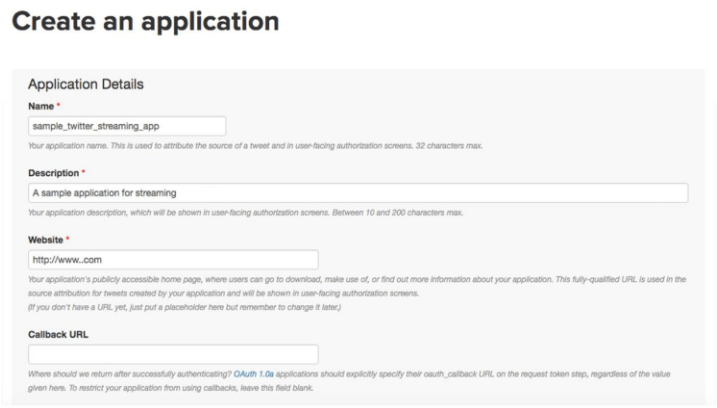
MD FAKRUL ISLAM (613839)

I will show how to build a simple application that reads online streams from Twitter using Python, then processes the tweets using Apache Spark Streaming to identify hashtags.

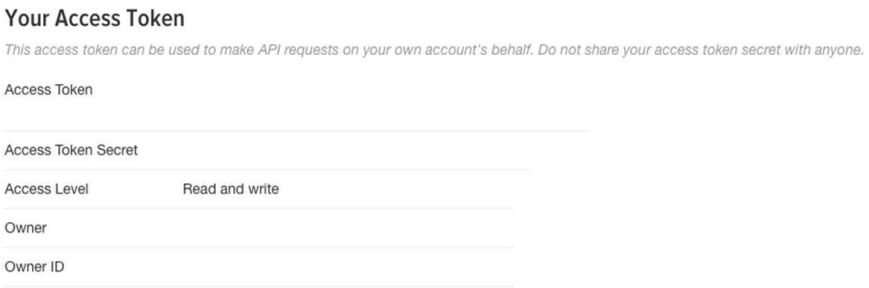
1. Creating Credentials for Twitter APIs

To get tweets from Twitter, I registered on **TwitterApps** by clicking on “Create new app” and then fill the below form click on “Create your Twitter app.”



2. Second, I went to my newly created app and opened the “Keys and Access Tokens” tab. Then click on “Generate my access token.”

The new access tokens will appear below.



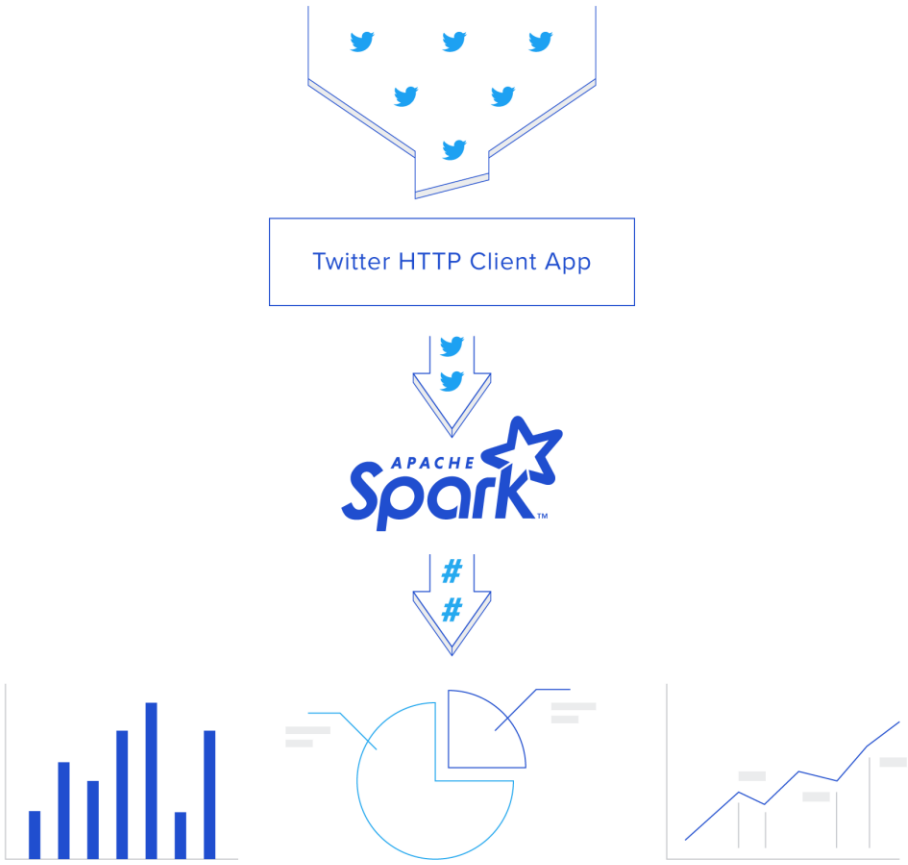
And now I am ready for the next step.

3. Building the Twitter HTTP Client

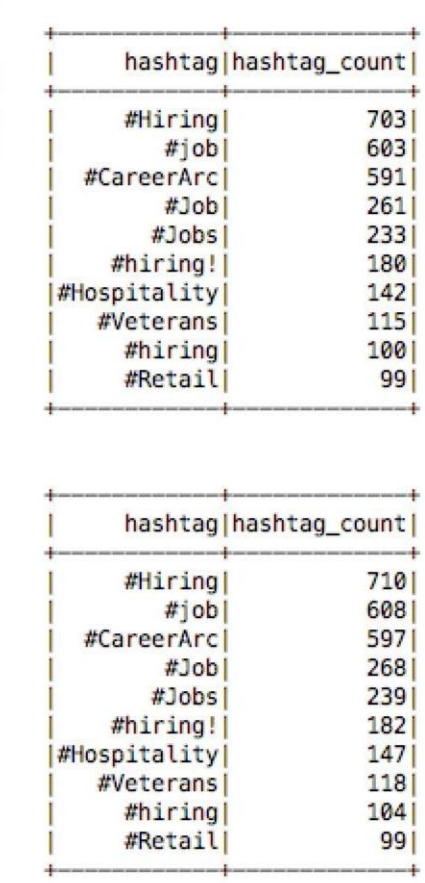
In this step, I build a simple client that will get the tweets from Twitter API using Python and pass them to the Spark Streaming instance.

4. Setting Up Our Apache Spark Streaming Application

I made Spark streaming app that will do real-time processing for the incoming tweets, extract the hashtags from them, and calculate how many hashtags have been mentioned. I configured all those in my azure account.



Finally, here is a sample output of the Spark Streaming while running and printing the hashtag_counts_df, the output is printed exactly every two seconds as per the batch intervals.



hashtag	hashtag_count
#Hiring	703
#job	603
#CareerArc	591
#Job	261
#Jobs	233
#hiring!	180
#Hospitality	142
#Veterans	115
#hiring	100
#Retail	99

hashtag	hashtag_count
#Hiring	710
#job	608
#CareerArc	597
#Job	268
#Jobs	239
#hiring!	182
#Hospitality	147
#Veterans	118
#hiring	104
#Retail	99