Big Data and Hadoop

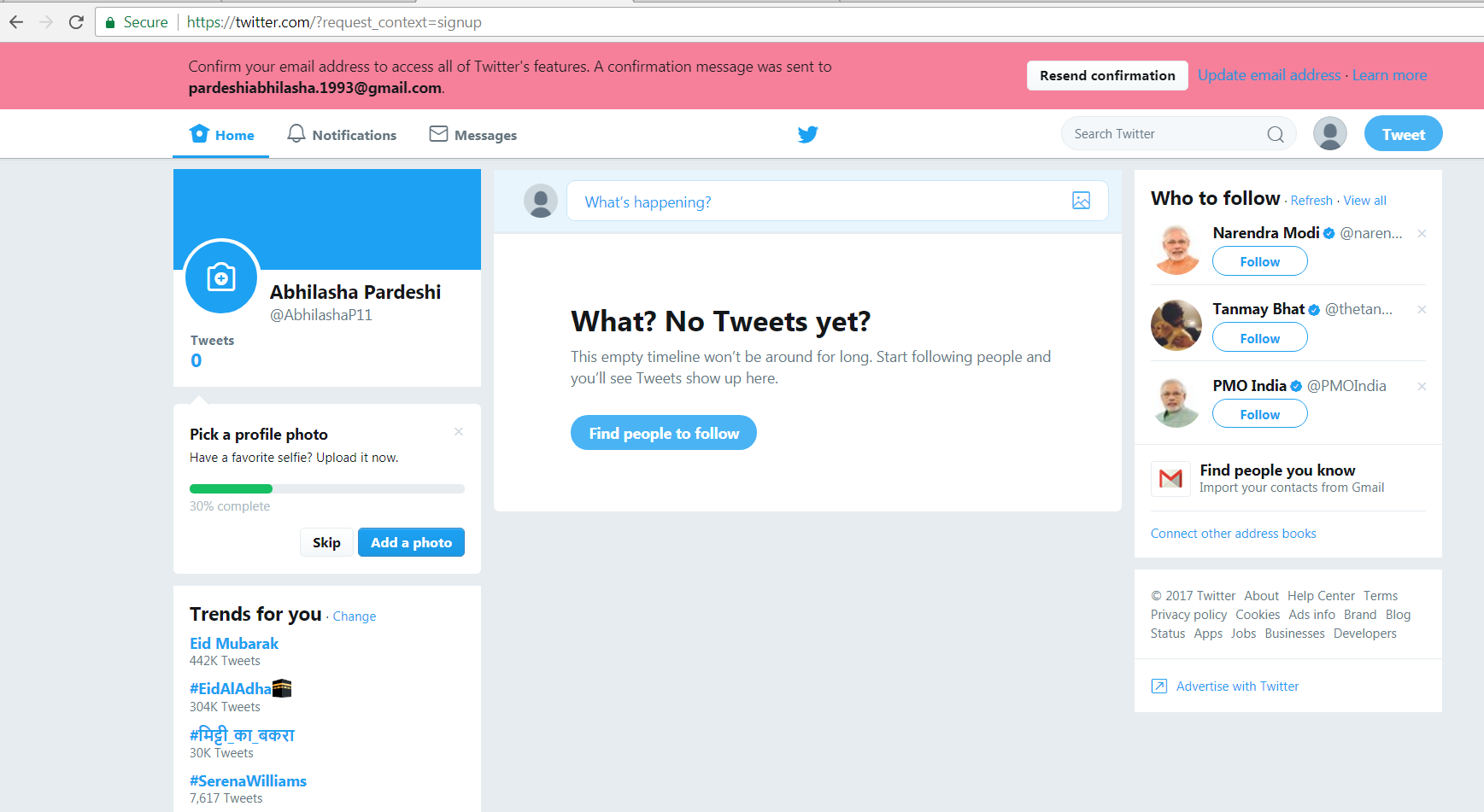
Session 11: Assignment1

**Problem Statement:**

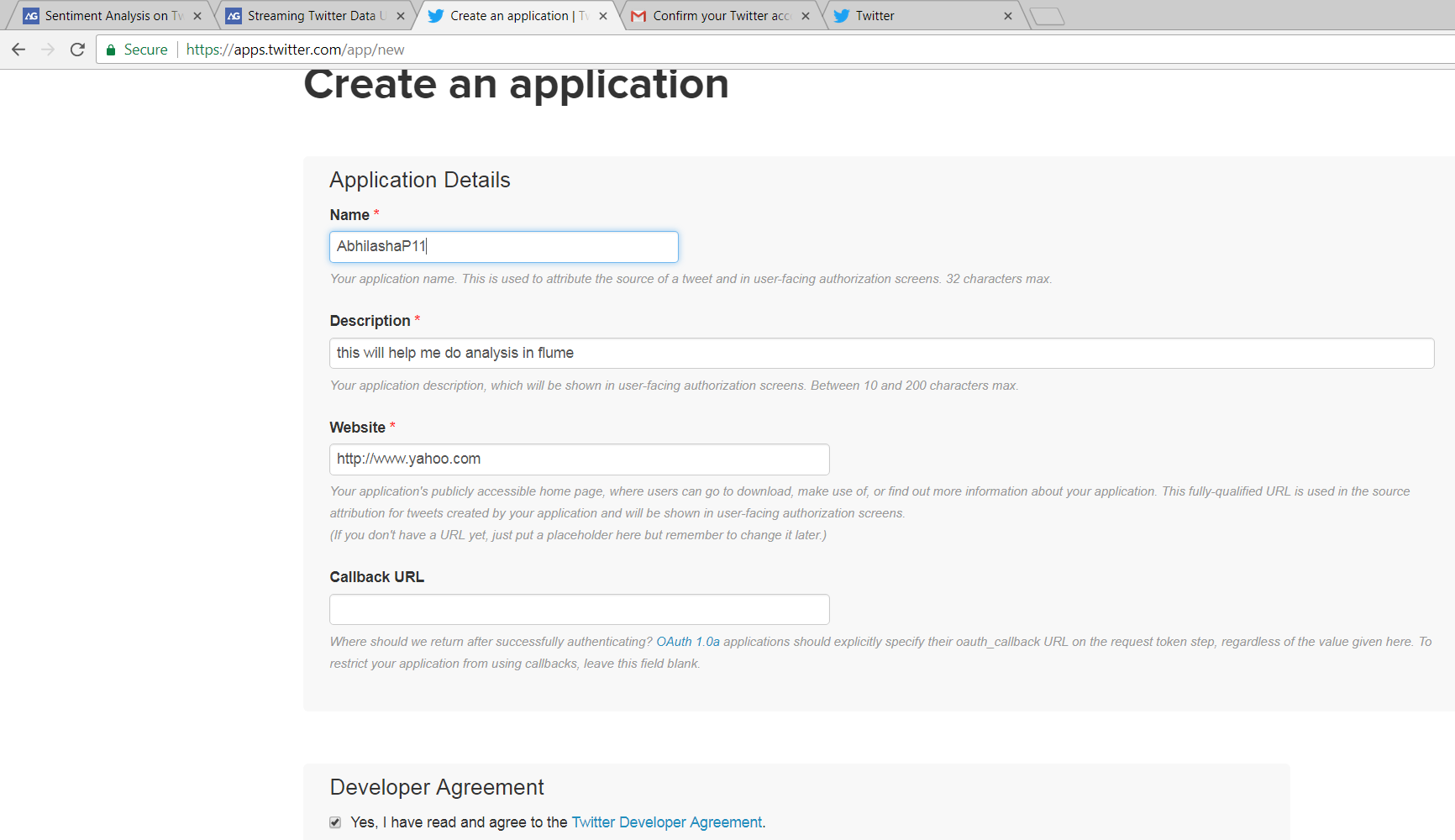
Create a flume agent that streams data from Twitter and stores in the HDFS.

**Solution:**

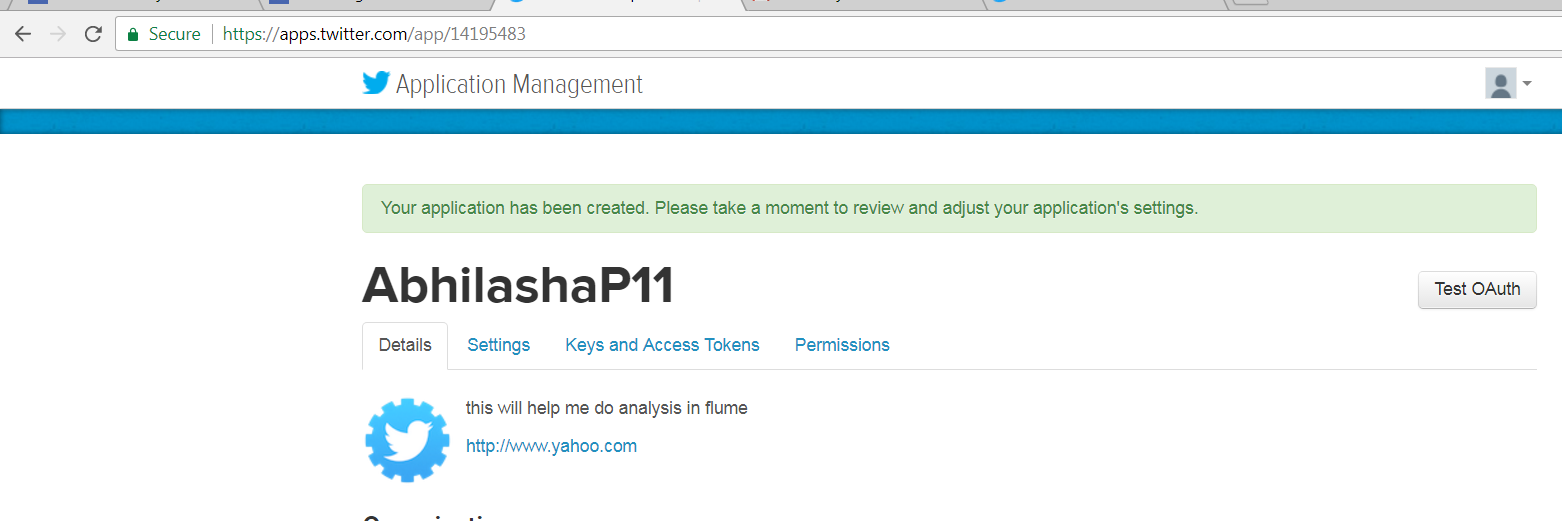
* To stream data to our database from twitter we should have Twitter account as the pre-requisite.
* So created a twitter account and logged into it. The snapshot shows the logged in twitter account.



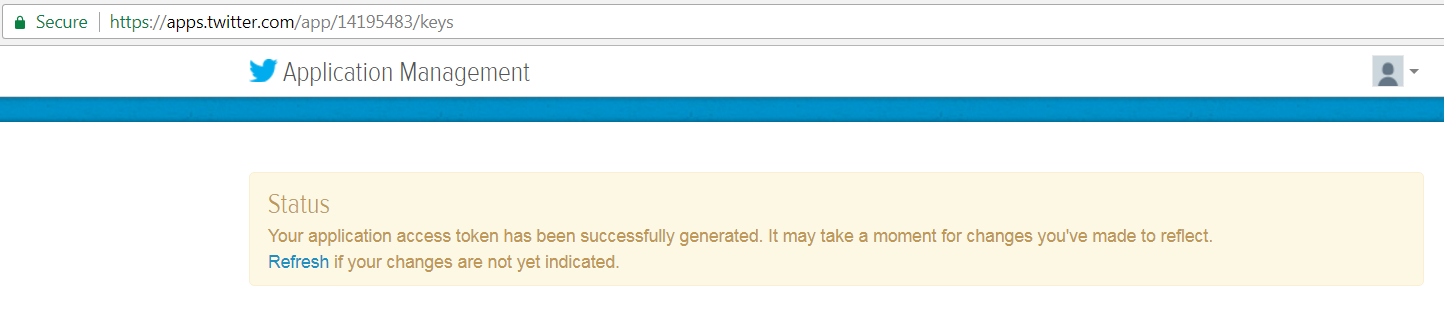
* Then, we create a new app in twitter as follows:
  1. We click on ‘Create New App’ button



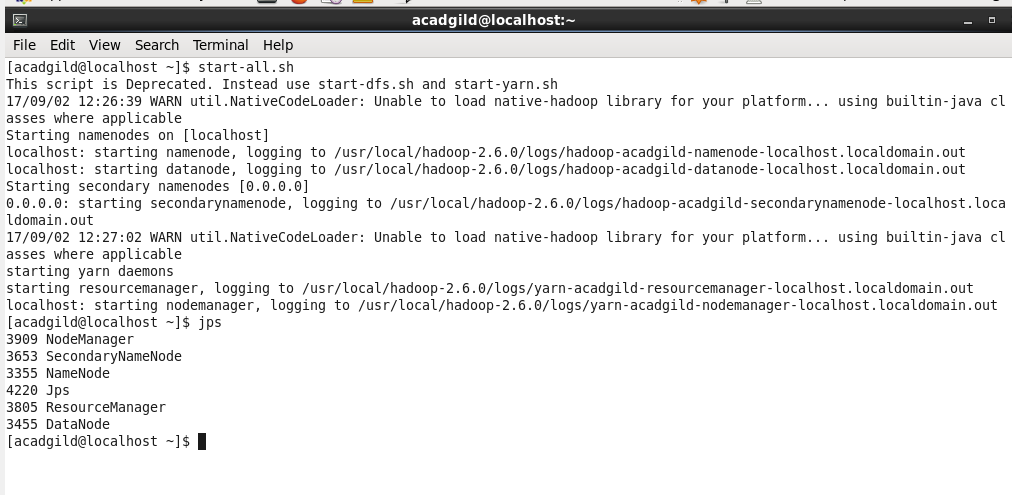
* 1. This shows new application got created



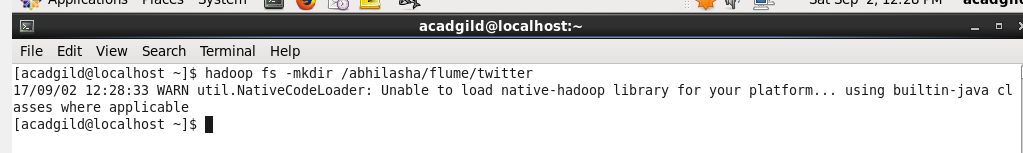
* 1. After the creation of the application, we consumer key, consumer secret code.
  2. Then we generate access token. Hence, we have access token and access token secret code.



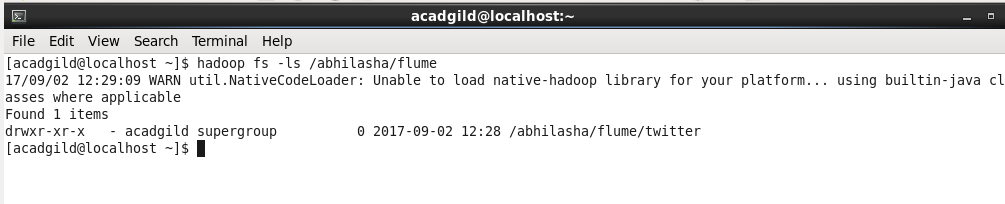
* Now, we start Hadoop cluster using the command start-all.sh



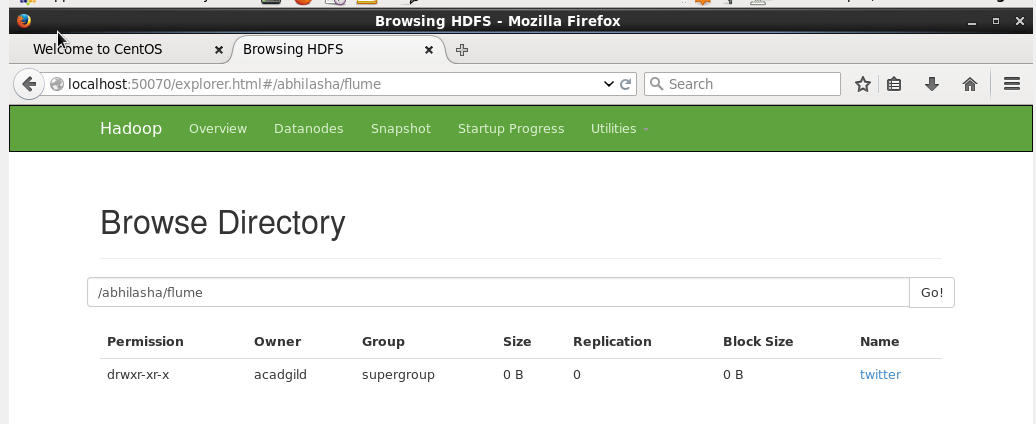
* Then we create a folder in hdfs where we want the data extracted from twitter, to be stored. We use mkdir command for the same. The destination folder is ‘/abhilasha/flume/twitter’.



* Ls command is used to confirm creation of destination folder.



We can also see the folder created in HDFS UI as follows:



* Next step is configuring the flume agent. Flume agent is made of three parts

1. Source
2. Channel
3. Sink

In our use-case, source is the twitter agent, sink in HDFS and the channel is memory channel.

The configuration file used is flume.conf

P**roperties of source** defined in it are as follows:

1. type = org.apache.flume.source.twitter.TwitterSource

This is to specify the we are using twitter as source

ii. Authorization attributes. These include consumerKey, consumerSecret, accessToken, accessTokenSecret. These values are obtained from twitter account as mentioned earlier.

iii. keywords = hadoop, primeMinister, Food, Travel, Trek

Used to identify the tweets to read.

P**roperties of the channel** defined in it are as follows:

1. type = memory

This is to specify the we are using memory channel

ii. capacity = 1000  
 The maximum number of events stored in the channel

iii. transactionCapacity = 1000

The maximum number of events the channel will take from a source or give to a sink per transaction

P**roperties of the sink** defined in it are as follows:

1. type = hdfs

This is to specify that hdfs is the sink

ii. path = /abhilasha/flume/twitter

The path on HDFS where the data will be transferred

iii. batchSize = 1000

Number of events written to file before it is flushed to HDFS

iv. rollSize = 0

File size to trigger roll, in bytes

v. rollInterval = 600

Number of seconds to wait before rolling current file

vi. rollCount = 1000

Number of events written to file before it rolled

vii. writeFormat=Text

Format for sequence file records. One of “Text” or “Writable”

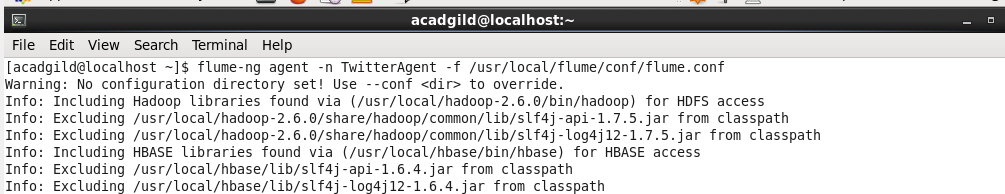
viii. fileType = DataStream

File format: currently SequenceFile, DataStream or CompressedStream

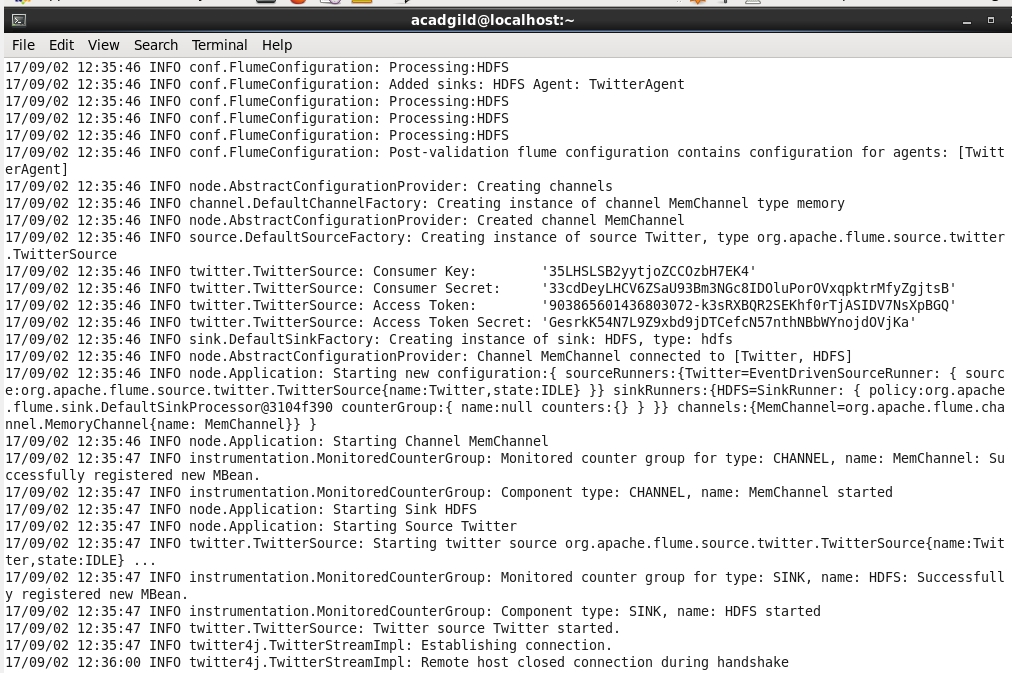
* Next is starting flume to start reading the data. The command used is

flume-ng agent -n TwitterAgent -f <location of created/edited conf file>

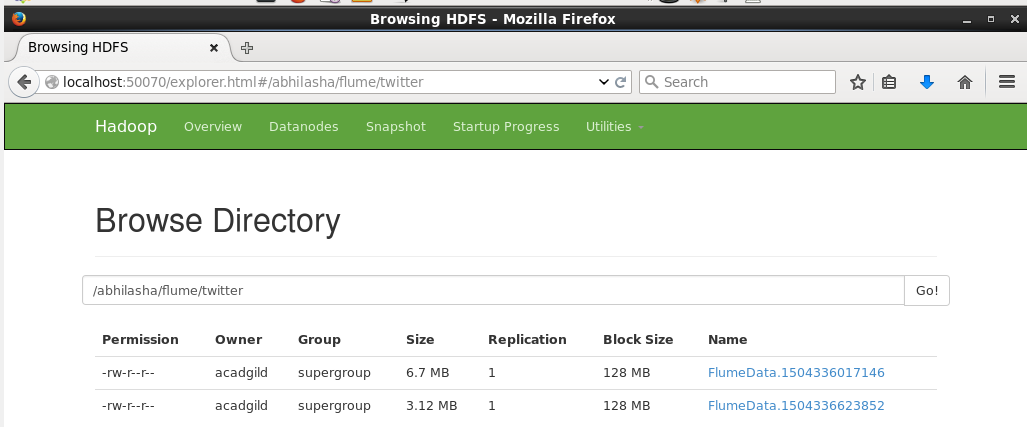
Conf file is ‘/usr/local/flume.conf/flume.conf’



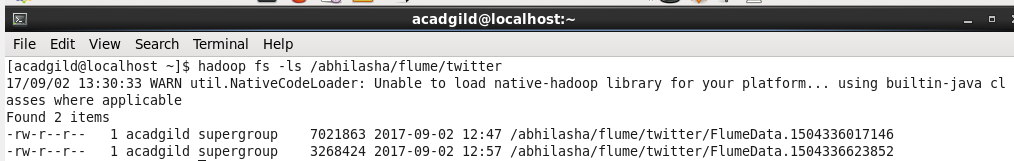
* Screenshot below shows the establishment of flume agent.



* The data extracted from twitter is stored in hdfs at the location specified in conf file. This is seen from the UI of HDFS as follows:



Ls command can also be used to verify the same as follows:



If used cat command to see the data, a snippet of the data is



Data is in JSon format