

**BUAN 6346.501**

**Bonus Project - Flume Twitter**

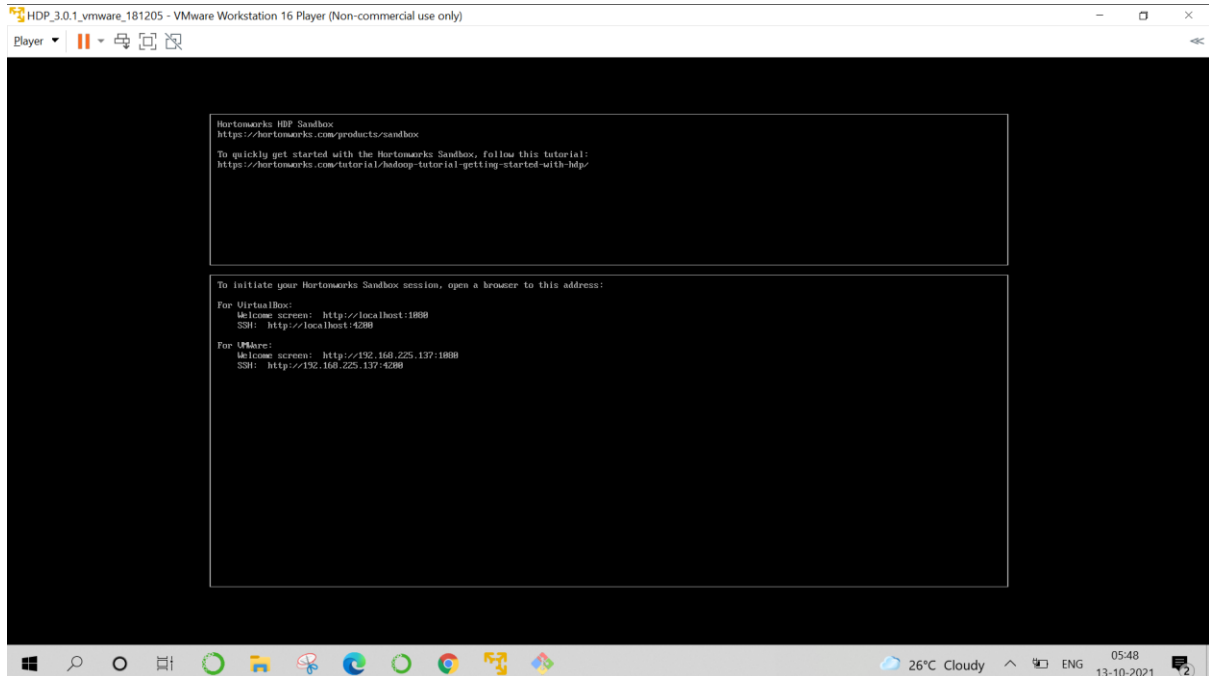
**Arul Chakravarthy**

## Twitter-Flume Integration

### Twitter-Flume Integration

**Step 1:** New Twitter Developer Account was created using <https://developer.twitter.com/>

**Step 2:** For Hadoop ecosystem, Sandbox HDP Horton VMWare V3.0.1 was downloaded and installed from Cloudera



From above, my Sandbox IP is 192.168.225.137

**Step 3:** For Windows, Hosts.cfg has been configured with above HDP Sandbox IP and its host name mapping as follows

```
192.168.225.137 localhost sandbox-hdp.hortonworks.com sandbox-hdf.hortonworks.com
```

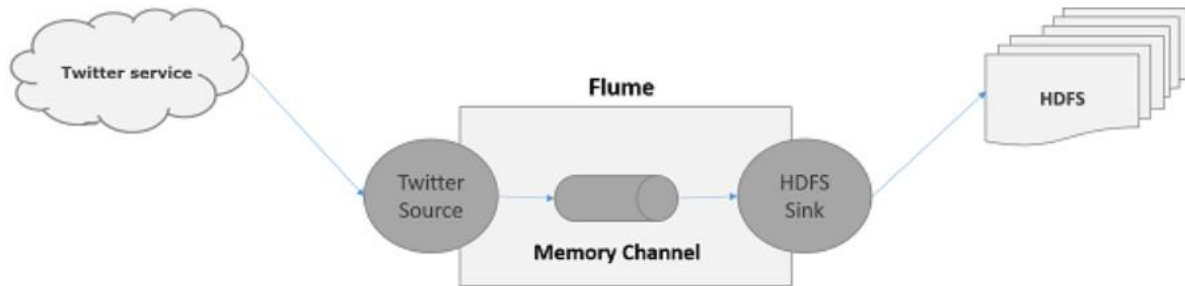
**Hosts.cfg configuration:**

```
# Copyright (c) 1993-2009 Microsoft Corp.
#
# This is a sample HOSTS file used by Microsoft TCP/IP for Windows.
#
# This file contains the mappings of IP addresses to host names. Each
# entry should be kept on an individual line. The IP address should
# be placed in the first column followed by the corresponding host name.
# The IP address and the host name should be separated by at least one
# space.
#
# Additionally, comments (such as these) may be inserted on individual
# lines or following the machine name denoted by a '#' symbol.
#
# For example:
#
#      102.54.94.97      rhino.acme.com          # source server
#      38.25.63.10     x.acme.com              # x client host
#
# localhost name resolution is handled within DNS itself.
#      127.0.0.1        localhost
#      ::1              localhost
192.168.225.137 localhost sandbox-hdp.hortonworks.com sandbox-hdf.hortonworks.com
```

## Twitter-Flume Integration

Let's see in this document on how we can set-up Flume to read the tweets into HDFS

**Destination HDFS Directory:** /loudacre/tweets



**Step 4:** For Windows, Git Bash was used throughout to login into HDP-Sandbox by SSH as root user via port 2222

```
arulc@LAPTOP-70QSL9Q7 MINGW64 ~  
$ ssh root@sandbox-hdp.hortonworks.com -p 2222  
root@sandbox-hdp.hortonworks.com's password:  
You are required to change your password immediately (root enforced)  
Last login: Wed Oct 13 22:33:18 2021 from 172.18.0.3  
Changing password for root.  
(current) UNIX password:  
New password:  
Retype new password:
```

**Step 4:** A simple check whether Hadoop is installed

```
[root@sandbox-hdp ~]# hadoop version  
Hadoop 3.1.1.3.0.1.0-187  
Source code repository git@github.com:hortonworks/hadoop.git -r 2820e4d6fc7ec31ac42187083ed5933c823e9784  
Compiled by jenkins on 2018-09-19T10:19Z  
Compiled with protoc 2.5.0  
From source with checksum 889327faf5a6ca5fc06fcf97c13af29  
This command was run using /usr/hdp/3.0.1.0-187/hadoop/hadoop-common-3.1.1.3.0.1.0-187.jar
```

**Step 5:** Latest version of Apache-Flume was downloaded from the below link in my local laptop <https://downloads.apache.org/flume/1.7.0/> . From local laptop, apache-flume zip file has been transferred into HDP Sandbox using SCP

```
arulc@LAPTOP-70QSL9Q7 MINGW64 ~/Downloads  
$ scp -P 2222 apache-flume-1.7.0-bin.tar.gz root@sandbox-hdp.hortonworks.com:/root  
root@sandbox-hdp.hortonworks.com's password:  
apache-flume-1.7.0-bin.tar.gz 100% 53MB 77.4MB/s 00:00
```

```
[root@sandbox-hdp ~]# ls  
anaconda-ks.cfg apache-flume-1.7.0-bin.tar.gz
```

**Step 6:** Apache-flume zip file has been unzipped as follows using sudo access

```
[root@sandbox-hdp ~]# sudo tar xzf apache-flume-1.7.0-bin.tar.gz  
[root@sandbox-hdp ~]# |
```

## Twitter-Flume Integration

**Step 7:** Unzipped apache-flume was moved into /usr/local/flume/ of my sandbox

```
[root@sandbox-hdp ~]# sudo mv apache-flume-1.7.0-bin /usr/local/flume/  
[root@sandbox-hdp ~]#
```

**Step 8:** Necessary environment variables has been set as FLUME\_PATH and the bin directory of flume has been added to the PATH variable which will be used in the further steps

```
[root@sandbox-hdp bin]# export FLUME_HOME=/usr/local/flume/apache-flume-1.7.0-bin/  
[root@sandbox-hdp bin]# export PATH=$PATH:$FLUME_HOME/bin  
[root@sandbox-hdp bin]#
```

**Step 9:** Check for flume-ng version

```
[root@sandbox-hdp bin]# flume-ng version  
Flume 1.7.0  
Source code repository: https://git-wip-us.apache.org/repos/asf/flume.git  
Revision: 511d868555dd4d16e6ce4fedc72c2d1454546707  
Compiled by bessbd on Wed Oct 12 20:51:10 CEST 2016  
From source with checksum 0d21b3ffdc55a07e1d08875872c00523
```

**Step 10:** Once Flume installation was done, Twitter application was created using the following settings

### Edit App details

#### App name

Maximum length 32 characters

2

#### App icon



[Upload](#)

Maximum size of 700k, JPG, GIF, PNG

#### Description

Briefly describe your App.

Between 10 and 200 characters

94

## Twitter-Flume Integration



### App permissions

[Edit](#)

Read and Write

Read + Post Tweets and profile information



### Authentication settings

[Edit](#)

3-legged OAuth is **disabled**



Use 3-legged OAuth for Sign in with Twitter, posting Tweets on behalf of other accounts and more. Get more information in the [docs](#).

**Step11:** Destination directory has been created as /loudacre/tweets in Sandbox HDFS to store the tweets as follows

```
[root@sandbox-hdp ~]# sudo -u hdfs hadoop fs -mkdir -p /loudacre/tweets
[root@sandbox-hdp ~]#
```

Changed the HDFS destination directory owner, as root for root user to write tweets into /loudacre/tweets

```
[root@sandbox-hdp conf]# sudo -u hdfs hadoop fs -chown root /loudacre/tweets
```

**Step 12:** flume-env.sh configuration file of Apache Flume was edited using flume-env.sh.template pre-created during installation as reference and added JAVA\_PATH and CLASSPATH as environment variables

```
[root@sandbox-hdp conf]# cd $FLUME_HOME/conf/
[root@sandbox-hdp conf]# cp flume-env.sh.template flume-env.sh
[root@sandbox-hdp conf]# vi flume-env.sh
```

## Flume-env.sh

**Step13:** Under conf folder of apache-flume a new configuration file has been created as twitter.conf with necessary Flume Source, sink and channel configurations

```

arulc@LAPTOP-70QSL9Q7 MINGW64 ~
$ ssh root@sandbox-hdp.hortonworks.com -p 2222
root@sandbox-hdp.hortonworks.com's password:
Last login: Thu Oct 14 00:22:30 2021 from 172.18.0.3
[root@sandbox-hdp ~]# cat /usr/local/flume/apache-flume-1.7.0-bin/conf/twitter.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 = sfGuz
TwitterAgent.sources.Twitter.consumerSecret = M3sty
TwitterAgent.sources.Twitter.accessToken = 1320
TwitterAgent.sources.Twitter.accessTokenSecret = FR0nt9C
TwitterAgent.sources.Twitter.keywords = hadoop,hive, bigdata, mapreduce, sqoop, noase, pig

# Describing/Configuring the sink
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = /loudacre/tweets/
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.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

```

Arul Chakravarthy



## Twitter-Flume Integration

**Step14:** Flume Agent was run using the below command with parameters as `-n TwitterAgent` and `-conf-file <path-to-twitter.conf-file>` as follows to read the tweets from twitter

```
flume-ng agent -name TwitterAgent --conf-file $FLUME_PATH/conf/twitter.conf
```

```
[root@sandbox-hdp conf]# flume-ng agent -name TwitterAgent --conf-file $FLUME_PATH/conf/twitter.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Warning: JAVA_HOME is not set!
Info: Including Hadoop libraries found via (/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/bin/hbase) for HBASE access
Info: Including Hive libraries found via () for Hive access
+ exec /bin/java -Xmx20m -cp "/usr/local/flume/apache-flume-1.7.0-bin/lib/*:/usr/hdp/3.0.1.0-187/hadoop/conf:/usr/hdp/3.0.1.0-187/hadoop/lib/*:/usr/hdp/3.0.1.0-187/hadoop/./:/usr/hdp/3.0.1.0-187/hadoop-hdfs/./:/usr/hdp/3.0.1.0-187/hadoop-hdfs/lib/*:/usr/hdp/3.0.1.0-187/hadoop-hdfs/./:/usr/hdp/3.0.1.0-187/hadoop-mapreduce/lib/*:/usr/hdp/3.0.1.0-187/hadoop-mapreduce/./:/usr/hdp/3.0.1.0-187/hadoop-yarn/./:/usr/hdp/3.0.1.0-187/tez/lib/*:/usr/hdp/3.0.1.0-187/tez/conf:/usr/hdp/3.0.1.0-187/tez/conf/lib/*:/usr/hdp/3.0.1.0-187/tez/doc:/usr/hdp/3.0.1.0-187/tez/hadoop-shim-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/hadoop-shim-2.8.0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/*:/usr/hdp/3.0.1.0-187/tez/man:/usr/hdp/3.0.1.0-187/tez/tez-api-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-common-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-dag-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-examples-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-history-parser-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-javac-tools-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-job-analyzer-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-mapreduce-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-protobuf-history-plugin-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-runtime-internals-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-runtime-library-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-tests-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-yarn-timeline-cache-plugin-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-yarn-timeline-history-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-yarn-timeline-history-with-acls-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez-yarn-timeline-history-with-fs-0.9.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/tez/ut:/usr/hdp/3.0.1.0-187/tez/lib/commons-io-2.4.jar:/usr/hdp/3.0.1.0-187/tez/lib/commons-lang-2.6.jar:/usr/hdp/3.0.1.0-187/tez/lib/commons-codec-1.4.jar:/usr/hdp/3.0.1.0-187/tez/lib/commons-collections-3.2.2.jar:/usr/hdp/3.0.1.0-187/tez/lib/commons-collections4-4.1.jar:/usr/hdp/3.0.1.0-187/tez/lib/commons-math3-3.1.1.jar:/usr/hdp/3.0.1.0-187/tez/lib/gcs-connector-1.9.0.3.0.1.0-187-shaded.jar:/usr/hdp/3.0.1.0-187/tez/lib/guava-11.0.2.jar:/usr/hdp/3.0.1.0-187/tez/lib/hadoop-aws-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/hadoop-azure-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/hadoop-azure-datalake-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/hadoop-hdfs-client-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/hadoop-mapreduce-client-common-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/hadoop-mapreduce-client-core-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/hadoop-mapreduce-yarn-server-timeline-plugin-storage-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/tez/lib/jersey-client-1.19.jar:/usr/hdp/3.0.1.0-187/tez/lib/jersey-json-1.19.jar:/usr/hdp/3.0.1.0-187/tez/lib/jettison-1.3.4.jar:/usr/hdp/3.0.1.0-187/tez/lib/jetty-server-9.3.22.v20171030.jar:/usr/hdp/3.0.1.0-187/tez/lib/jetty-util-9.3.22.v20171030.jar:/usr/hdp/3.0.1.0-187/tez/lib/jsr305-3.0.0.jar:/usr/hdp/3.0.1.0-187/tez/lib/metrics-core-3.1.0.jar:/usr/hdp/3.0.1.0-187/tez/lib/protobuf-java-2.5.0.jar:/usr/hdp/3.0.1.0-187/tez/lib/roaringbitmap-0.4.9.jar:/usr/hdp/3.0.1.0-187/tez/lib/servlet-api-2.5.jar:/usr/hdp/3.0.1.0-187/tez/lib/slf4j-api-1.7.10.jar:/usr/hdp/3.0.1.0-187/tez/lib/tez.tar.gz:/usr/hdp/3.0.1.0-187/hbase/conf:/usr/hdp/3.0.1.0-187/hbase/lib/tools.jar:/usr/hdp/3.0.1.0-187/hbase/lib/accessors-smart-1.2.jar:/usr/hdp/3.0.1.0-187/hbase/lib/aopalliance-1.0.jar:/usr/hdp/3.0.1.0-187/hbase/lib/aopalliance-repackaged-2.5.0-b32.jar:/usr/hdp/3.0.1.0-187/hbase/lib/asm-5.0.4.jar:/usr/hdp/3.0.1.0-187/hbase/lib/atlas-plugin-classloader-1.0.0.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/hbase/lib/audience-annotations-0.5.0.jar:/usr/hdp
```

TwitterAgent was started as below.,

```
21/10/14 00:22:16 INFO node.Application: Starting Source Twitter
21/10/14 00:22:16 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name=Twitter,state=IDLE} ...
21/10/14 00:22:16 INFO twitter.TwitterSource: Twitter source Twitter started.
21/10/14 00:22:16 INFO twitter4j.TwitterStreamImpl: Establishing connection.
21/10/14 00:22:17 INFO twitter4j.TwitterStreamImpl: Connection established.
21/10/14 00:22:17 INFO twitter4j.TwitterStreamImpl: Receiving stream.
21/10/14 00:22:17 INFO hdfs.HDFSDataStream: Serializer = TEXT, UserRawLocalFileSystem = false
21/10/14 00:22:18 INFO hdfs.BucketWriter: Creating /loudacre/tweets//FlumeData.1634170937877.tmp
21/10/14 00:22:19 INFO twitter.TwitterSource: Processed 100 docs
21/10/14 00:22:22 INFO twitter.TwitterSource: Processed 200 docs
21/10/14 00:22:25 INFO twitter.TwitterSource: Processed 300 docs
21/10/14 00:22:28 INFO twitter.TwitterSource: Processed 400 docs
21/10/14 00:22:31 INFO twitter.TwitterSource: Processed 500 docs
21/10/14 00:22:34 INFO twitter.TwitterSource: Processed 600 docs
21/10/14 00:22:36 INFO twitter.TwitterSource: Processed 700 docs
21/10/14 00:22:39 INFO twitter.TwitterSource: Processed 800 docs
21/10/14 00:22:42 INFO twitter.TwitterSource: Processed 900 docs
21/10/14 00:22:45 INFO twitter.TwitterSource: Processed 1,000 docs
21/10/14 00:22:45 INFO twitter.TwitterSource: Total docs indexed: 1,000, total skipped docs: 0
21/10/14 00:22:45 INFO twitter.TwitterSource: 35 docs/second
21/10/14 00:22:45 INFO twitter.TwitterSource: Run took 28 seconds and processed:
21/10/14 00:22:45 INFO twitter.TwitterSource: 0.01 MB/sec sent to index
21/10/14 00:22:45 INFO twitter.TwitterSource: 0.268 MB text sent to index
21/10/14 00:22:45 INFO twitter.TwitterSource: There were 0 exceptions ignored:
21/10/14 00:22:47 INFO twitter.TwitterSource: Processed 1,100 docs
21/10/14 00:22:50 INFO hdfs.BucketWriter: Closing /loudacre/tweets//FlumeData.1634170937877.tmp
21/10/14 00:22:50 INFO hdfs.BucketWriter: Renaming /loudacre/tweets/FlumeData.1634170937877.tmp to /loudacre/tweets/FlumeData.1634170937877
```

**Step15:** Successfully Flume has ingested Tweets from Twitter into HDFS /loudacre/tweets directory as follows

```
[root@sandbox-hdp ~]# hdfs dfs -ls /loudacre/tweets/
Found 9 items
-rw-r--r-- 1 root hdfs 504429 2021-10-14 00:22 /loudacre/tweets/FlumeData.1634170937877
-rw-r--r-- 1 root hdfs 480563 2021-10-14 00:23 /loudacre/tweets/FlumeData.1634170970748
-rw-r--r-- 1 root hdfs 545926 2021-10-14 00:23 /loudacre/tweets/FlumeData.1634171001770
-rw-r--r-- 1 root hdfs 6864542 2021-10-14 00:24 /loudacre/tweets/FlumeData.1634171032781
-rw-r--r-- 1 root hdfs 10074552 2021-10-14 00:24 /loudacre/tweets/FlumeData.1634171065990
-rw-r--r-- 1 root hdfs 10074552 2021-10-14 00:25 /loudacre/tweets/FlumeData.1634171096190
-rw-r--r-- 1 root hdfs 10074552 2021-10-14 00:25 /loudacre/tweets/FlumeData.1634171126386
-rw-r--r-- 1 root hdfs 10074552 2021-10-14 00:26 /loudacre/tweets/FlumeData.1634171156754
-rw-r--r-- 1 root hdfs 5037276 2021-10-14 00:26 /loudacre/tweets/FlumeData.1634171187080
[root@sandbox-hdp ~]# |
```

**Step16:** Finally, a sanity check to display the contents of tweets from above HDFS destination directory using `hdfs dfs -cat /loudacre/tweets/FlumeData.1634170937877 | head`

\*\*\* **END** \*\*\*