

Kalyan Big Data Projects – Project 2

How To Stream Twitter Data Into Hadoop in JSON format Using Apache Flume

Pre-Requisites of Flume Project:

hadoop-2.6.0
flume-1.6.0
java-1.7

NOTE: Make sure that install all the above components

Flume Project Download Links:

`hadoop-2.6.0.tar.gz` ==> [link](https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-2.6.0.tar.gz)
(<https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-2.6.0.tar.gz>)

`apache-flume-1.6.0-bin.tar.gz` ==> [link](https://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz)
(<https://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz>)

`kalyan-twitter-hdfs-agent.conf` ==> [link](https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/flume/project2-twitter-hadoop-json/kalyan-twitter-hdfs-agent.conf)
(<https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/flume/project2-twitter-hadoop-json/kalyan-twitter-hdfs-agent.conf>)

`kalyan-flume-project-0.1.jar` ==> [link](https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-flume-project-0.1.jar)
(<https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-flume-project-0.1.jar>)

Learnings of this Project:

- We will learn Flume Configurations and Commands
 - Flume Agent
 1. Source (Twitter Source)
 2. Channel (Memory Channel)
 3. Sink (Hdfs Sink)
 - Major project in Real Time `Social Media (Twitter) Sentiment Analysis`
 1. We are extracting the data from twitter using twitter api credentials
 2. This data will be useful to do sentiment analysis on twitter tweets
 3. JSON is the output format
 - We can use hive / pig / mapreduce to analyze this data
 1. explore hive query to analysis
 2. explore pig scripts to analysis
 3. explore mapreduce to analysis
-

1. create "**kalyan-twitter-hdfs-agent.conf**" file with below content

```
agent.sources = Twitter
agent.channels = MemChannel
agent.sinks = HDFS

agent.sources.Twitter.type = com.orientit.kalyan.flume.source.KalyanTwitterSource
agent.sources.Twitter.channels = MemChannel
agent.sources.Twitter.consumerKey = *****
agent.sources.Twitter.consumerSecret = *****
agent.sources.Twitter.accessToken = *****
agent.sources.Twitter.accessTokenSecret = *****
agent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data
scientiest, business intelligence, mapreduce, data warehouse, data warehousing, mahout, hbase,
nosql, newsql, businessintelligence, cloudcomputing

agent.sinks.HDFS.type = hdfs
agent.sinks.HDFS.channel = MemChannel
agent.sinks.HDFS.hdfs.path = hdfs://localhost:8020/user/flume/tweets
agent.sinks.HDFS.hdfs.fileType = DataStream
agent.sinks.HDFS.hdfs.writeFormat = Text
agent.sinks.HDFS.hdfs.batchSize = 100
agent.sinks.HDFS.hdfs.rollSize = 0
agent.sinks.HDFS.hdfs.rollCount = 100
agent.sinks.HDFS.hdfs.useLocalTimeStamp = true

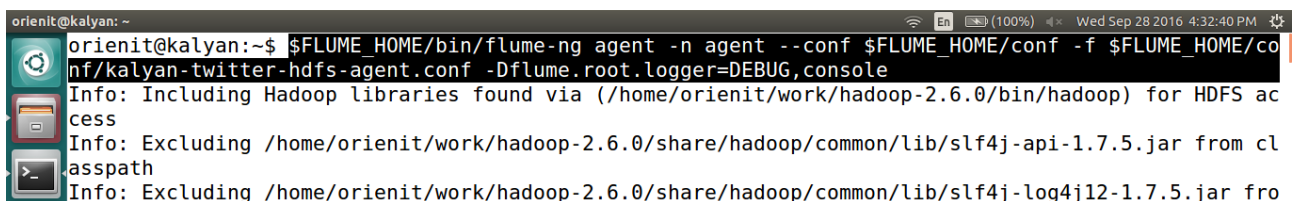
agent.channels.MemChannel.type = memory
agent.channels.MemChannel.capacity = 1000
agent.channels.MemChannel.transactionCapacity = 100
```

2. Copy "**kalyan-twitter-hdfs-agent.conf**" file into "**\$FUME_HOME/conf**" folder

3. Copy "**kalyan-flume-project-0.1.jar**" file into "**\$FLUME_HOME/lib**" folder

4. Execute the below command to **Extract data from Twitter into Hadoop using Flume`**

```
$FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/co
nf/kalyan-twitter-hdfs-agent.conf -Dflume.root.logger=DEBUG,console
```



```
orientit@kalyan: ~
orientit@kalyan:~$ $FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/co
nf/kalyan-twitter-hdfs-agent.conf -Dflume.root.logger=DEBUG,console
Info: Including Hadoop libraries found via (/home/orientit/work/hadoop-2.6.0/bin/hadoop) for HDFS ac
cess
Info: Excluding /home/orientit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from cl
asspath
Info: Excluding /home/orientit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar fro
```

5. Verify the data in console

```

gger.debug(SLF4JLogger.java:75)] Received:{"filter_level":"low","retweeted":false,"in_reply_to_scre
en_name":null,"possibly_sensitive":false,"truncated":false,"lang":"en","in_reply_to_status_id_str":
null,"id":781086360635580416,"in_reply_to_user_id_str":null,"timestamp_ms":"1475060474915","in_repl
y_to_status_id":null,"created_at":"Wed Sep 28 11:01:14 +0000 2016","favorite_count":0,"place":null,
"coordinates":null,"text":"Customer Journey Analytics Workforce Analytics Workforce Analytics Know
ledge Resources Passed On From Mavens https://t.co/qa05aU9zhG","contributors":null,"geo":null,"enti
ties":{"symbols":[],"urls":[{"expanded_url":"http://buff.ly/2cZNUUQ","indices":[110,133],"display_u
rl":"buff.ly/2cZNUUQ","url":"https://t.co/qa05aU9zhG"}],"hashtags":[],"user_mentions":[]},"is_quote
_status":false,"source":"<a href='\"http://bufferapp.com\"' rel='\"nofollow\">Buffer</a>","favorited"
:false,"in_reply_to_user_id":null,"retweet_count":0,"id_str":"781086360635580416","user":{"location
":null,"default_profile":true,"profile_background_tile":false,"statuses_count":95414,"lang":"en","p
rofile_link_color":"0084B4","id":571553244,"following":null,"protected":false,"favourites_count":0,
"profile_text_color":"333333","verified":false,"description":"Technical issues of information syste
ms and applications."},"contributors_enabled":false,"profile_sidebar_border_color":"C0C0C0","name":

```

6. Verify the data in hdfs location is "hdfs://localhost:8020/user/flume/tweets"

HDFS:/user/flume/tweets - Google Chrome

localhost:50075/browseDirectory.jsp?dir=%2Fuser%2Fflume%2Ftweets&namenodeinfoPort=50070&nnaddr=127.0.0.1:8020

Contents of directory /user/flume/tweets

Goto: /user/flume/tweets go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1475060467729	file	173.07 KB	1	128 MB	2016-09-28 16:31	rw-r--r--	orientit	supergroup
FlumeData.1475060499869.tmp	file	13.68 KB	1	128 MB	2016-09-28 16:31	rw-r--r--	orientit	supergroup

Go back to DFS home

HDFS:/user/flume/tweets/FlumeData.1475060467729 - Google Chrome

localhost:50075/browseBlock.jsp?blockId=1073741829&blockSize=177225&genStamp=1099511628781&filename=%2Fuser%2Fflume%2Ftweets%2FFlumeData.1475

File: /user/flume/tweets/FlumeData.1475060467729

Goto: /user/flume/tweets go

Go back to dir listing
Advanced view/download options
View Next chunk

```

{"filter_level":"low","retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive":false,"truncated":false,"lang":"en","in_reply_to_status_id_str":
null,"id":781086360635580416,"in_reply_to_user_id_str":null,"timestamp_ms":"1475060474915","in_repl
y_to_status_id":null,"created_at":"Wed Sep 28 11:01:14 +0000 2016","favorite_count":0,"place":null,
"coordinates":null,"text":"Customer Journey Analytics Workforce Analytics Workforce Analytics Know
ledge Resources Passed On From Mavens https://t.co/qa05aU9zhG","contributors":null,"geo":null,"enti
ties":{"symbols":[],"urls":[{"expanded_url":"http://buff.ly/2cZNUUQ","indices":[110,133],"display_u
rl":"buff.ly/2cZNUUQ","url":"https://t.co/qa05aU9zhG"}],"hashtags":[],"user_mentions":[]},"is_quote
_status":false,"source":"<a href='\"http://bufferapp.com\"' rel='\"nofollow\">Buffer</a>","favorited"
:false,"in_reply_to_user_id":null,"retweet_count":0,"id_str":"781086360635580416","user":{"location
":null,"default_profile":true,"profile_background_tile":false,"statuses_count":95414,"lang":"en","p
rofile_link_color":"0084B4","id":571553244,"following":null,"protected":false,"favourites_count":0,
"profile_text_color":"333333","verified":false,"description":"Technical issues of information syste
ms and applications."},"contributors_enabled":false,"profile_sidebar_border_color":"C0C0C0","name":

```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh