# Flume--Twitter-to-HDFS

Streaming the live data from Twitter to Hadoop Distributed File System(HDFS) using Flume

***FLUME:***

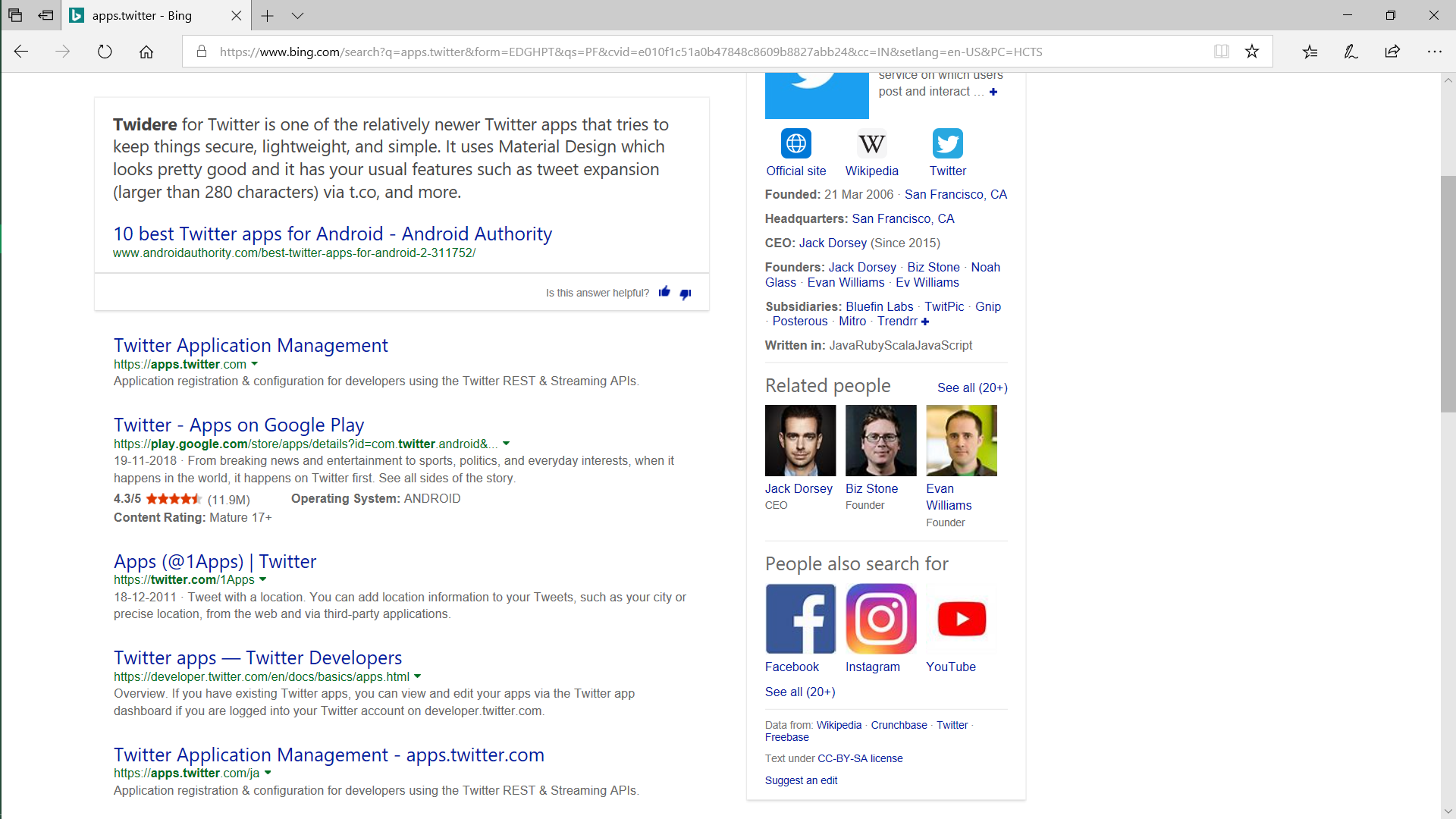
Flume is a tool for streaming the real-time data or live data,we can capture the live data from any sources to HDFS or any other data storages.Flume has capability to capture real-time tweets from Twitter and messages,posts etc from facebook and so on.It works very fast and effective.So in this repository ,i explained how to capture the real-time tweets from twitter and store into HDFS and also shared the code,please check it.Alright lets jump into explanation of this activity.

CREATING TWITTER ACCOUNT :

STEP 1: Create the twitter-app account .Go to below link.

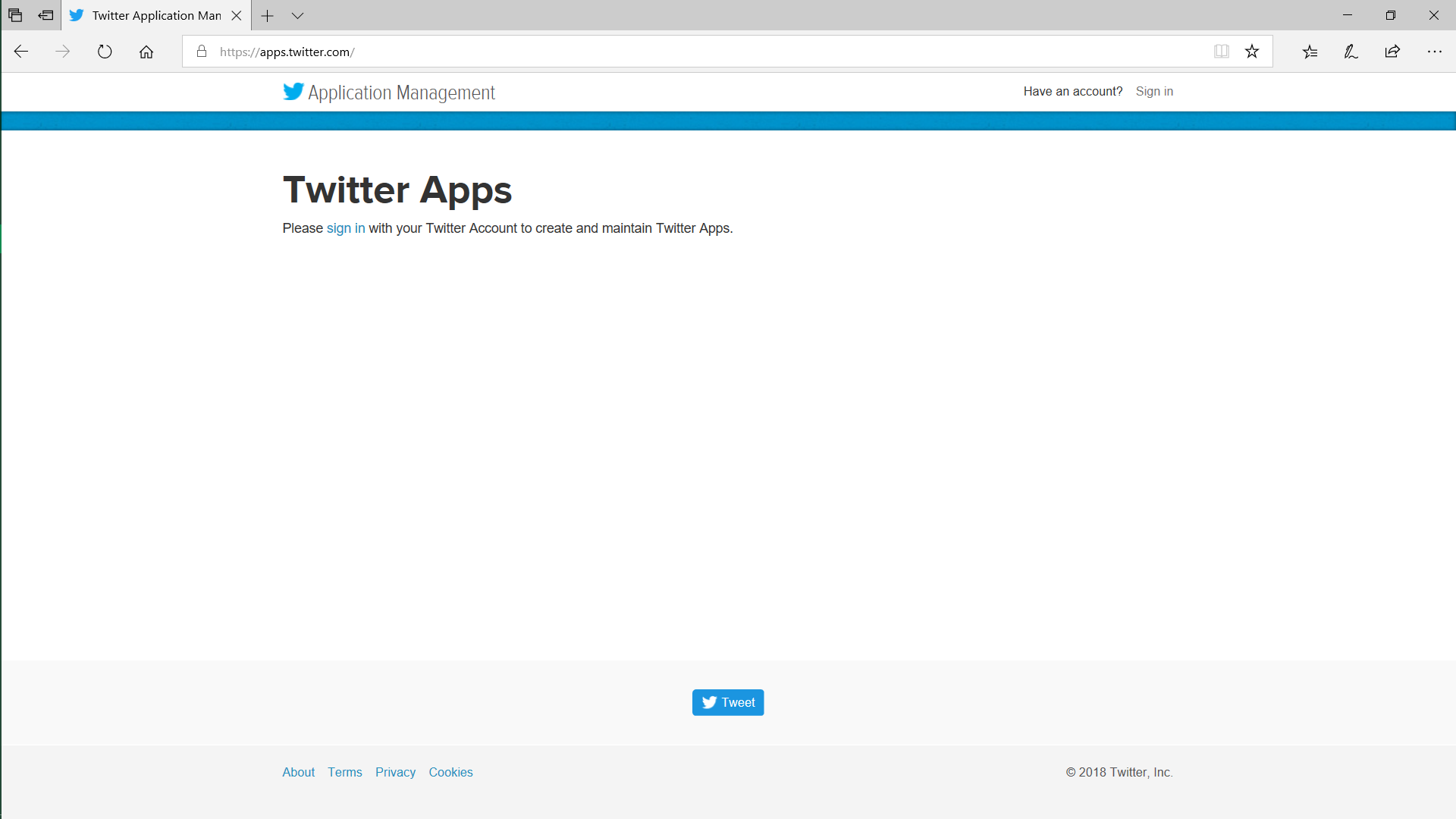
<https://apps.twitter.com/>

you can see as below and click on the highlighted link.



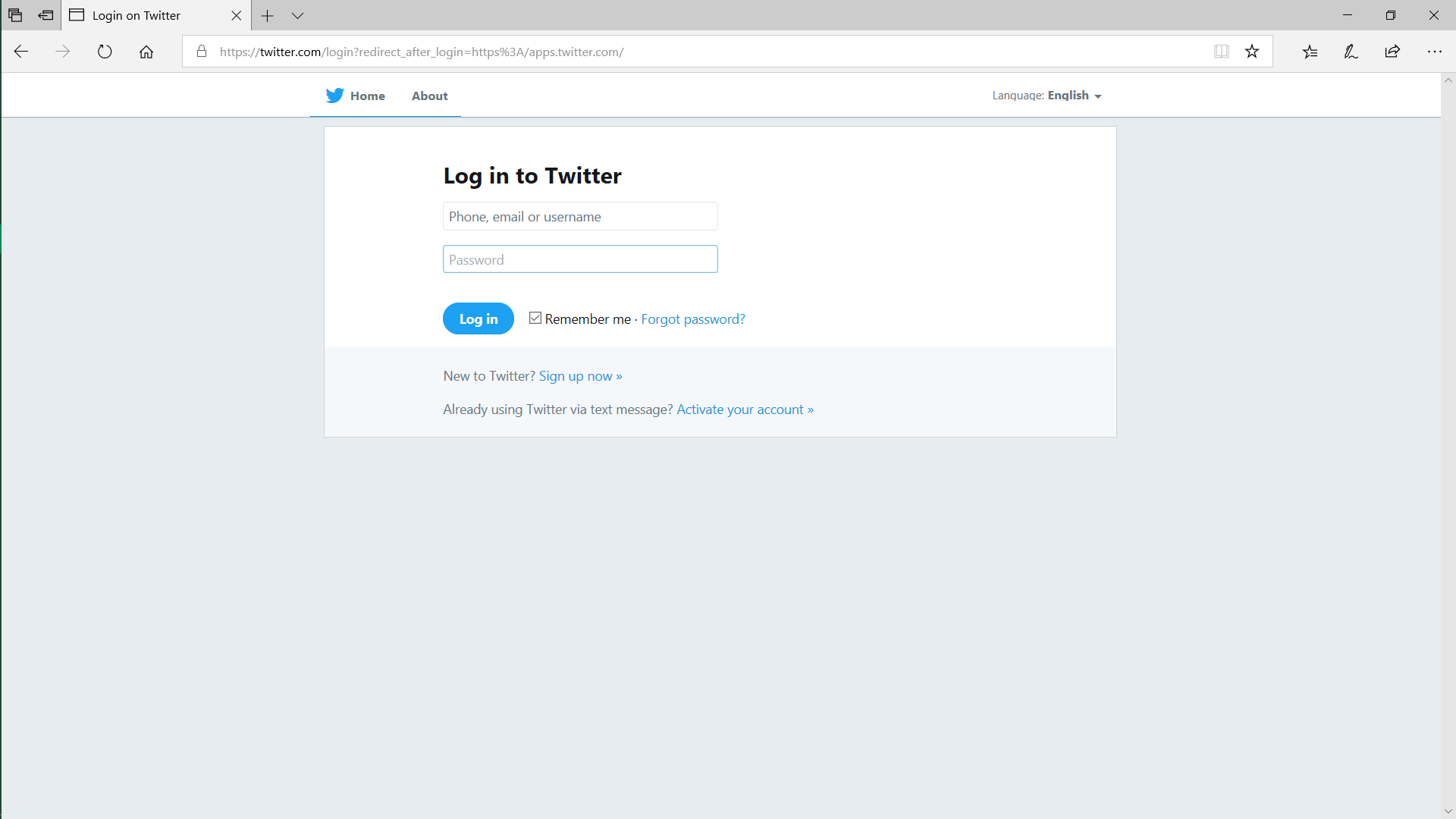


STEP 2: you can see as below and click on the sign in



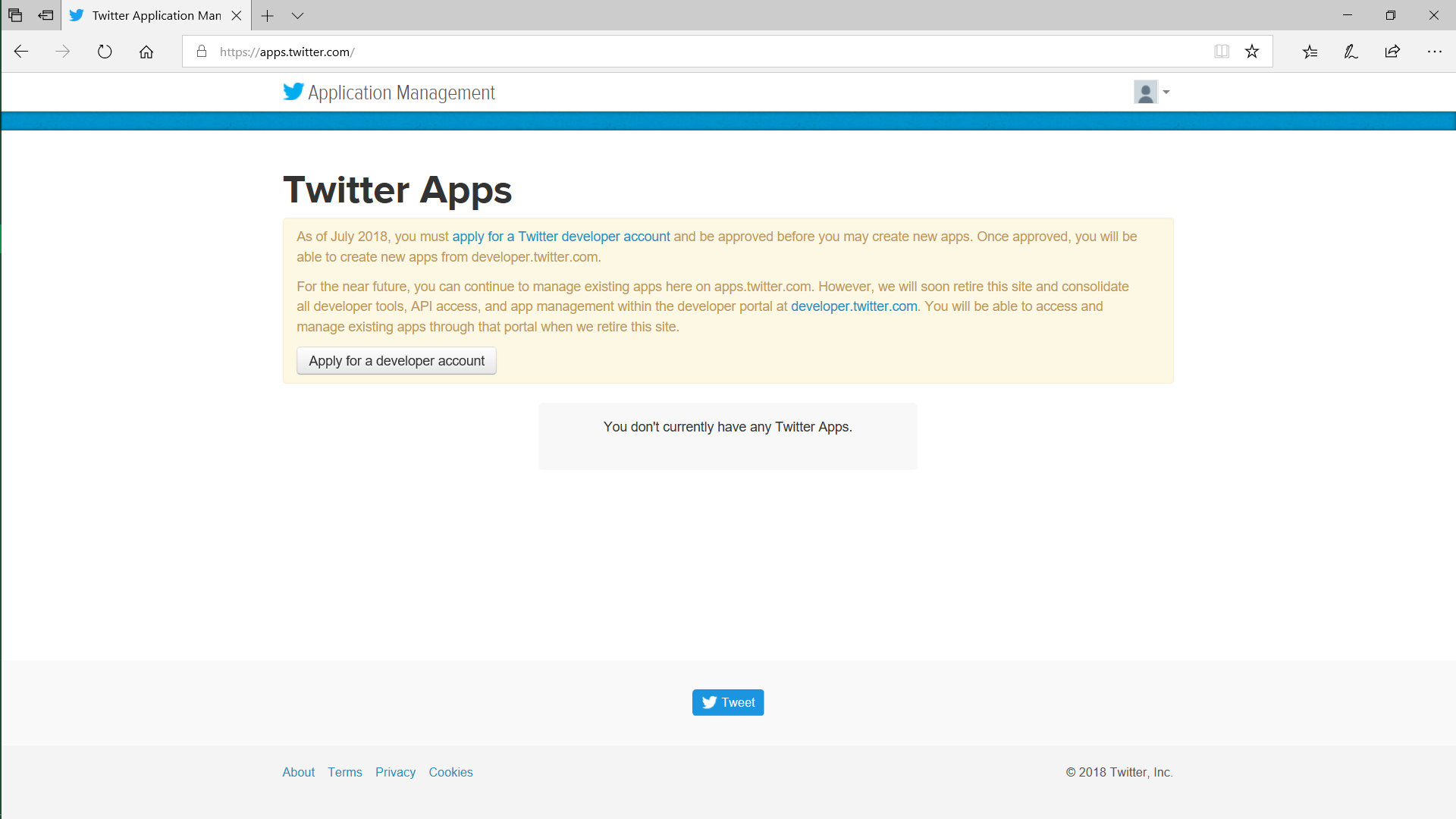


STEP 3: After filling the details ,login with registered “username”,and”password”



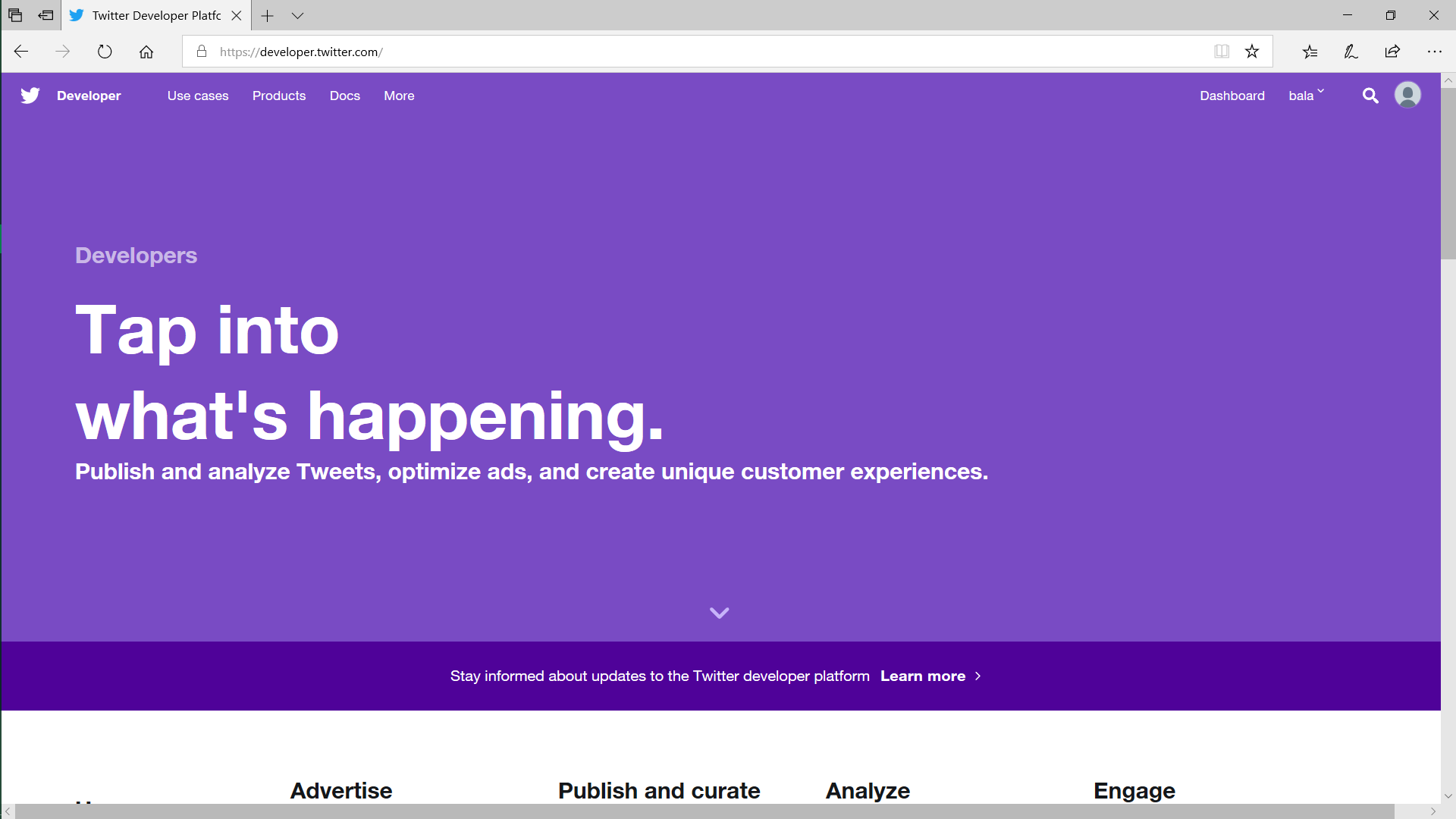


STEP 4: Now click on the developer.twitter.com

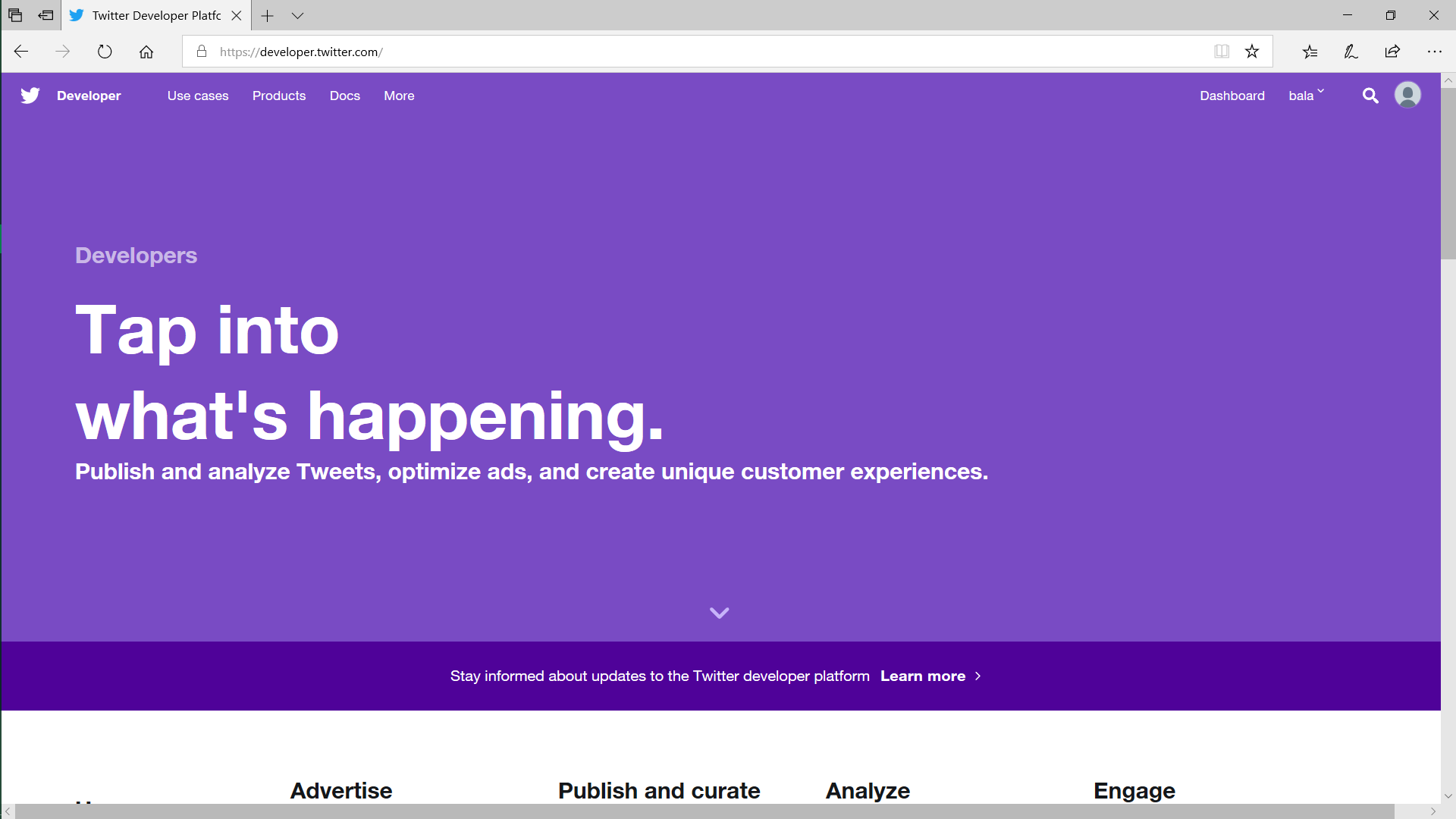




STEP 5: you can see as below

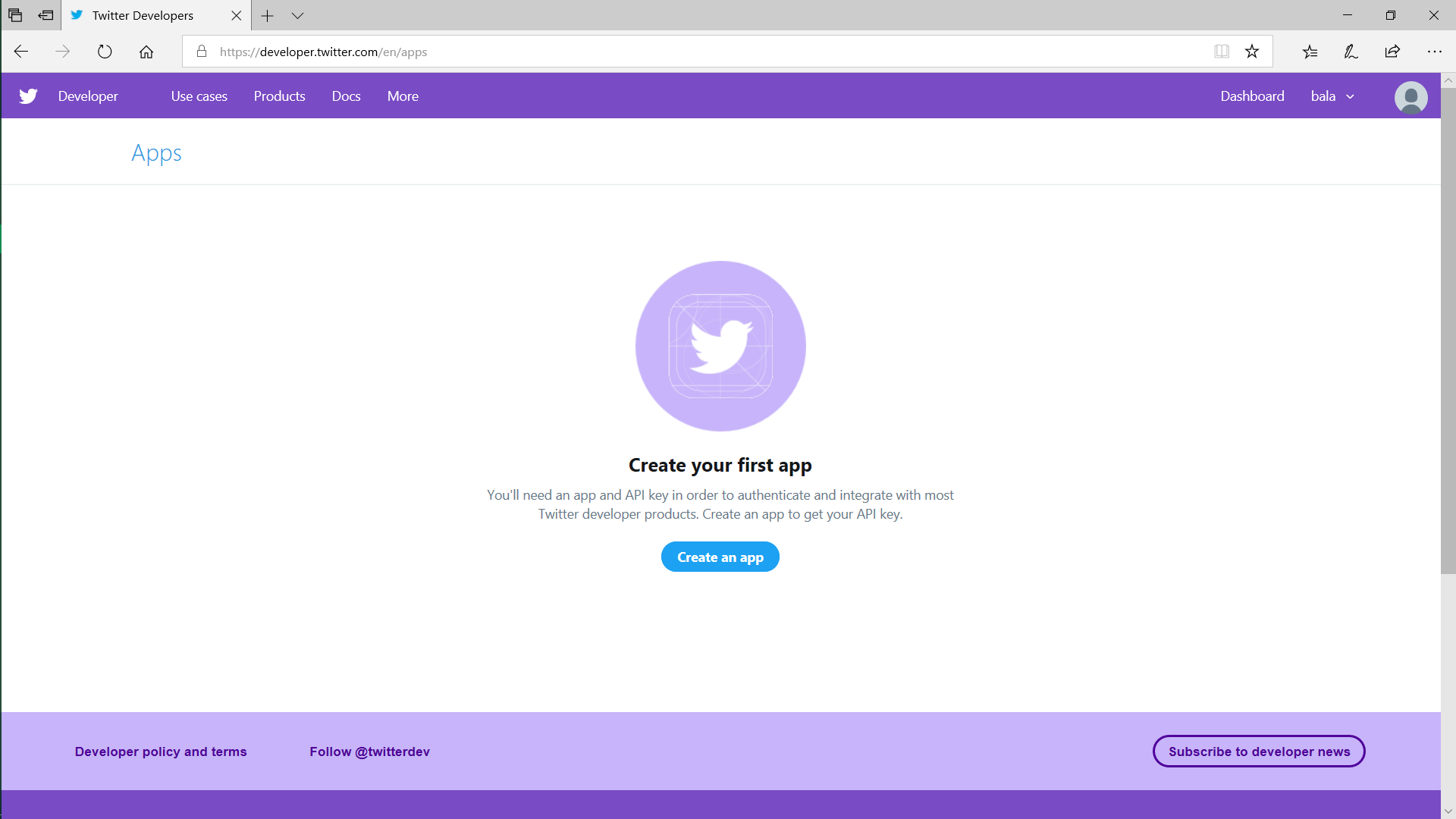


STEP 6: select the highlighted icon and go to “Apps”



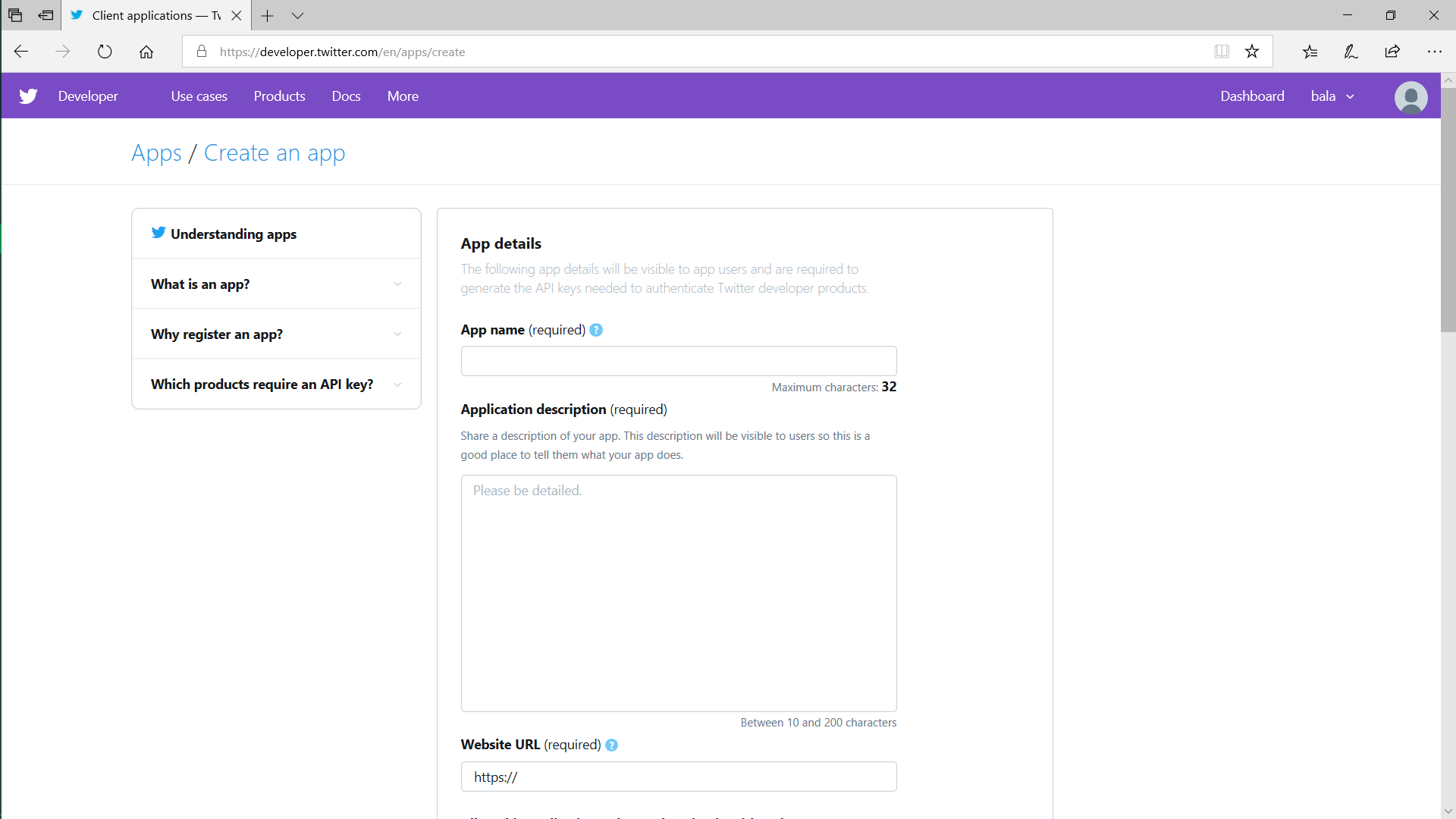


STEP 7: click on the “create an app“ button

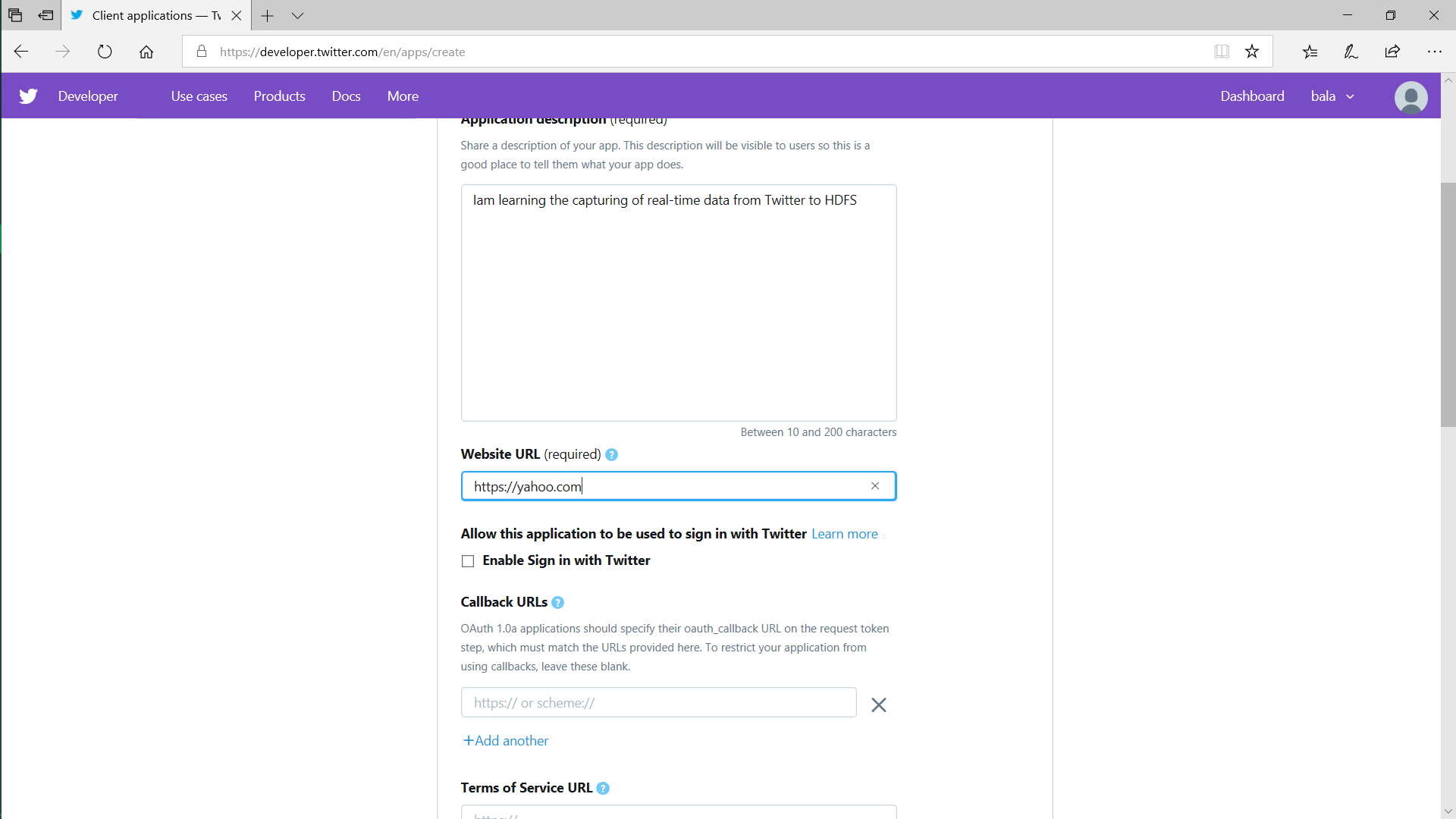




STEP 8: Fill the below given details

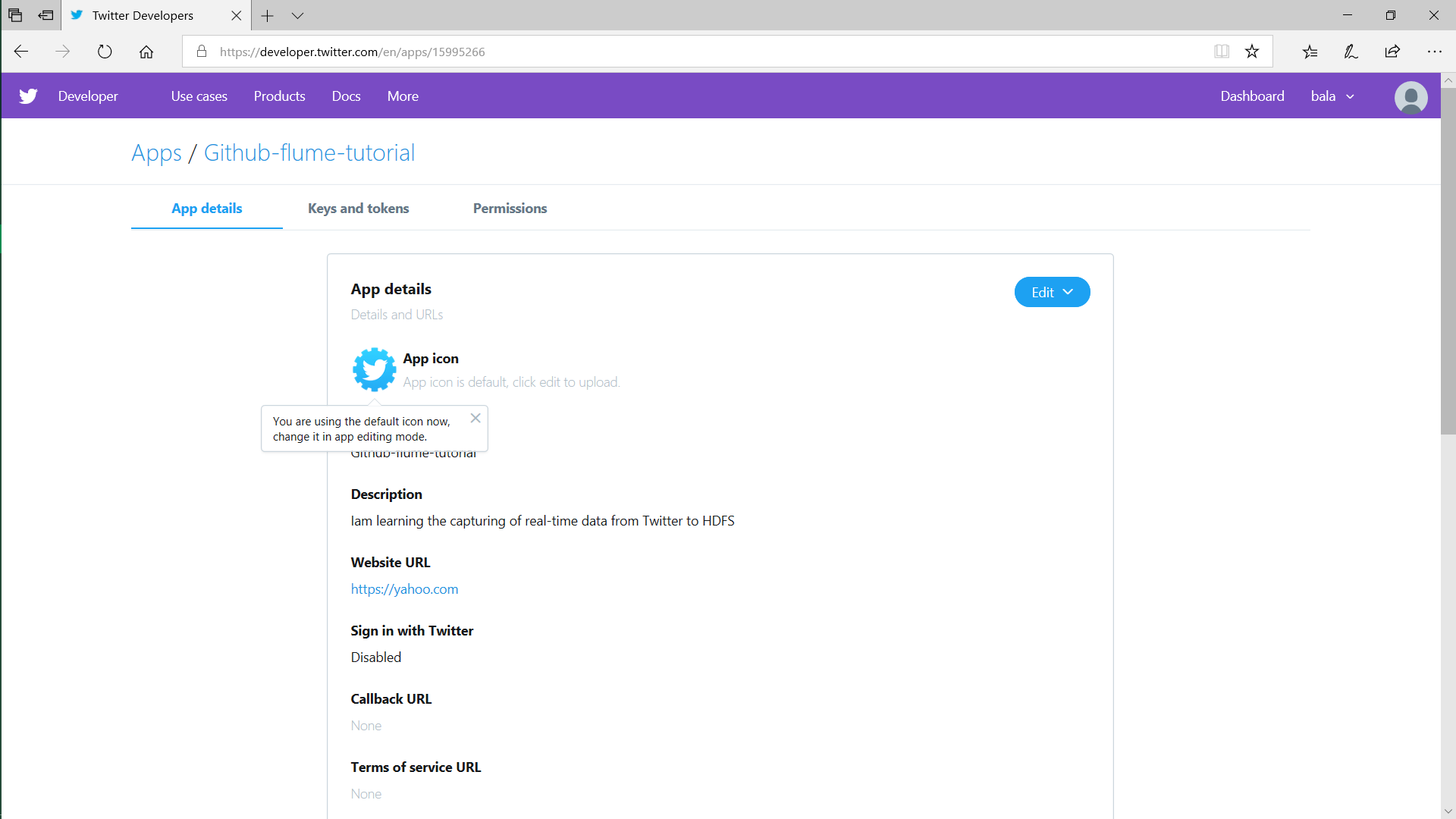


STEP 9: keep “yahoo.com” in textfield as given below because it is placeholder for downloading real-time data into websites, but here we are using our own systems to store data,so no need to give any website link.As it is required field,fill it as “yahoo.com” which is standalone link and click on create button.



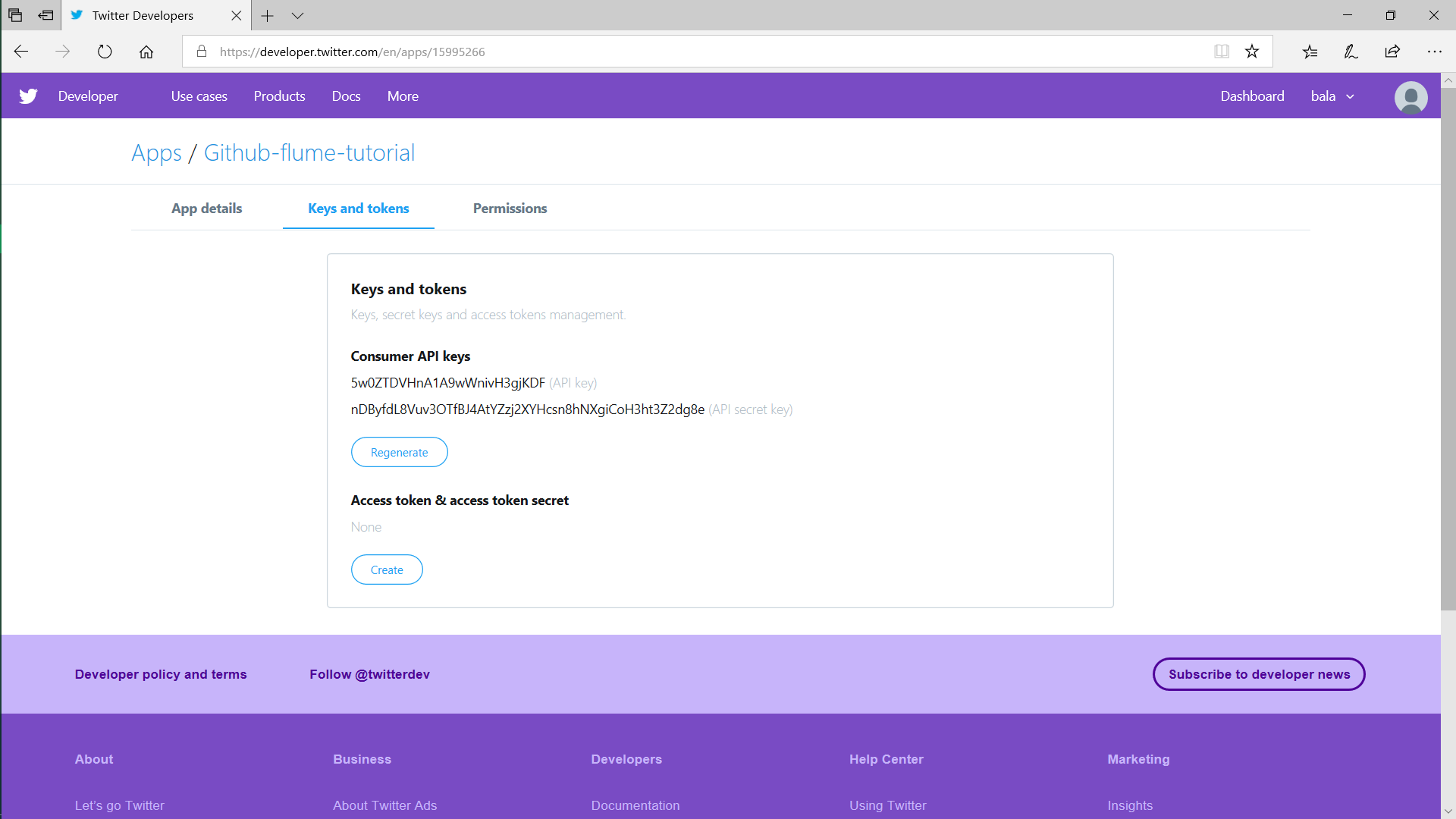


STEP 10: Congrats! you application is created and now go to “Keys and tokens” tab to access the required keys and tokens.



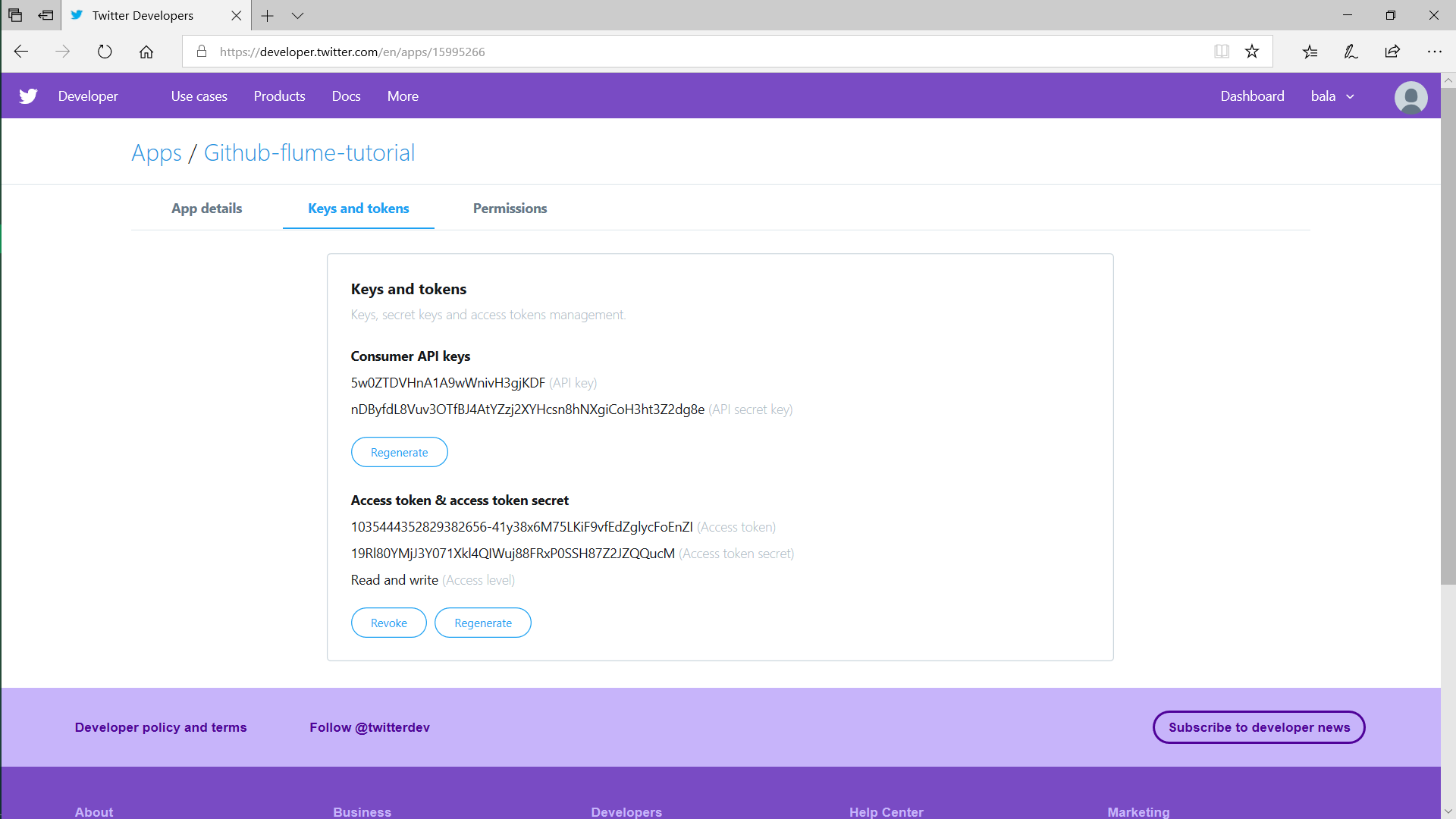


STEP 11: you can see your “API keys” as highlighted below and click on create button to access your “tokens” and save it for further use



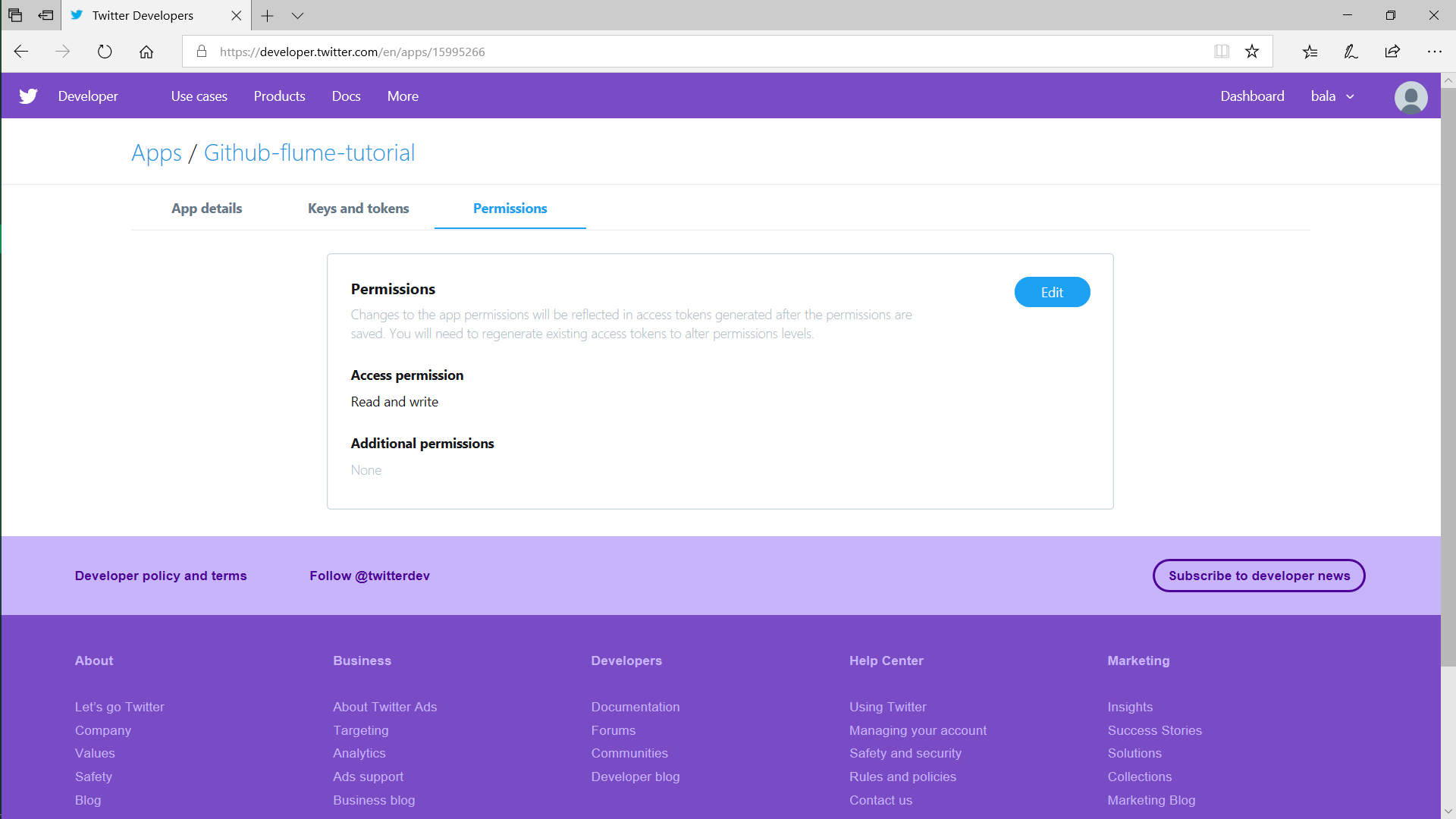


STEP 12: you can see your “tokens” as below and copy the tokens and save it for further use.



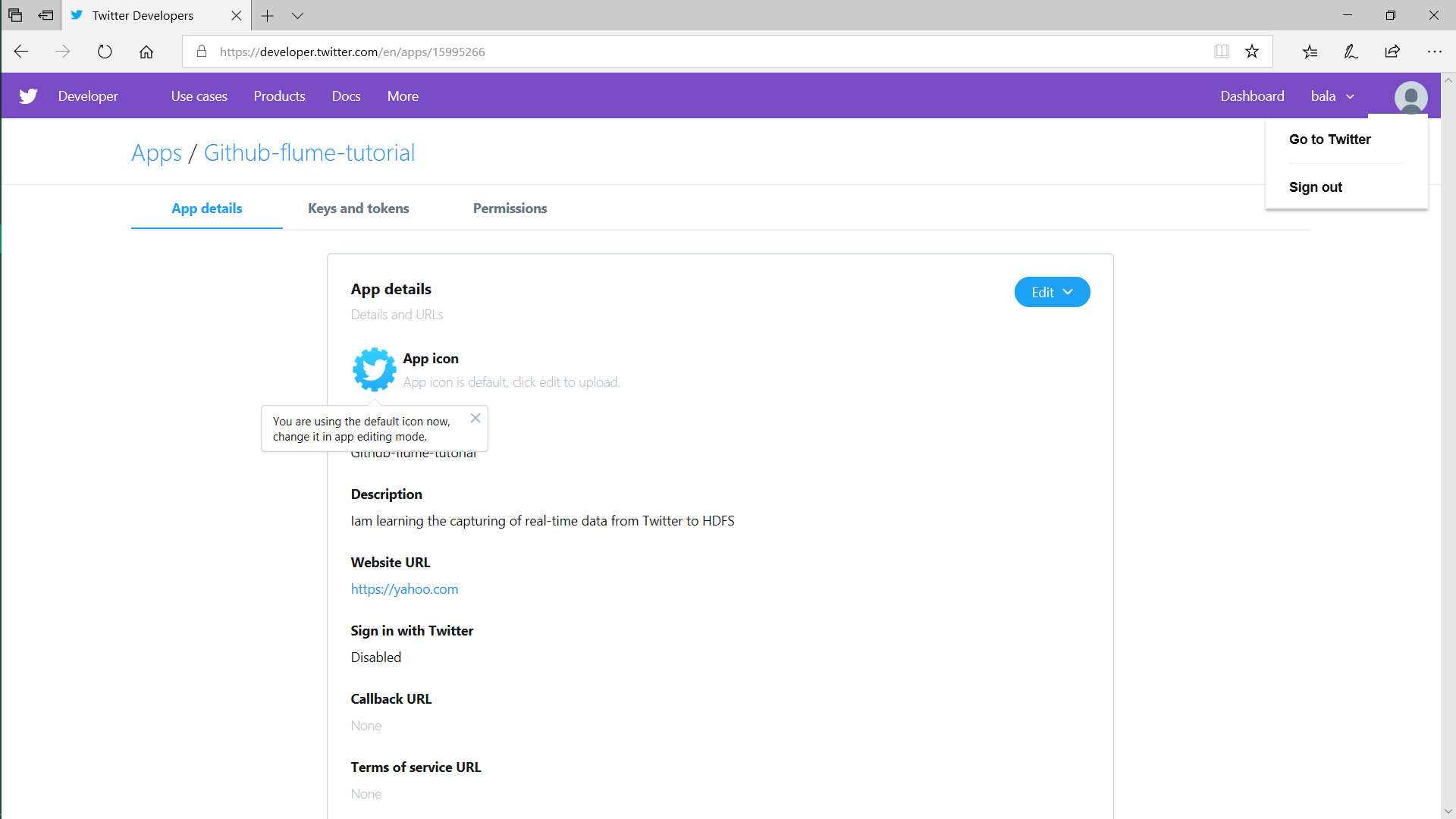


STEP 13: you can change the permissions by clicking on the permissions button(no need)





STEP 14: Signout from twitter (optional)





***CODE EXPLANATION:***

STEP 1: Creating the Source,Sink,Channel

Here TwitterAgent is user defined you can keep any name as you like .source,sink,channel names are also user defined.

TwitterAgent.sources=Twitter

TwitterAgent.sinks=HDFS

TwitterAgent.channels=Memorychannel

STEP 2: Configure the Source

Here “org.apache.flume.source.twitter.TwitterSource” is package , the consumerkey,consumersecret,accesstoken,accesstokensecret are API’s from Twitter ,how to get this keys and tokens are explained above.

TwitterAgent.sources.Twitter.type=org.apache.flume.source.twitter.TwitterSource

TwitterAgent.sources.Twitter.consumerKey=<copy from twitter as shown above>

TwitterAgent.sources.Twitter.consumerSecret=<copy from twitter as shown above>

TwitterAgent.sources.Twitter.accessToken=<copy from twitter as shown above>

TwitterAgent.sources.Twitter.accessTokenSecret=<copy from twitter as shown above>

STEP 3: Configure the Sink

1. This is configuration part of HDFS, hdfs is constant and case sensitive.

TwitterAgent.sinks.HDFS.type=hdfs

1. we have to give path to store data and remember only directory name has to give

TwitterAgent.sinks.HDFS.hdfs.path=testflume

1. we have to mention in what way the data will come like DataStream,BatchProcess etc

TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream

1. Here we have to tell to flume in which format data has to store like Text,Avro etc

TwitterAgent.sinks.HDFS.hdfs.writeFormat=Text

1. Batchsize is how many events has to store for one batch

TwitterAgent.sinks.HDFS.hdfs.batchSize=1000

STEP 4: Configure the Channel

1. Three Types of channels are available “Memory”,”File”,”JDBC”, among three iam using memory channel which is faster than other two.

TwitterAgent.channels.Memorychannel.type=memory

1. The capacity is how many events has to store per batch

TwitterAgent.channels.Memorychannel.capacity=10000

1. The transaction capacity is how many transactions(tweets) has to transfer from twitter .

TwitterAgent.channels.Memorychannel.transactionCapacity=100

STEP 5: Connecting Source to Channel and Sink to Channel

1. Above source,sink,channel configuration is completed successfully, here we have to connect source and channel so that data from twitter is transferred to channel

TwitterAgent.sources.Twitter.channels=Memorychannel

1. This is connecting the channel with sink that is HDFS

TwitterAgent.sinks.HDFS.channel=Memorychannel

STEP 6: Running the Code

Go to “Cloudera Manager” or any other platforms where flume is installed to run the code,then follow below steps:

cd /opt -> cd cloudera -> cd parcels -> cd CDH -> cd bin

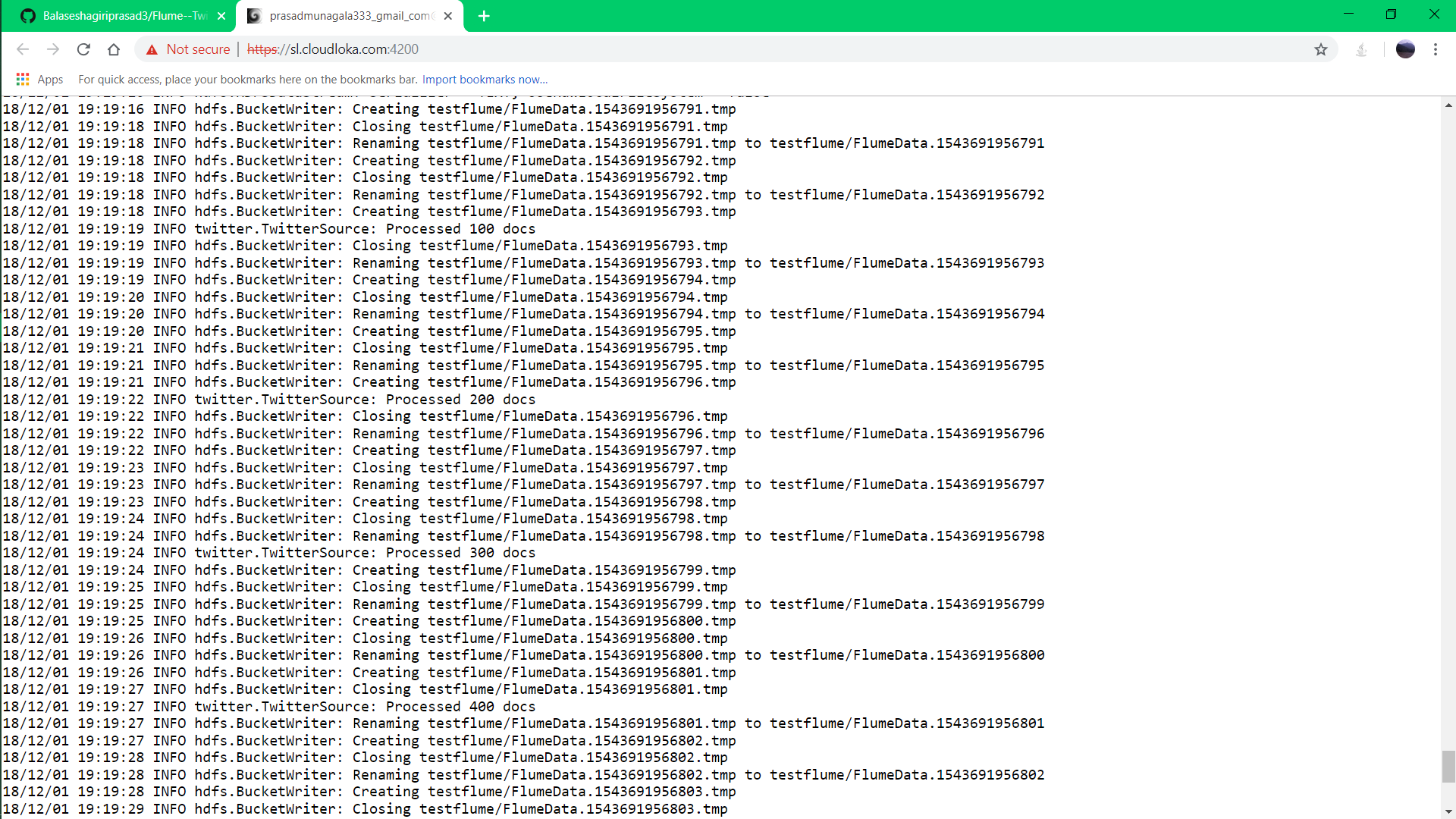
then you have to type the following command:

flume-ng agent --name TwitterAgent --conf githubb --conf-file CODE.conf

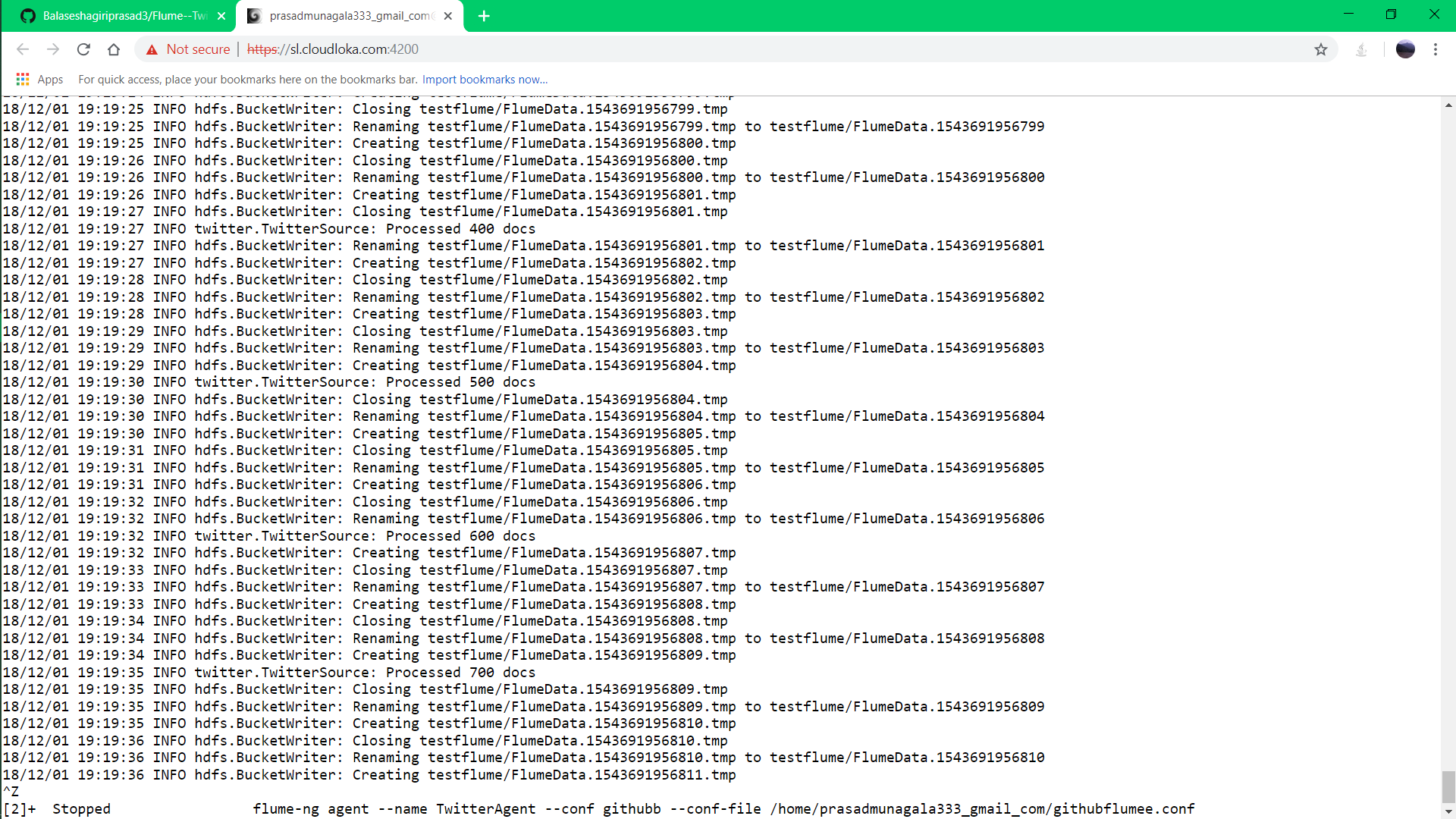
By running above command the tweets from twitter start transferring to HDFS through Memorychannel.

NOTE:The data will transfer continuously from Twitter to HDFS until you the stop, it occupies your system storage completely ,so be careful and press “ctrl+z” on shell to break the streaming after experiencing the real-time data transferring.

OUTPUT:



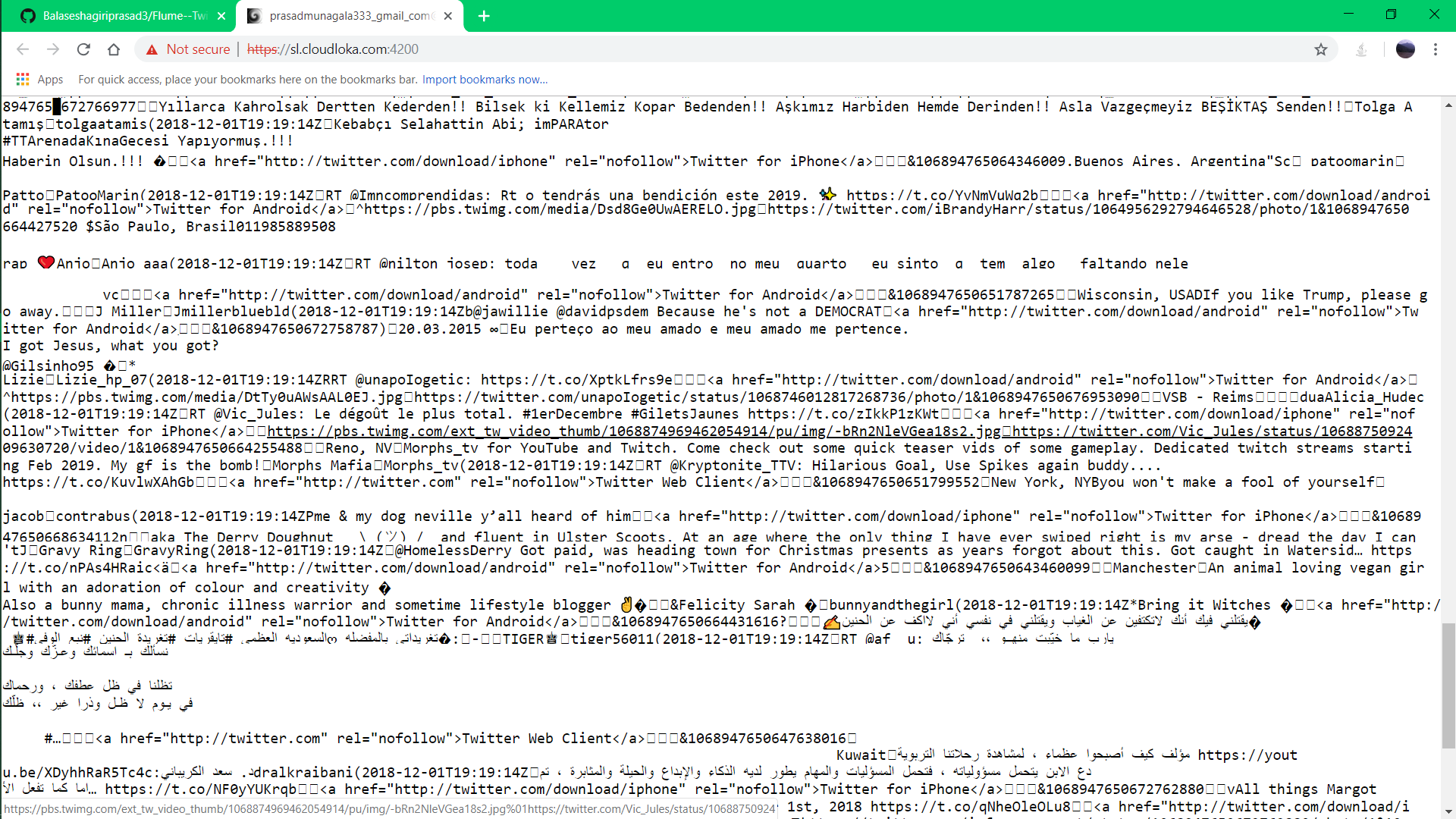






Here you can see stored data in HDFS but the format is not understand, this is because we are using free version of Twitter,that is why the data is not in understandable format,for this we have to pay amount and get API’s from twitter and then the data is in understandable format and also we can get data relate to some topics like football,cricket,machine learning etc











I think this is helpful to you, if you like this, follow me on “Github”.

**THANK YOU**