**Assignment4.5**

**Problem Statement**

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

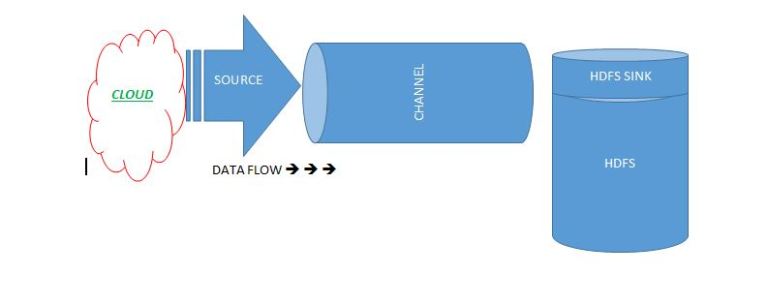
**What is Apache Flume?**

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**

Before moving forward to know the working of flume tool, It is mandatory to know the Flume architecture first.



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.

A Flume event will be in the following structure.

|  |  |
| --- | --- |
| Header | Byte Payload |

**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

If both prerequisites are available we can move to our further step.

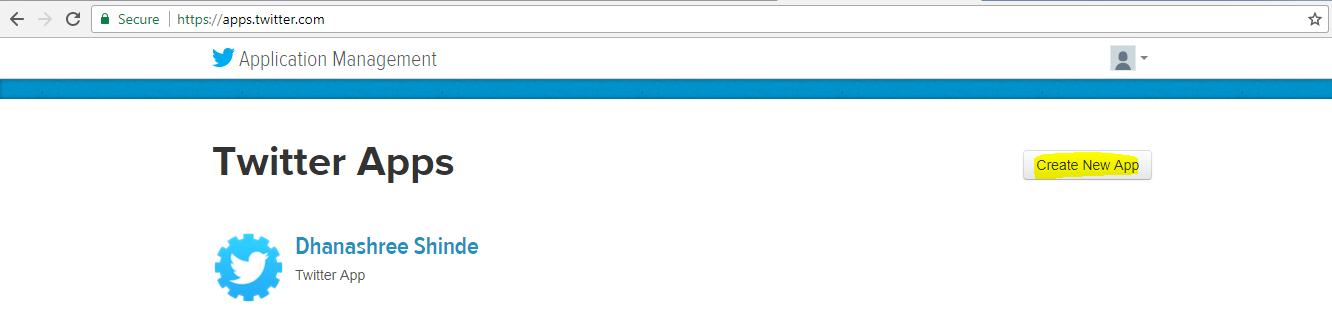
**Step 1:**

Login to  the twitter account

**Step 2:**

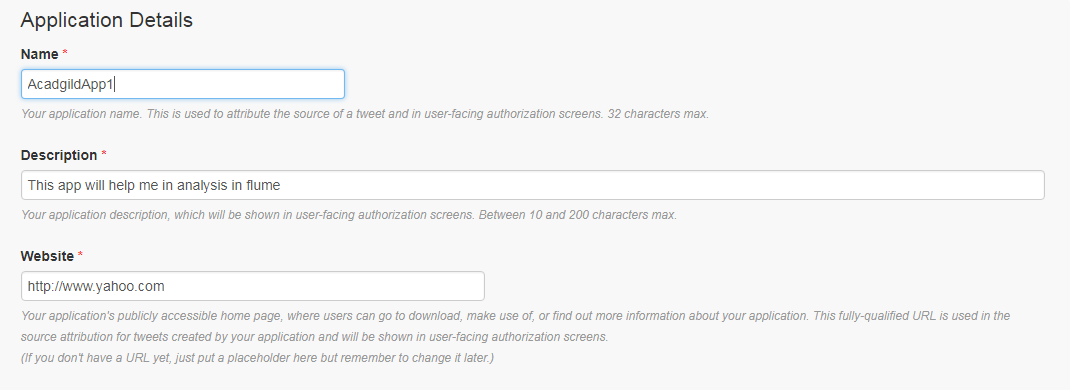
**Go to the following link and click the  ‘create new app’ button.**

[**https://apps.twitter.com/app**](https://apps.twitter.com/app)



**Step 3:**

**Enter the necessary details.**

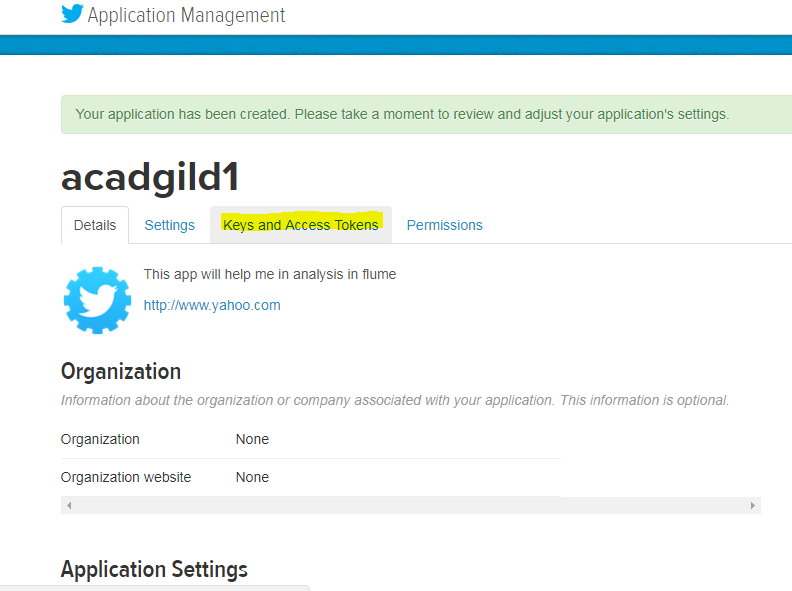


**Step 4:**

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

**Step 5:**

Select the ‘Keys and Access Token’ tab.



**Step 6:**

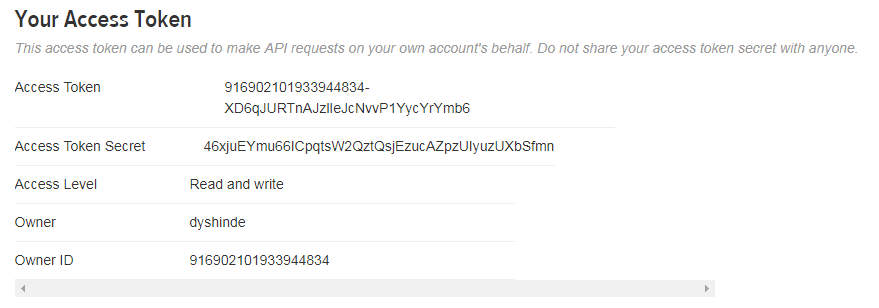
Copy the consumer key and the consumer secret code.

**Step 7:**

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

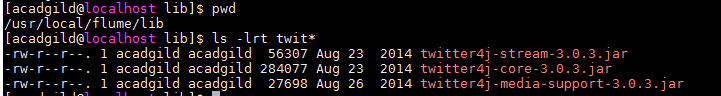
**Step 8:**

Copy the Access Token and Access token Secret code.

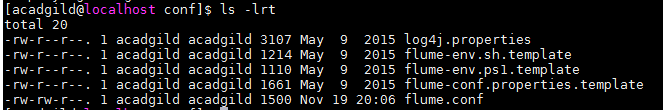


**Settings to be done for flume:**

Check if the jar files related to twitter and flume are in lib directory of flume



And main file .conf should be created as required under conf of flume directory.

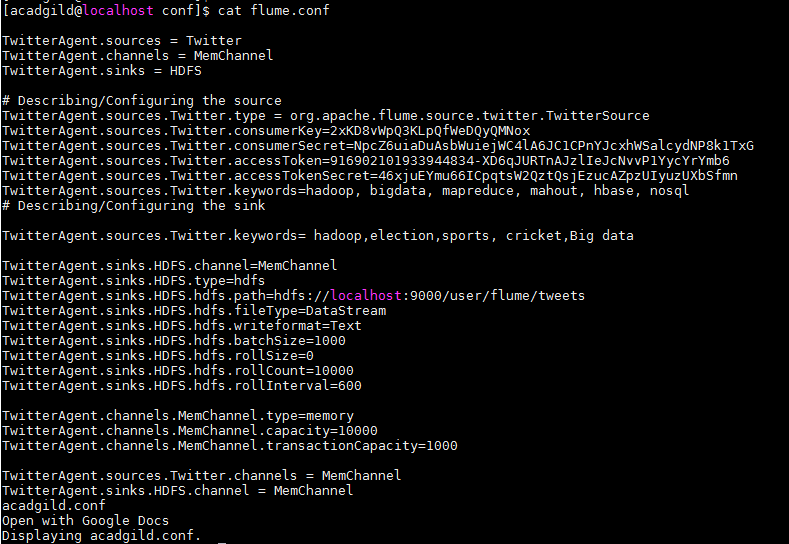


**Flume.conf:**

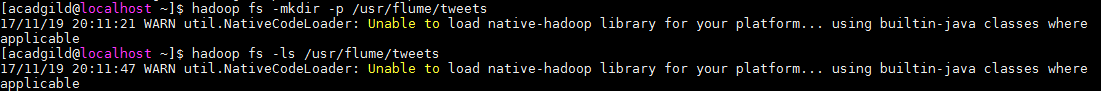
In this file

* defined Agent(TwitterAgent)
* defined source,channel,sink
  1. where Source is twitter – from where data is sourced
  2. and Shannel is intermediate storage until sink reads the source data(Memory channel is used in this case)
  3. and Sink is HDFS – where data pulled from source(Twitter source)would be placed for further analysis
* define the cosumer and access key,token of twitter application(for streaming)
* define keywords to be checked for
* define the properties of Sink
  1. rollsize - maximum file size that can be created in Sink
  2. rollcount – number of files that can be created in Sink
  3. writeformat & file type – type of file format(Text ex:ORC,RC)
  4. path – which is hdfs path(where data streamed from twitter resides)
* define the properties of both source and sink channel

1. type – memory channel hold the data to be passed in memory
2. capacity – number of events stored in channel
3. transactionCapacity – maximum number of events per transaction



Created the directory “tweets” in HDFS for storing the twitter data from flume





Data from twitter is stored in HDFS directory.

