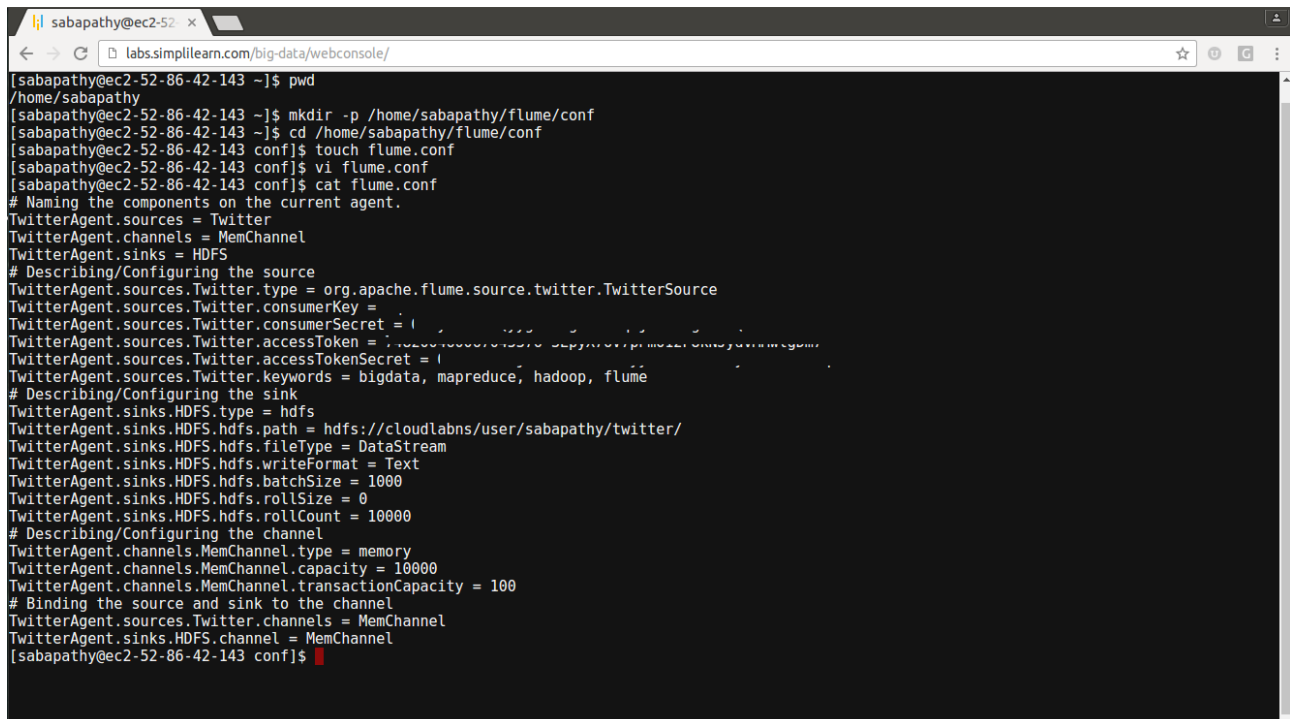
```
TwitterAgent.sources.Twitter.consumerKey = CONSUMER KEY
```

```
TwitterAgent.sources.Twitter.consumerSecret = CONSUMER SECRET
```

```
TwitterAgent.sources.Twitter.accessToken = ACCESS TOKEN
```

```
TwitterAgent.sources.Twitter.accessTokenSecret = TOKEN SECRET
```

```
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://cloudlabns/user/USERNAME/twitter/
```



```
sabapathy@ec2-52-86-42-143 ~]$ pwd
/home/sabapathy
[sabapathy@ec2-52-86-42-143 ~]$ mkdir -p /home/sabapathy/flume/conf
[sabapathy@ec2-52-86-42-143 ~]$ cd /home/sabapathy/flume/conf
[sabapathy@ec2-52-86-42-143 conf]$ touch flume.conf
[sabapathy@ec2-52-86-42-143 conf]$ vi flume.conf
[sabapathy@ec2-52-86-42-143 conf]$ cat flume.conf
# Naming the components on the current agent.
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey = 
TwitterAgent.sources.Twitter.consumerSecret = 
TwitterAgent.sources.Twitter.accessToken = 
TwitterAgent.sources.Twitter.accessTokenSecret = 
TwitterAgent.sources.Twitter.keywords = bigdata, mapreduce, hadoop, flume
# Describing/Configuring the sink
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://cloudlabns/user/sabapathy/twitter/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
# Describing/Configuring the channel
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
# Binding the source and sink to the channel
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
[sabapathy@ec2-52-86-42-143 conf]$
```

Create a new folder in hdfs to get twitter access log

\$hadoop fs -mkdir /user/sabapathy/twitter

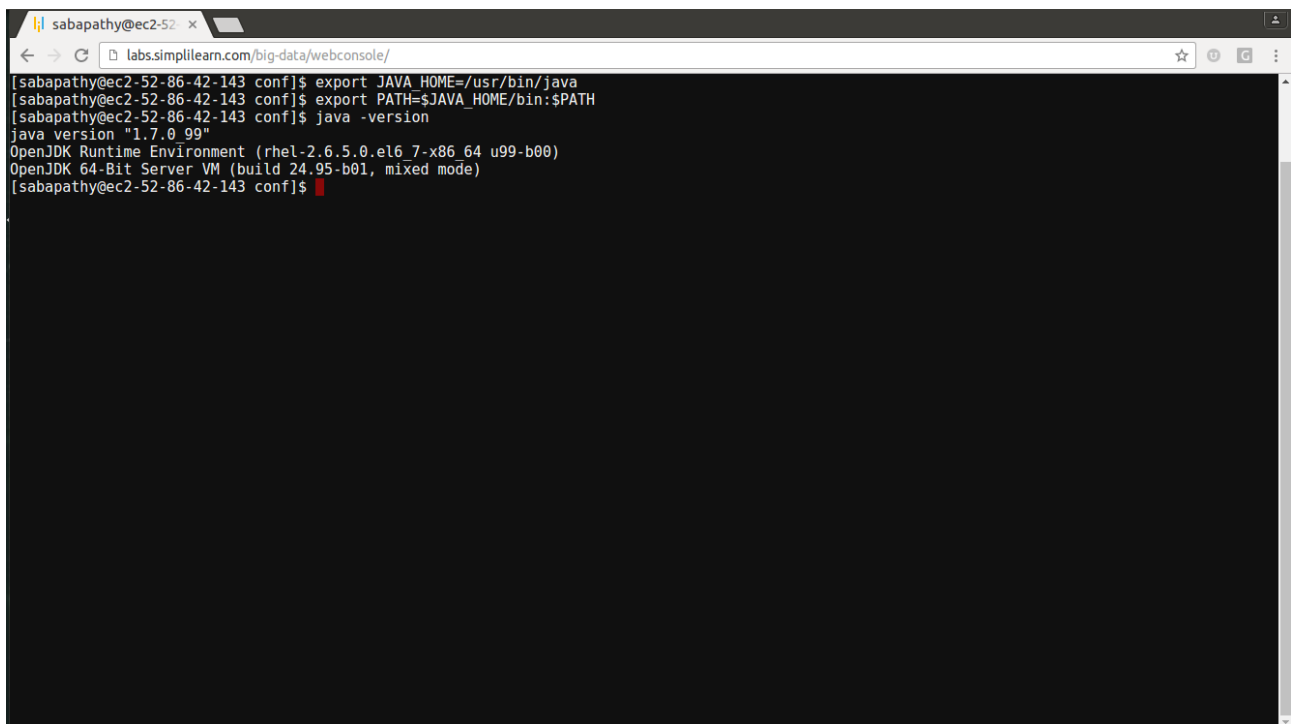
Before starting the flume agent make sure java home is set in your shell environment by issuing the command given below.

java -version

If java home is not set before, export the two environment variables to your shell or your ~/.bash_profile

export JAVA_HOME=/usr/bin/java

export PATH=\$JAVA_HOME/bin:\$PATH

A screenshot of a web browser window displaying a terminal session. The browser's address bar shows 'labs.simplilearn.com/big-data/webconsole/'. The terminal output shows the following commands and results:









```
[sabapathy@ec2-52-86-42-143 conf]$ export JAVA_HOME=/usr/bin/java
[sabapathy@ec2-52-86-42-143 conf]$ export PATH=$JAVA_HOME/bin:$PATH
[sabapathy@ec2-52-86-42-143 conf]$ java -version
java version "1.7.0_99"
OpenJDK Runtime Environment (rhel-2.6.5.0.el6_7-x86_64 u99-b08)
OpenJDK 64-Bit Server VM (build 24.95-b01, mixed mode)
[sabapathy@ec2-52-86-42-143 conf]$
```

Run flume agent with your configuration file

\$flume-ng agent --conf /home/sabapathy/flume/conf -f /home/sabapathy/flume/conf/flume.conf
Dflume.root.logger=DEBUG,console -n TwitterAgent

```
labs.simplilearn.com/big-data/webconsole/
16/09/22 13:06:49 INFO conf.FlumeConfiguration: Processing:HDFS
16/09/22 13:06:49 INFO conf.FlumeConfiguration: Processing:HDFS
16/09/22 13:06:49 INFO conf.FlumeConfiguration: Processing:HDFS
16/09/22 13:06:50 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [TwitterAgent]
16/09/22 13:06:50 INFO node.AbstractConfigurationProvider: Creating channels
16/09/22 13:06:50 INFO channel.DefaultChannelFactory: Creating instance of channel MemChannel type memory
16/09/22 13:06:50 INFO node.AbstractConfigurationProvider: Created channel MemChannel
16/09/22 13:06:50 INFO source.DefaultSourceFactory: Creating instance of source Twitter, type org.apache.flume.source.twitter.TwitterSource
16/09/22 13:06:50 INFO twitter.TwitterSource: Consumer Key: 'fq0xMFsN11570AikFhnB05C8f'
16/09/22 13:06:50 INFO twitter.TwitterSource: Consumer Secret: '017y8PEAcE0yygB0wngkkuIqfjP96DGgdcZEQLfWheYm1BRra'
16/09/22 13:06:50 INFO twitter.TwitterSource: Access Token: '748200468067045376-3EpyX76V7oPm012F8kNjydvMwlgDm7'
16/09/22 13:06:50 INFO twitter.TwitterSource: Access Token Secret: '0cu17cnOLjTTb9T2hVkunoytVcc5C1Nb7y0YTtYtnCcq'
16/09/22 13:06:50 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs
16/09/22 13:06:50 INFO hdfs.HDFSEventSink: Hadoop Security enabled: false
16/09/22 13:06:50 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]
16/09/22 13:06:51 INFO node.Application: Starting new configuration: { sourceRunners: {Twitter-EventDrivenSourceRunner: { source:org.apache.flume.s
source.twitter.TwitterSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS-SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessorg52a8f44
9 counterGroup:{ name:null counters:{ } }} channels:{MemChannel-org.apache.flume.channel.MemoryChannel{name: MemChannel}} }
16/09/22 13:06:51 INFO node.Application: Starting Channel MemChannel
16/09/22 13:06:51 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Successfully registered new MBean.
16/09/22 13:06:51 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
16/09/22 13:06:51 INFO node.Application: Starting Sink HDFS
16/09/22 13:06:51 INFO node.Application: Starting Source Twitter
16/09/22 13:06:51 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} ...
16/09/22 13:06:51 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully registered new MBean.
16/09/22 13:06:51 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
16/09/22 13:06:51 INFO twitter.TwitterSource: Twitter source Twitter started.
16/09/22 13:06:51 INFO twitter4j.TwitterStreamImpl: Establishing connection.
16/09/22 13:06:52 INFO twitter4j.TwitterStreamImpl: Connection established.
16/09/22 13:06:52 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
16/09/22 13:06:53 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
16/09/22 13:06:53 INFO hdfs.BucketWriter: Creating hdfs://cloudlabs/user/sabapathy/twitter//FlumeData.1474549613307.tmp
16/09/22 13:06:55 INFO twitter.TwitterSource: Processed 100 docs
16/09/22 13:06:56 INFO twitter.TwitterSource: Processed 200 docs
```

Check the files generated in hdfs sink (Example: /home/sabapathy/twitter)



sabapathy

File Browser

Rename

Move

Copy

Change Permissions

Download

Delete

New

Upload

Home / user / sabapathy / twitter

Trash

Type	Name	Size	User	Group	Permissions	Date
Folder	..		sabapathy	hadoop	drwxr-xr-x	September 22, 2016 06:08 AM
Folder	..		sabapathy	hadoop	drwxr-xr-x	September 22, 2016 06:06 AM
File	FlumeData.1474549613307	719.8 KB	sabapathy	hadoop	-rw-r--r--	September 22, 2016 06:07 AM
File	FlumeData.1474549645487	675.9 KB	sabapathy	hadoop	-rw-r--r--	September 22, 2016 06:07 AM
File	FlumeData.1474549676363	703.4 KB	sabapathy	hadoop	-rw-r--r--	September 22, 2016 06:08 AM
File	FlumeData.1474549707294.tmp	0 bytes	sabapathy	hadoop	-rw-r--r--	September 22, 2016 06:08 AM

Show 45 items per page. Showing 1 to 4 of 4 items, page 1 of 1.

Home

/ user / sabapathy / twitter / FlumeData.1474549613307

ACTIONS

View As Binary

Download

View File Location

Refresh

INFO

Last Modified
Sept 22, 2016
6:07 a.m.

User
sabapathy

Group
hadoop

Size
719.8 KB

Mode
100644

First BlockPrevious BlockNext BlockLast Block

Viewing Bytes: 1 - 4096 of 737053 (4096 B block size)

Warning: some binary data has been masked out with '�'.

```
Obj      avro.schema{
  { "type": "record", "name": "Doc", "doc": "adoc", "fields": [ { "name": "id", "type": "string", { "name": "user_friends_count", "type": [ "int", "nul
1" ] }, { "name": "user_location", "type": [ "string", "null" ] }, { "name": "user_description", "type": [ "string", "null" ] }, { "name": "user_statuses_c
ount", "type": [ "int", "null" ] }, { "name": "user_followers_count", "type": [ "int", "null" ] }, { "name": "user_name", "type": [ "string", "null" ] }, { "n
ame": "user_screen_name", "type": [ "string", "null" ] }, { "name": "created_at", "type": [ "string", "null" ] }, { "name": "text", "type": [ "string", "nu
ll" ] }, { "name": "retweet_count", "type": [ "long", "null" ] }, { "name": "retweeted", "type": [ "boolean", "null" ] }, { "name": "in_reply_to_user_i
d", "type": [ "long", "null" ] }, { "name": "source", "type": [ "string", "null" ] }, { "name": "in_reply_to_status_id", "type": [ "long", "null" ] }, { "nam
e": "media_url_https", "type": [ "string", "null" ] }, { "name": "expanded_url", "type": [ "string", "null" ] } ] }
    ,Magdeburg, Deutschland{ Franz1 W.FranziW6(2016-09-22T13:06:43ZRT @heuteshow: Saarlands AfD-Spitzenkandidat verkauft in se
inem Antiquitätenshop #Hakenkreuz. Er wusste angeblich nicht, dass das noch ni
        <a href="http://twitter.com/
download/iphone" rel="nofollow">Twitter for iPhone</a>
Obj      avro.schema{
  { "type": "record", "name": "Doc", "doc": "adoc", "fields": [ { "name": "id", "type": "string", { "name": "user_friends_count", "type": [ "int", "nul
1" ] }, { "name": "user_location", "type": [ "string", "null" ] }, { "name": "user_description", "type": [ "string", "null" ] }, { "name": "user_statuses_c
ount", "type": [ "int", "null" ] }, { "name": "user_followers_count", "type": [ "int", "null" ] }, { "name": "user_name", "type": [ "string", "null" ] }, { "n
ame": "user_screen_name", "type": [ "string", "null" ] }, { "name": "created_at", "type": [ "string", "null" ] }, { "name": "text", "type": [ "string", "nu
ll" ] }, { "name": "retweet_count", "type": [ "long", "null" ] }, { "name": "retweeted", "type": [ "boolean", "null" ] }, { "name": "in_reply_to_user_i
d", "type": [ "long", "null" ] }, { "name": "source", "type": [ "string", "null" ] }, { "name": "in_reply_to_status_id", "type": [ "long", "null" ] }, { "nam
e": "media_url_https", "type": [ "string", "null" ] }, { "name": "expanded_url", "type": [ "string", "null" ] } ] }
    Beatriz Durãesbaliittlecandy(2016-09-22T13:06:43ZFilme da minha vida
        <a href="//twitter.com/download/android" rel="nofollow">Twitter for Android</a>
$778943611295657985Hollywood, FL0of
for Twitter (@allthingsloveof HARDY BUTTER online no 95/92/45 https://t.co/gKtWmFz7 https://t.co/iOWfEz77
```

After verifying this operation, Stop the process and remove the logs generated in hdfs.