**Assignment 11.3**

**Problem Statement**

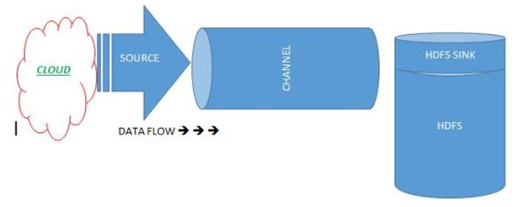
**Create a flume agent that streams data from Twitter and stores in the HDFS.**

**Solution:**

Apache Flume is a Hadoop ecosystem component used to collect, aggregate and moves a large amount of log data from different sources to a centralized data store.

It is an open source component which is designed to locate and store the data in a distributed environment and collects the data as per the specified input key(s).

**Flume Architecture**

****

Flume is composed of the following components.

**Flume Event**: It is the main unit of the data that is transported inside the Flume (Typically a single log entry). It contains a payload of the byte array that is to be transported from the source path to the destination path which could be accompanied by optional headers.

**Flume Agent**: Is an independent Java virtual machine daemon process which receives the data (events) from clients and transports to the subsequent destination (sink or agent).

**Source**: Is the component of Flume agent which receives data from the data generators say, twitter, facebook, weblogs from different sites and transfers this data to one or more channels in the form of Flume event.The external source sends data to Flume in a format that is recognized by the target Flume source. Example, an Avro Flume source can be used to receive Avro data from Avro clients or other Flume agents in the flow that send data from an Avro sink, or the Thrift Flume source will receive data from a Thrift sink, or a Flume Thrift RPC client or Thrift Clients are written in any language generated from the Flume thrift protocol.

**Channel**: Once, the Flume source receives an Event, it stores this data into one or more channel and buffers them till they are consumed by sinks. It acts as a bridge between the source and sinks. These channels are implemented to handle any number of sources and sinks.

**Sink**: It stores the data into the centralized stores like HDFS and HBase.

**Streaming Twitter Data**

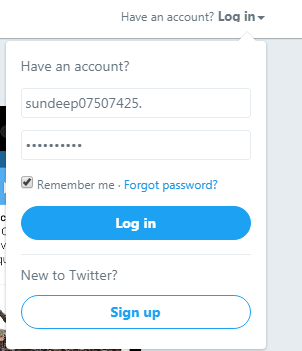
To stream data to our database from twitter we should have the following pre-requisites.

Twitter account

Hadoop cluster

**Step 1:**

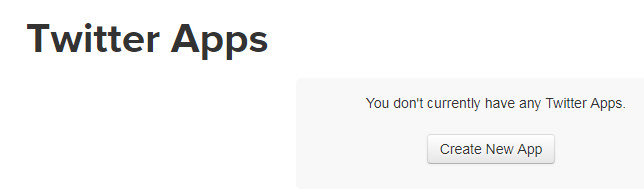
Login to the twitter account



**Step 2:**

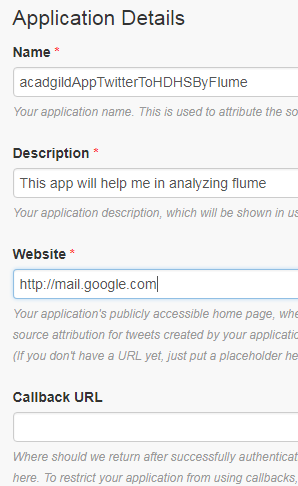
Go to the following link and click the ‘create new app’ button.

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



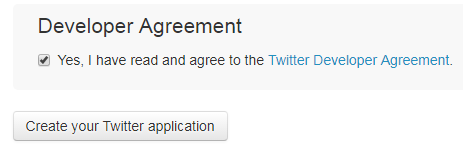
**Step 3:**

Enter the necessary details

. 

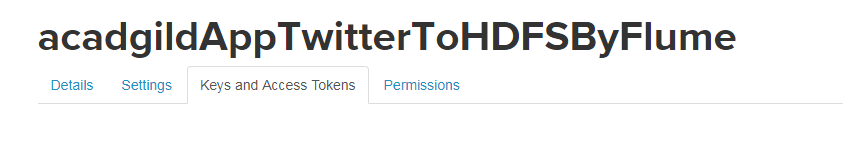
**Step 4:**

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



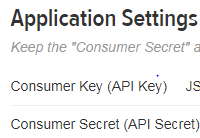
**Step 5:**

Select the ‘Keys and Access Token’ tab.



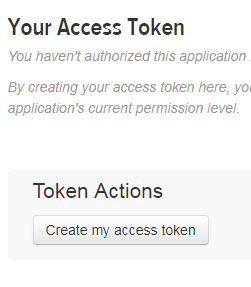
**Step 6:**

Copy the consumer key and the consumer secret code.

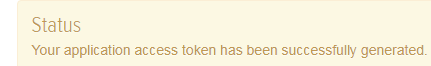


**Step 7:**

Scroll down further and select the ‘create my access token’ button.



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



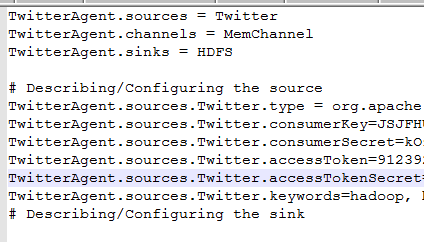
**Step 8:**

Copy the Access Token and Access token Secret code.

**step 9:**

create flume.cof for flume configuration

Change the twitter api keys with the keys generated as shown in the step no 6 and step number 8.



**Step 10:**

We have to decide which keywords tweet data to be collected from the twitter application. So, you can change the keywords in the TwitterAgent.sources.Twitter.keywords command.

In our example, we are fetching tweet data related to Hadoop, election, sports, cricket and Big data.



**Complete Contents of 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=JSJFHUXHmbXwUd18V1ZkCsqSP

TwitterAgent.sources.Twitter.consumerSecret=kOz5akHEyzD1hRUK1DK3enfb1BIxxI8MebWSybfAmmLhqxPCYq

TwitterAgent.sources.Twitter.accessToken=912392212553490432-37Uuo9ZaUDFU1lUTd0aKIx5zkqQQtT5

TwitterAgent.sources.Twitter.accessTokenSecret=O9EzvKvZVLkyZxsSOj2qmmFa3dreKy0g1xDVbbo5jXlxS

TwitterAgent.sources.Twitter.keywords=Hadoop, election, sports, cricket and Big data

# Describing/Configuring the sink

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path =hdfs://localhost:9000/user/flume/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.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

# Binding the source and sink to the channel

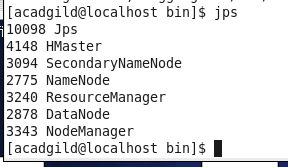
TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sinks.HDFS.channel = MemChannel

**Step 11:**

Open a new terminal and start all the Hadoop daemons, before running the flume command to fetch the twitter data.

Use the ‘jps’ command to see the running Hadoop daemons.

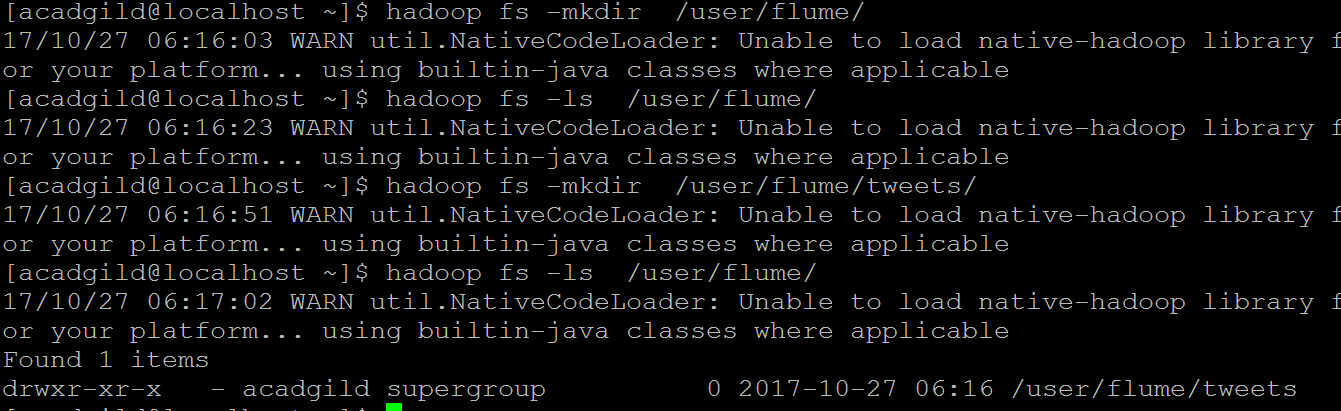


**Step 12:**

Create a new directory inside HDFS path, where the Twitter tweet data should be stored.

hadoop fs –mkdir /user/flume/

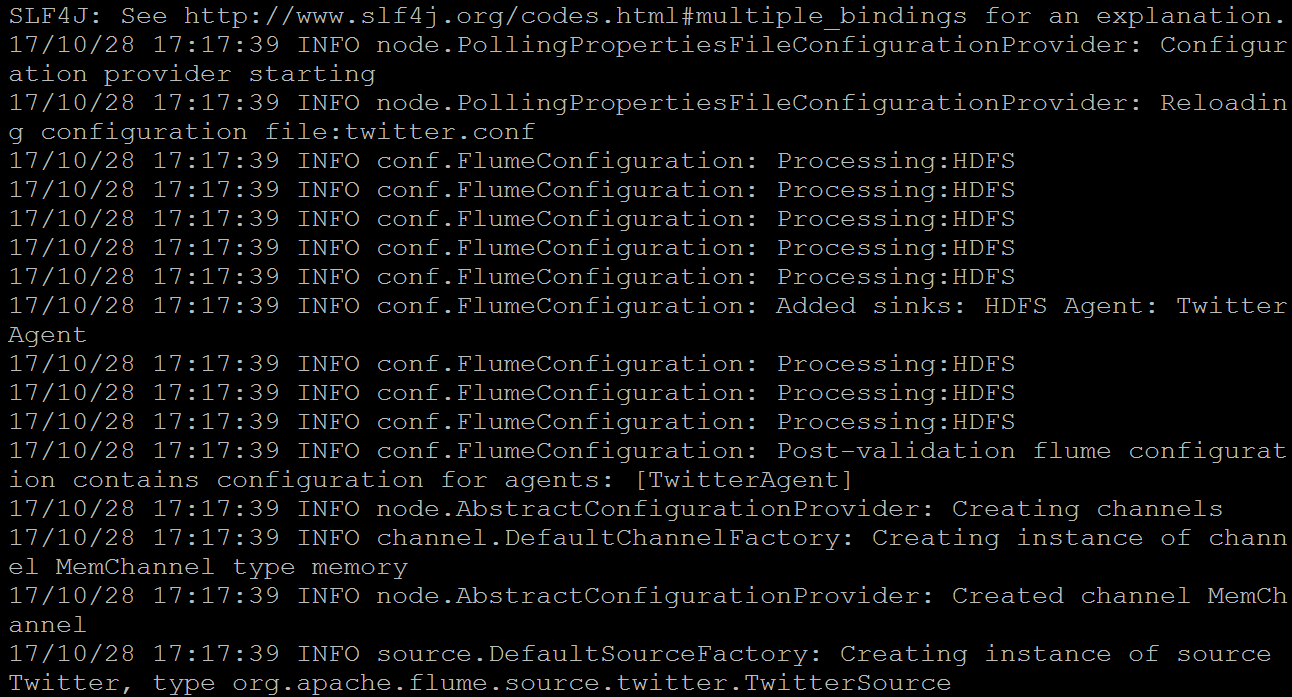
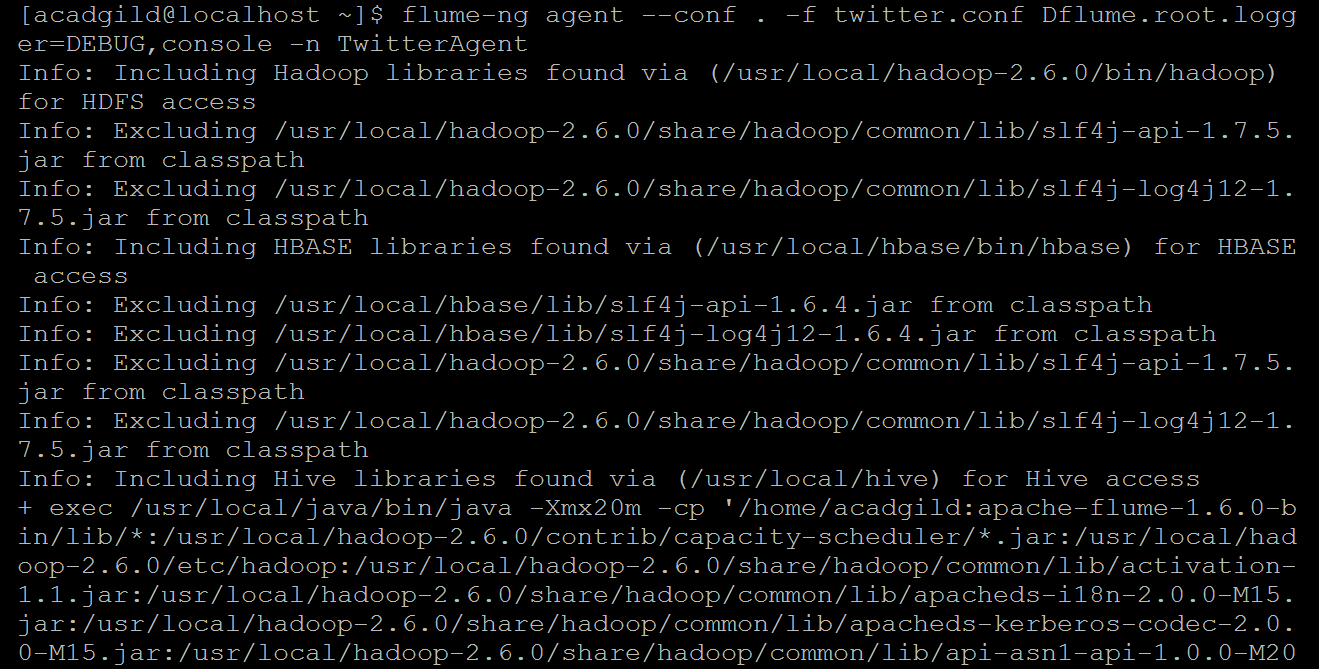
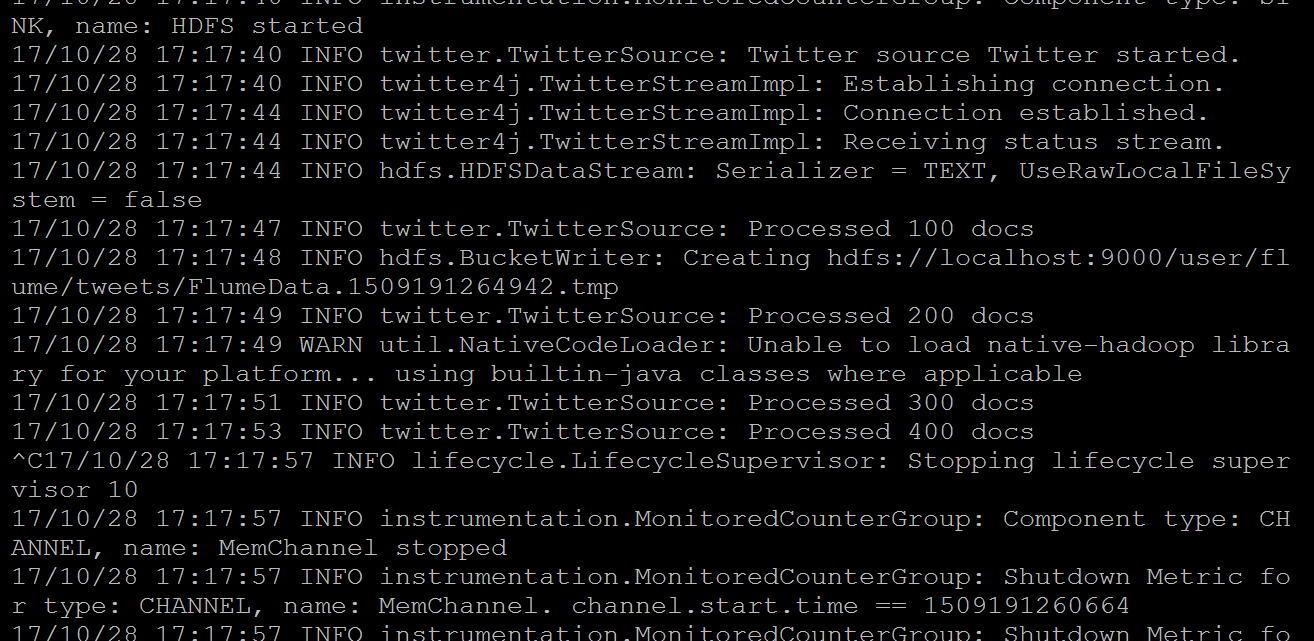
hadoop fs –mkdir /user/flume/tweets



**Step 13:**

For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path.

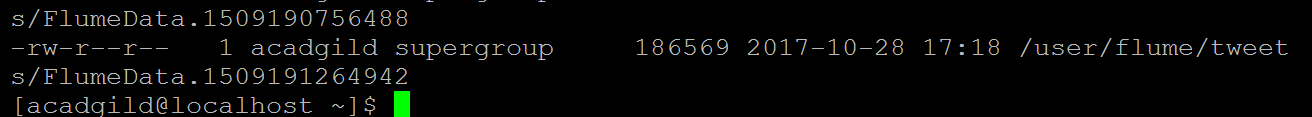
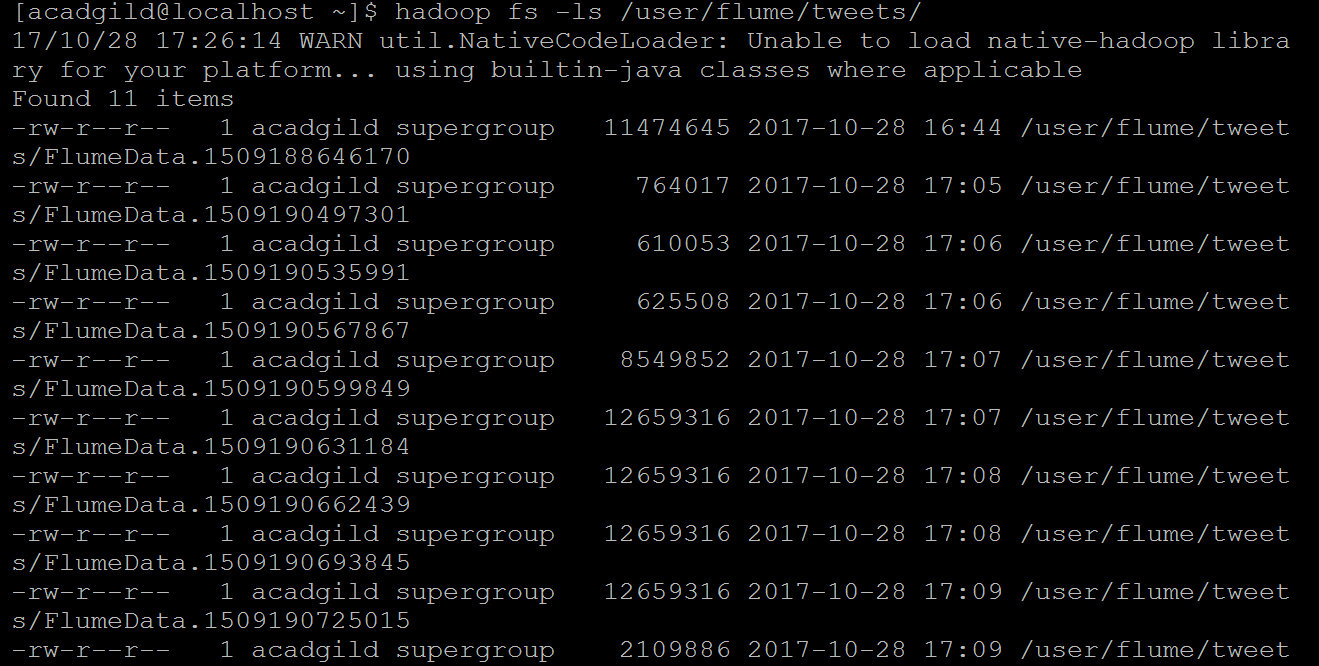
flume-ng agent --conf . -f twitter.conf Dflume.root.logger=DEBUG,console -n TwitterAgent

**Step 14:**

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

hadoop fs -ls /user/flume/tweets/



**Step 15:**

use the ‘cat’ command to display the tweet data

hadoop fs -cat /user/flume/tweets/FlumeData.15091908 |head -10

