BIG DATA HADOOP AND SPARK DEVLOPMENT

ASSIGNMENT 12

Table of Contents:

BIG DATA HADOOPAND SPARK DEVELOPMENT

1. Introduction

In this assignment, the given tasks are performed and Output of the tasks are recorded in the form of Screenshots.

2. Objective

This Assignment consolidates the deeper understanding of the Session – 12 Introduction to Oozie and Flume

3. Prerequisite

To stream data to our database from twitter we should have the following

pre-requisites.

- Twitter account
- Hadoop cluster

4. Problem Statement

- ### Task 1

Create a flume agent that streams data from Twitter and stores in the HDFS.

5. Expected Output

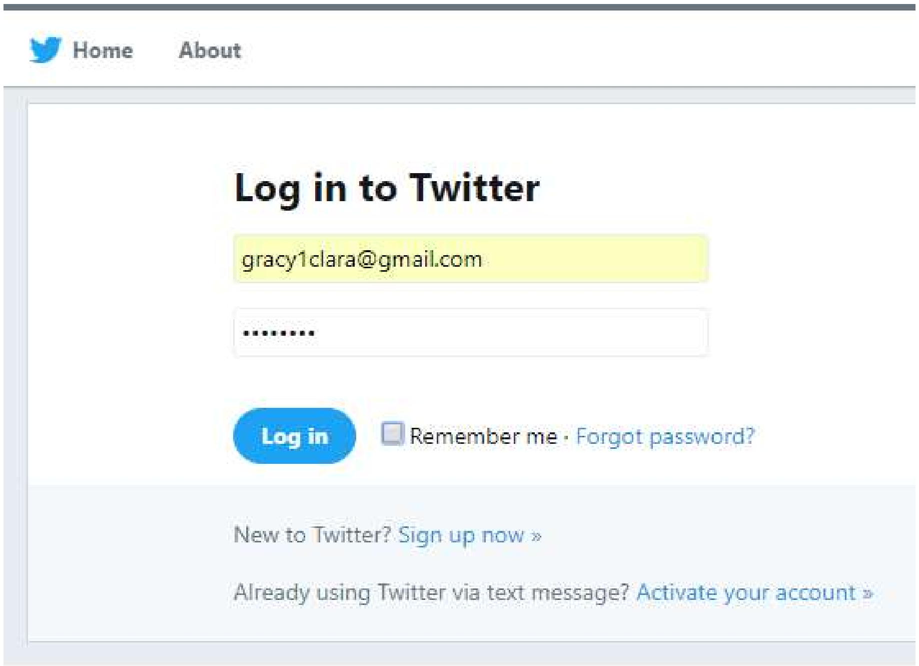
- ## Task 1

Create a flume agent that streams data from Twitter and stores in the HDFS.

Below are the steps to be followed to create a Flume agent to stream Twitter data into HDFS: -

1. Login to Twitter account.
2. Create a new app in this url - https://apps.twitter.com/app
3. Accept the terms and conditions and proceed further.
4. From “Keys and Access Token” tab get the Consumer Key, Consumer Secret, Access Key and Access Secret and hit on “Create my access token”.
5. Now update the keys and access tokens attained in Step4 in the configuration file as mentioned in the Screen shot.

Login to the twitter account



Log in to Twitter

gracy1clara@gmail.com

•••••••••

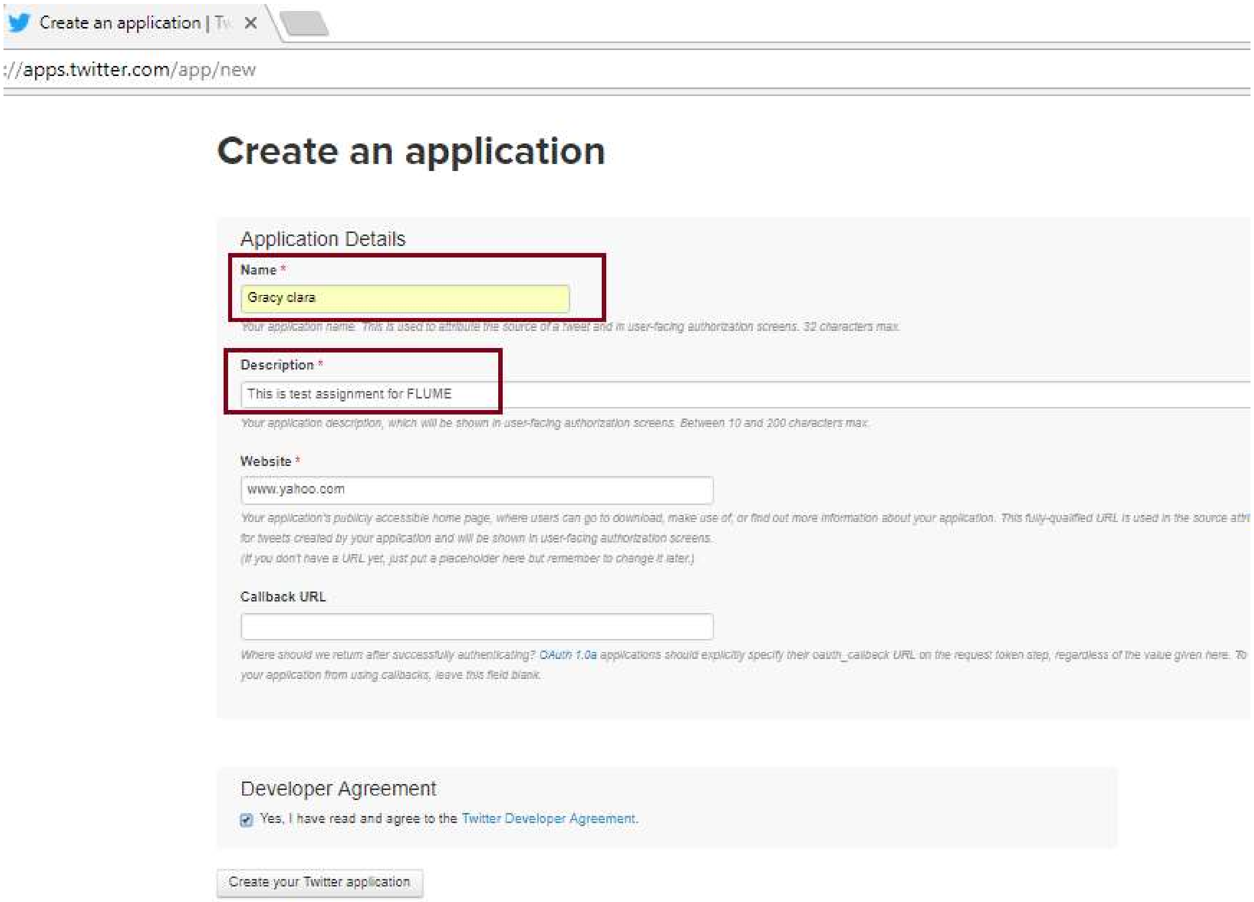
□ Remember me · Forgot password?

New to Twitter? Sign up now »

Already using Twitter via text message? Activate your account »

Go to the following link and click the ‘create new app’ button.

https://apps.twitter.com/app



Create an application

Application Details

Name *

Gracy clara

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

This is test assignment for FLUME

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

www.yahoo.com

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attri for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

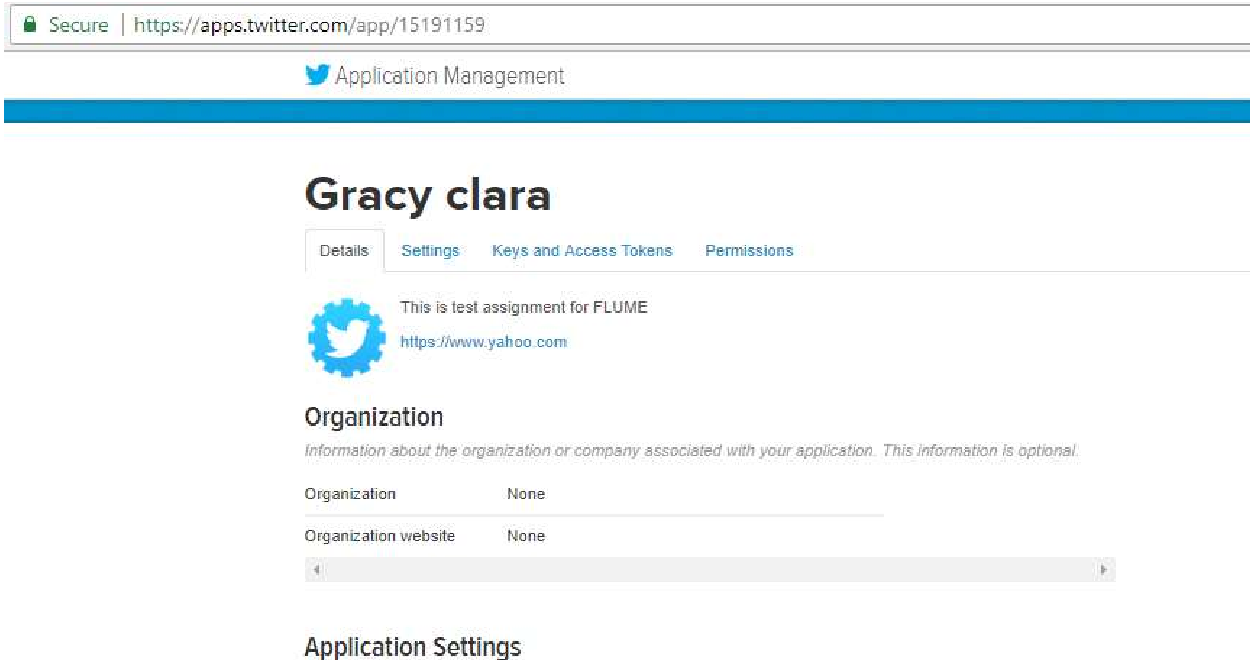
Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To your application from using callbacks, leave this field blank.

Developer Agreement

☑ Yes, I have read and agree to the Twitter Developer Agreement.

Accept the developer agreement and select the ‘create your Twitter application’ button’

Select the ‘Keys and Access Token’ tab.



Gracy clara

This is test assignment for FLUME

https://www.yahoo.com

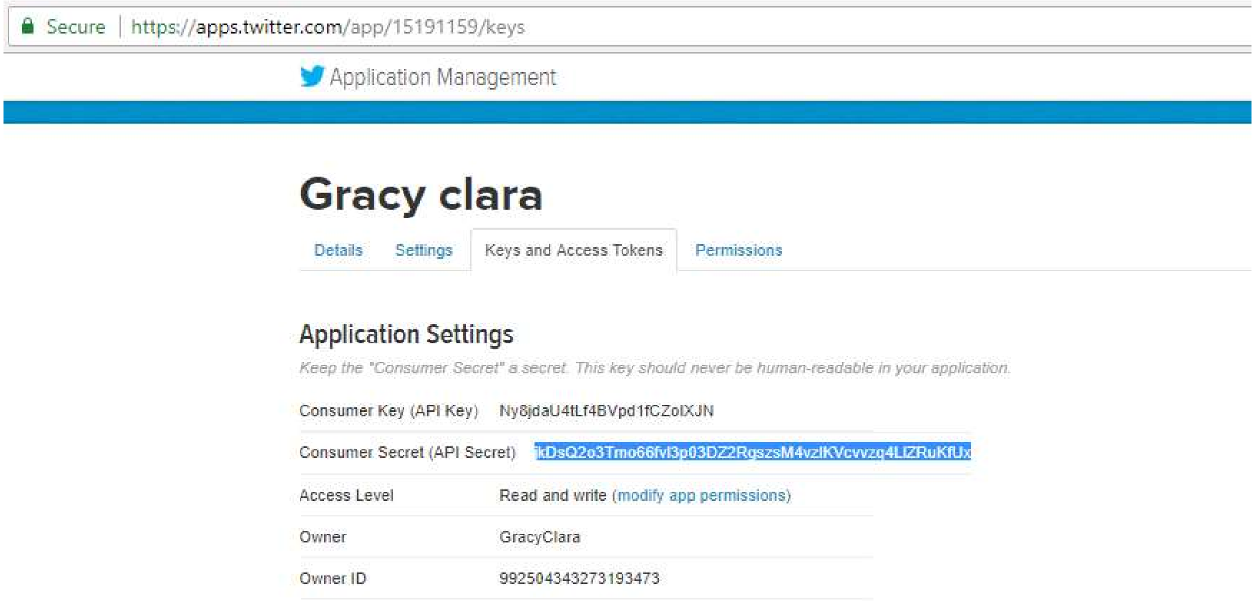
Organization

Information about the organization or company associated with your application. This information is optional.

Organization	None
Organization website	None

Application Settings

Copy the consumer key and the consumer secret code, scroll down further and select the ‘create my access token’ button.



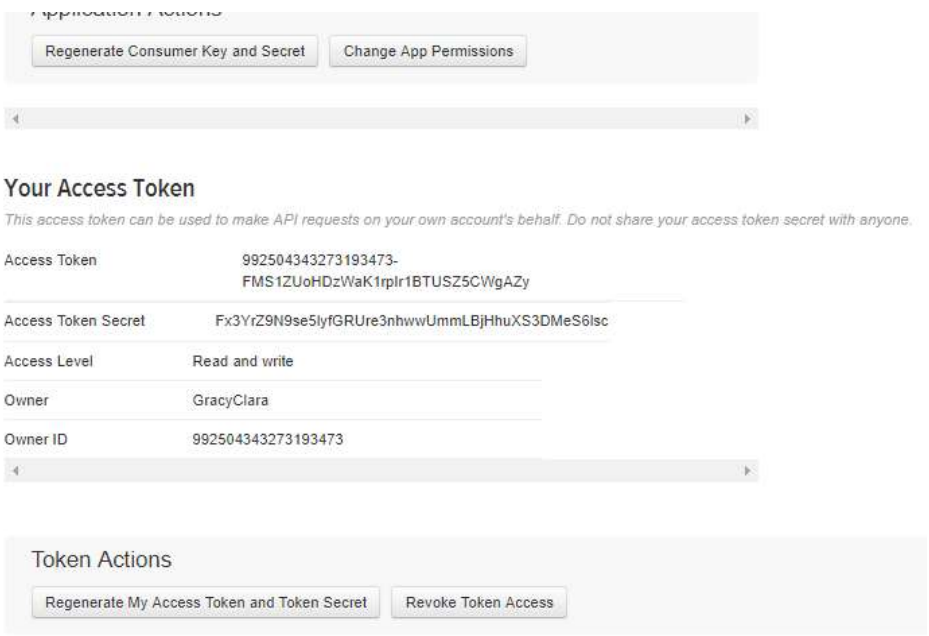
Gracy clara

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	Ny8jdaU4tLf4BVpd1fCZoIXJN
Consumer Secret (API Secret)	ikDsQ2o3Tmo66fvl3p03DZ2RgszsM4vzlKVcvvzq4LlZRuKfUx
Access Level	Read and write (modify app permissions)
Owner	GracyClara
Owner ID	992504343273193473

Now, you will receive a message “you have successfully generated your application access token”.


Application Actions
Regenerate Consumer Key and Secret
Change App Permissions
Token Actions
Regenerate My Access Token and Token Secret
Revoke Token Access

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	992504343273193473- FMS1ZUoHDzWaK1rplr1BTUSZ5CWgAZy
Access Token Secret	Fx3YrZ9N9se5lyfGRUre3nhwwUmmLBjHhuXS3DMeS6lsc
Access Level	Read and write
Owner	GracyClara
Owner ID	992504343273193473

Copy the Flume configuration code from the below link and paste it in the newly created file in the location,

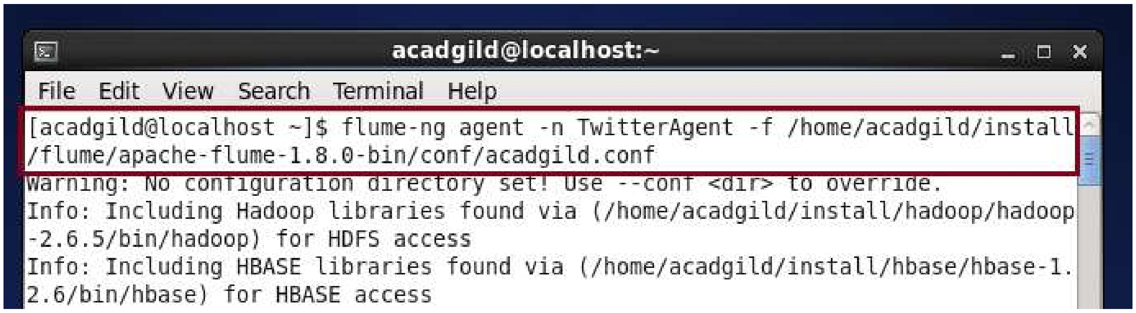
/home/acadgild/apache-flume-1.6.0-bin/conf/acadgild.conf

https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk

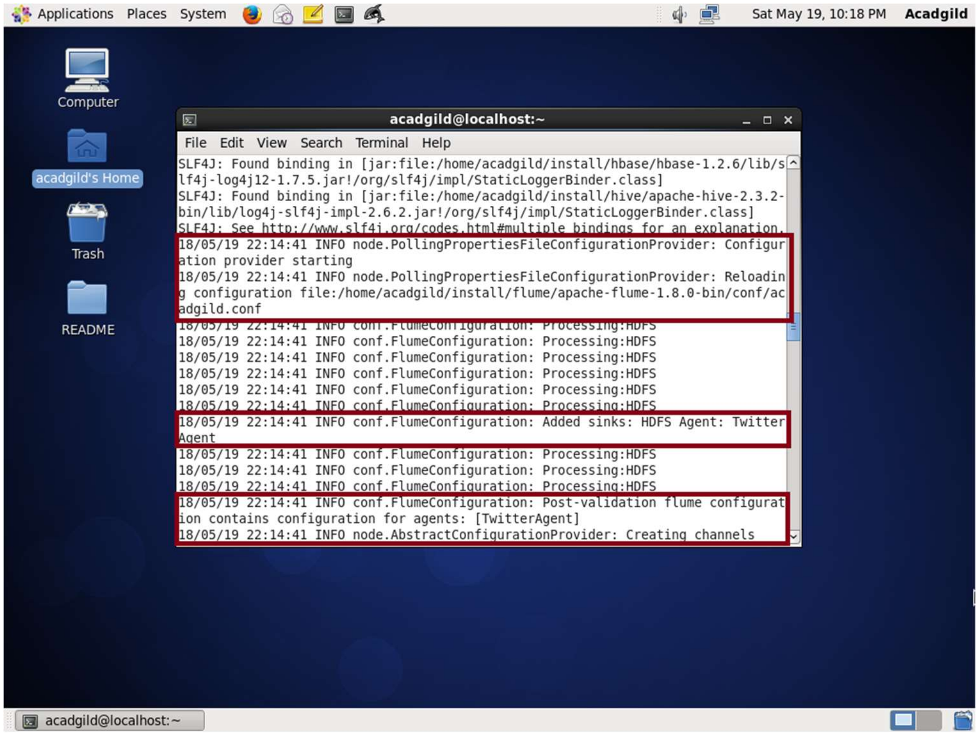
Update the newly created file with twitter **api** keys like consumer key, Consumer token, Access token and the access token secret code and with the **key words.**

hadoop dfs –mkdir /user/acadgild/hadoop/tweets

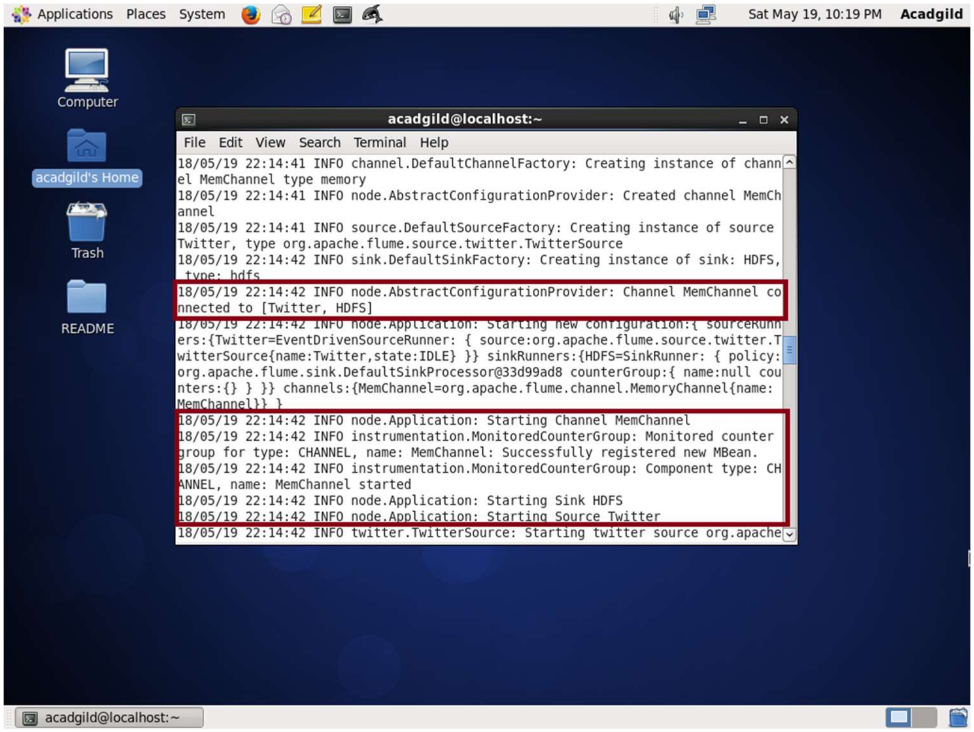
flume-ng agent -n TwitterAgent -f /home/acadgild/apache-flume-1.6.0-bin/conf/acadgild.conf



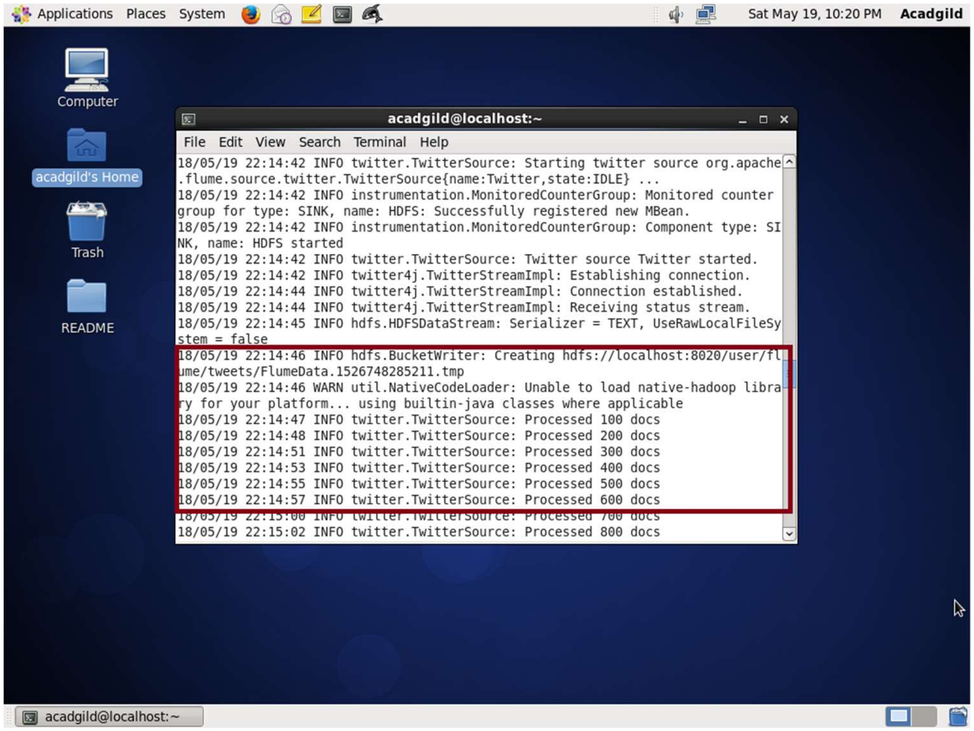
```
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/install
/flume/apache-flume-1.8.0-bin/conf/acadgild.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/install/hadoop/hadoop
-2.6.5/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/home/acadgild/install/hbase/hbase-1.
2.6/bin/hbase) for HBASE access
```



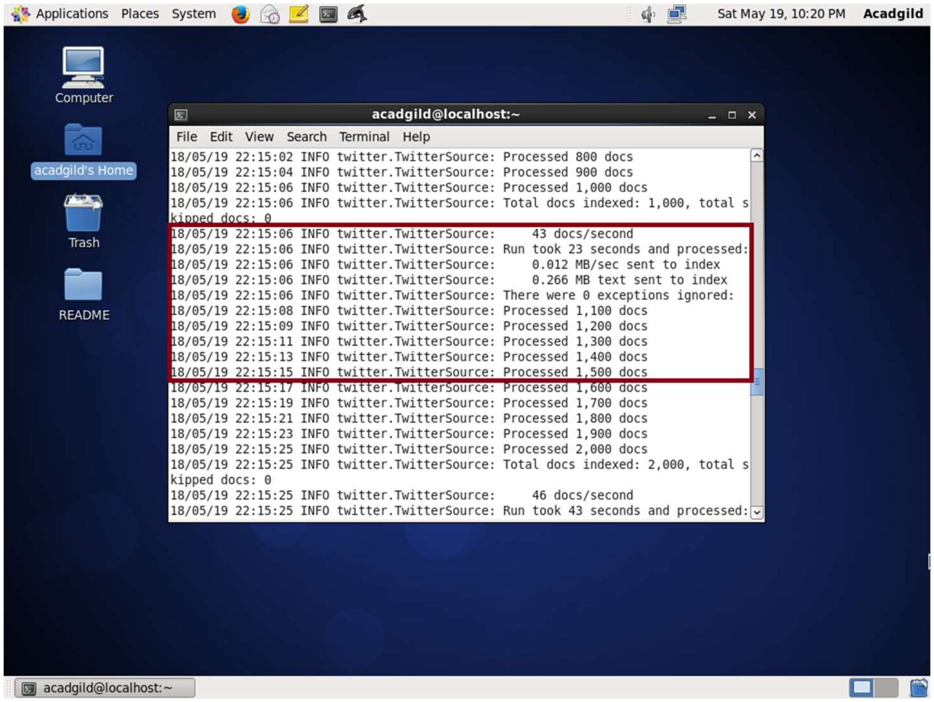
```
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/s
lf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-
bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
18/05/19 22:14:41 INFO node.PollingPropertiesFileConfigurationProvider: Configur
ation provider starting
18/05/19 22:14:41 INFO node.PollingPropertiesFileConfigurationProvider: Reloadin
g configuration file:/home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/ac
adgild.conf
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Added sinks: HDFS Agent: Twitter
Agent
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/19 22:14:41 INFO conf.FlumeConfiguration: Post-validation flume configurat
ion contains configuration for agents: [TwitterAgent]
18/05/19 22:14:41 INFO node.AbstractConfigurationProvider: Creating channels
```



```
18/05/19 22:14:41 INFO channel.DefaultChannelFactory: Creating instance of chann
el MemChannel type memory
18/05/19 22:14:41 INFO node.AbstractConfigurationProvider: Created channel MemCh
annel
18/05/19 22:14:41 INFO source.DefaultSourceFactory: Creating instance of source
Twitter, type org.apache.flume.source.twitter.TwitterSource
18/05/19 22:14:42 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS,
 type: hdfs
18/05/19 22:14:42 INFO node.AbstractConfigurationProvider: Channel MemChannel co
nnected to [Twitter, HDFS]
18/05/19 22:14:42 INFO node.Application: Starting new configuration:{ sourceRunn
ers:{Twitter=EventDrivenSourceRunner: { source:org.apache.flume.source.twitter.T
witterSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS=SinkRunner: { policy:
org.apache.flume.sink.DefaultSinkProcessor@33d99ad8 counterGroup:{ name:null cou
nters:{} } }} channels:{MemChannel=org.apache.flume.channel.MemoryChannel{name:
MemChannel}} }
18/05/19 22:14:42 INFO node.Application: Starting Channel MemChannel
18/05/19 22:14:42 INFO instrumentation.MonitoredCounterGroup: Monitored counter
group for type: CHANNEL, name: MemChannel: Successfully registered new MBean.
18/05/19 22:14:42 INFO instrumentation.MonitoredCounterGroup: Component type: CH
ANNEL, name: MemChannel started
18/05/19 22:14:42 INFO node.Application: Starting Sink HDFS
18/05/19 22:14:42 INFO node.Application: Starting Source Twitter
18/05/19 22:14:42 INFO twitter.TwitterSource: Starting twitter source org.apache
```

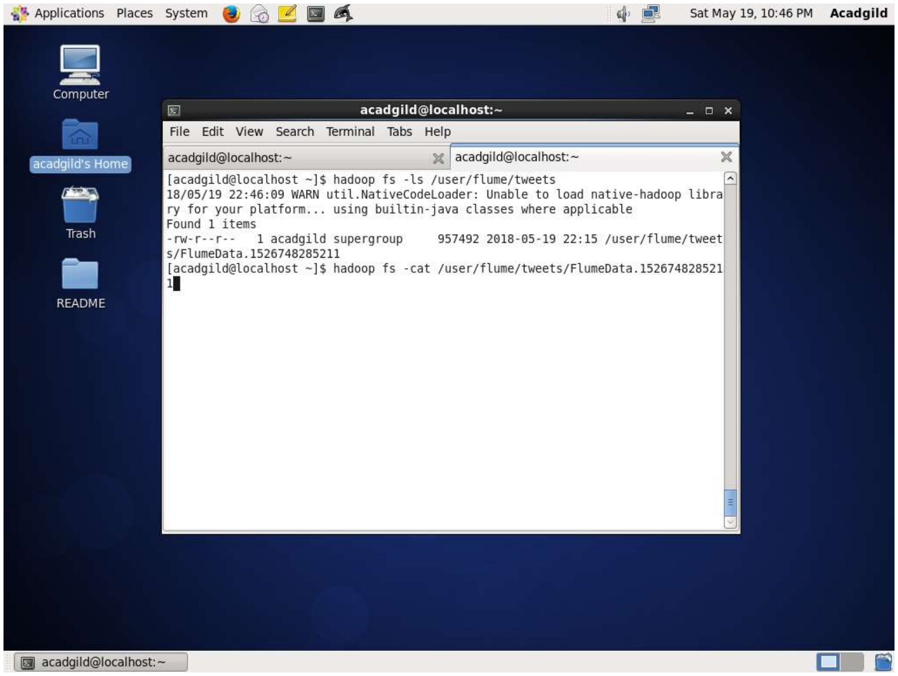


```
18/05/19 22:14:42 INFO twitter.TwitterSource: Starting twitter source org.apache
.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} ...
18/05/19 22:14:42 INFO instrumentation.MonitoredCounterGroup: Monitored counter
group for type: SINK, name: HDFS: Successfully registered new MBean.
18/05/19 22:14:42 INFO instrumentation.MonitoredCounterGroup: Component type: SI
NK, name: HDFS started
18/05/19 22:14:42 INFO twitter.TwitterSource: Twitter source Twitter started.
18/05/19 22:14:42 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/05/19 22:14:44 INFO twitter4j.TwitterStreamImpl: Connection established.
18/05/19 22:14:44 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/05/19 22:14:45 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSy
stem = false
18/05/19 22:14:46 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/fl
ume/tweets/FlumeData.1526748285211.tmp
18/05/19 22:14:46 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
18/05/19 22:14:47 INFO twitter.TwitterSource: Processed 100 docs
18/05/19 22:14:48 INFO twitter.TwitterSource: Processed 200 docs
18/05/19 22:14:51 INFO twitter.TwitterSource: Processed 300 docs
18/05/19 22:14:53 INFO twitter.TwitterSource: Processed 400 docs
18/05/19 22:14:55 INFO twitter.TwitterSource: Processed 500 docs
18/05/19 22:14:57 INFO twitter.TwitterSource: Processed 600 docs
18/05/19 22:15:00 INFO twitter.TwitterSource: Processed 700 docs
18/05/19 22:15:02 INFO twitter.TwitterSource: Processed 800 docs
```



```
18/05/19 22:15:02 INFO twitter.TwitterSource: Processed 800 docs
18/05/19 22:15:04 INFO twitter.TwitterSource: Processed 900 docs
18/05/19 22:15:06 INFO twitter.TwitterSource: Processed 1,000 docs
18/05/19 22:15:06 INFO twitter.TwitterSource: Total docs indexed: 1,000, total s
kipped docs: 0
18/05/19 22:15:06 INFO twitter.TwitterSource:     43 docs/second
18/05/19 22:15:06 INFO twitter.TwitterSource: Run took 23 seconds and processed:
18/05/19 22:15:06 INFO twitter.TwitterSource:     0.012 MB/sec sent to index
18/05/19 22:15:06 INFO twitter.TwitterSource:     0.266 MB text sent to index
18/05/19 22:15:06 INFO twitter.TwitterSource: There were 0 exceptions ignored:
18/05/19 22:15:08 INFO twitter.TwitterSource: Processed 1,100 docs
18/05/19 22:15:09 INFO twitter.TwitterSource: Processed 1,200 docs
18/05/19 22:15:11 INFO twitter.TwitterSource: Processed 1,300 docs
18/05/19 22:15:13 INFO twitter.TwitterSource: Processed 1,400 docs
18/05/19 22:15:15 INFO twitter.TwitterSource: Processed 1,500 docs
18/05/19 22:15:17 INFO twitter.TwitterSource: Processed 1,600 docs
18/05/19 22:15:19 INFO twitter.TwitterSource: Processed 1,700 docs
18/05/19 22:15:21 INFO twitter.TwitterSource: Processed 1,800 docs
18/05/19 22:15:23 INFO twitter.TwitterSource: Processed 1,900 docs
18/05/19 22:15:25 INFO twitter.TwitterSource: Processed 2,000 docs
18/05/19 22:15:25 INFO twitter.TwitterSource: Total docs indexed: 2,000, total s
kipped docs: 0
18/05/19 22:15:25 INFO twitter.TwitterSource:     46 docs/second
18/05/19 22:15:25 INFO twitter.TwitterSource: Run took 43 seconds and processed:
```

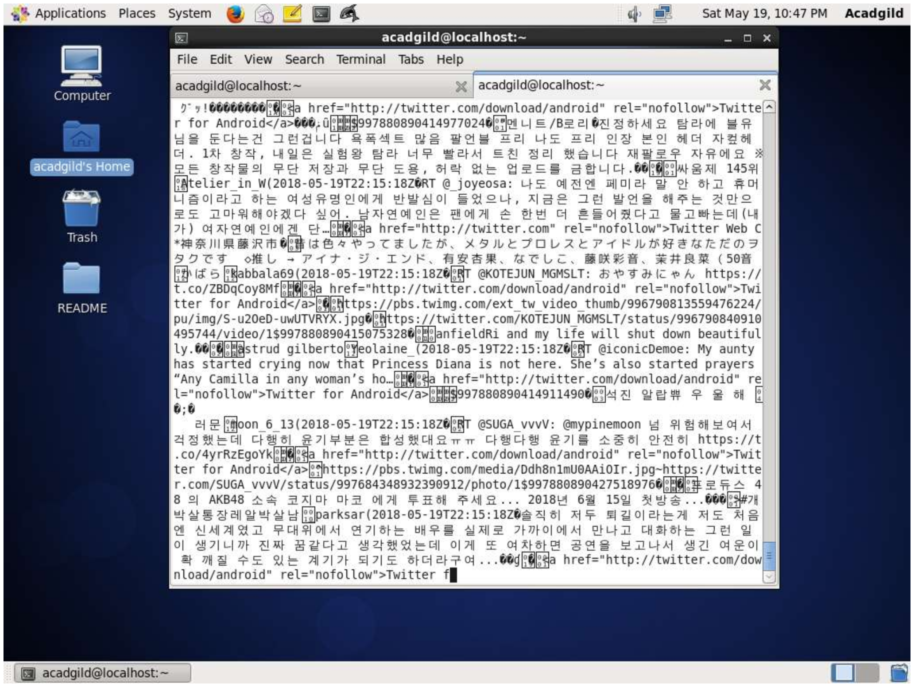
Once, the tweet data started streaming it into the given HDFS path we can use ‘Ctrl+c’ command to stop the streaming process.



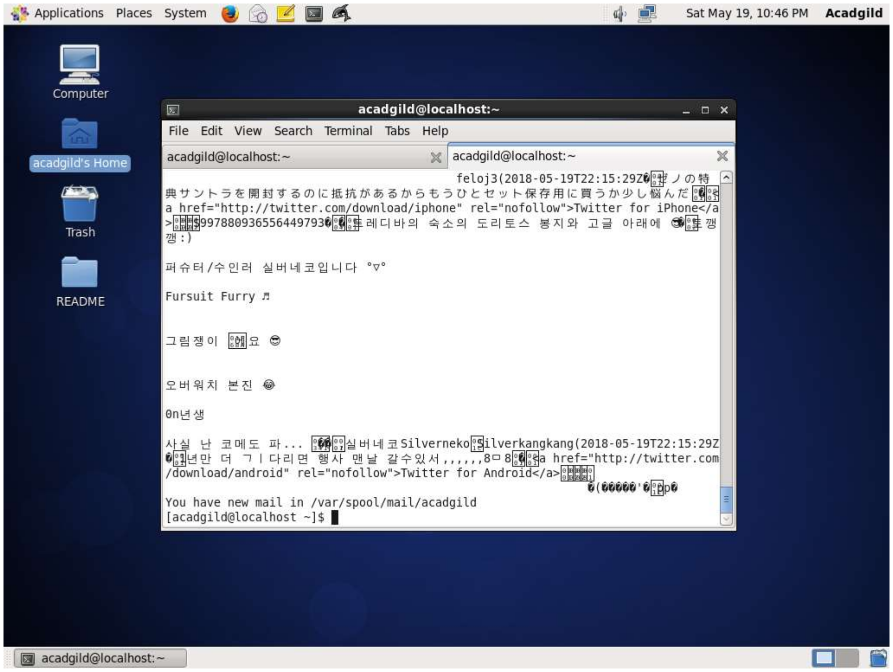
```
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets
18/05/19 22:46:09 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup     957492 2018-05-19 22:15 /user/flume/tweet
s/FlumeData.1526748285211
[acadgild@localhost ~]$ hadoop fs -cat /user/flume/tweets/FlumeData.152674828521
1
```

To check the contents of the tweet data we can use the following command:

hadoop fs -cat /user/acadgild/hadoop/tweets/FlumeData.1526748285211



```
 ｸﾞｯ!��������▯�▯a href="http://twitter.com/download/android" rel="nofollow">Twitte
r for Android</a>���,û▯▯$997880890414977024�▯멘니트/B로리�진정하세요 탐라에 블유
님을 둔다는건 그런겁니다 욕폭섹트 많음 팔언블 프리 나도 프리 인장 본인 헤더 자컴헤
더. 1차 창작, 내일은 실험왕 탐라 너무 빨라서 트친 정리 했습니다 재팔로우 자유에요 ※
모든 창작물의 무단 저장과 무단 도용, 허락 없는 업로드를 금합니다.��▯�▯싸움제 145위
▯Atelier_in_W(2018-05-19T22:15:18Z�RT @_joyeosa: 나도 예전엔 페미라 말 안 하고 휴머
니즘이라고 하는 여성유명인에게 반발심이 들었으나, 지금은 그런 발언을 해주는 것만으
로도 고마워해야겠다 싶어. 남자연예인은 팬에게 손 한번 더 흔들어줬다고 물고빠는데(내
가) 여자연예인에겐 단…▯�▯a href="http://twitter.com" rel="nofollow">Twitter Web C
*神奈川県藤沢市�▯昔は色々やってましたが、メタルとプロレスとアイドルが好きなただのヲ
タクです ◇推し → アイナ・ジ・エンド、有安杏果、なでしこ、藤咲彩音、茉井良菜（50音
▯かばら▯kabbala69(2018-05-19T22:15:18Z�▯RT @KOTEJUN_MGMSLT: おやすみにゃん https://
t.co/ZBDqCoy8Mf▯�▯a href="http://twitter.com/download/android" rel="nofollow">Twi
tter for Android</a>▯�▯https://pbs.twimg.com/ext_tw_video_thumb/996790813559476224/
pu/img/S-u2OeD-uwUTVRYX.jpg�▯https://twitter.com/KOTEJUN_MGMSLT/status/996790840910
495744/video/1$997880890415075328�▯anfieldRi and my life will shut down beautiful
ly.��▯�▯astrud gilberto▯Yeolaine_(2018-05-19T22:15:18Z�▯RT @iconicDemoe: My aunty
has started crying now that Princess Diana is not here. She’s also started prayers
“Any Camilla in any woman’s ho…▯�▯a href="http://twitter.com/download/android" re
l="nofollow">Twitter for Android</a>▯$997880890414911490�▯석진 알랍뿌 우 울 해 ▯
�;�
   러문▯moon_6_13(2018-05-19T22:15:18Z�▯RT @SUGA_vvvV: @mypinemoon 넘 위험해보여서
걱정했는데 다행히 윤기부분은 합성했대요ㅠㅠ 다행다행 윤기를 소중히 안전히 https://t
.co/4yrRzEgoYk▯�▯a href="http://twitter.com/download/android" rel="nofollow">Twit
ter for Android</a>▯https://pbs.twimg.com/media/Ddh8n1mU0AAiOIr.jpg~https://twitte
r.com/SUGA_vvvV/status/997684348932390912/photo/1$997880890427518976�▯�▯프로듀스 4
8 의 AKB48 소속 코지마 마코 에게 투표해 주세요... 2018년 6월 15일 첫방송...���▯#개
박살통장레알박살남▯parksar(2018-05-19T22:15:18Z�솔직히 저두 퇴길이라는게 저도 처음
엔 신세계였고 무대위에서 연기하는 배우를 실제로 가까이에서 만나고 대화하는 그런 일
이 생기니까 진짜 꿈같다고 생각했었는데 이게 또 여차하면 공연을 보고나서 생긴 여운이
 확 깨질 수도 있는 계기가 되기도 하더라구여...��g▯�▯a href="http://twitter.com/dow
nload/android" rel="nofollow">Twitter f
```



```
                                            feloj3(2018-05-19T22:15:29Z�▯ザノの特
典サントラを開封するのに抵抗があるからもうひとセット保存用に買うか少し悩んだ▯�▯
a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a
>▯$997880936556449793�▯�▯트레디바의 숙소의 도리토스 봉지와 고글 아래에 ☻▯트깽
깽:)

퍼슈터/수인러 실버네코입니다 °▽°

Fursuit Furry ♬


그림쟁이 ▯요 😎


오버워치 본진 😂

0n년생

사실 난 코메도 파... ▯�▯실버네코Silverneko▯Silverkangkang(2018-05-19T22:15:29Z
�▯1년만 더 ㄱㅣ다리면 행사 맨날 갈수있서,,,,,,8ㅁ8▯�▯a href="http://twitter.com
/download/android" rel="nofollow">Twitter for Android</a>▯▯
                                                                   �(�����'�▯pp�

You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ █
```