

MD FAKRUL ISLAM (613839)

Master's in computer science, Data Science Track

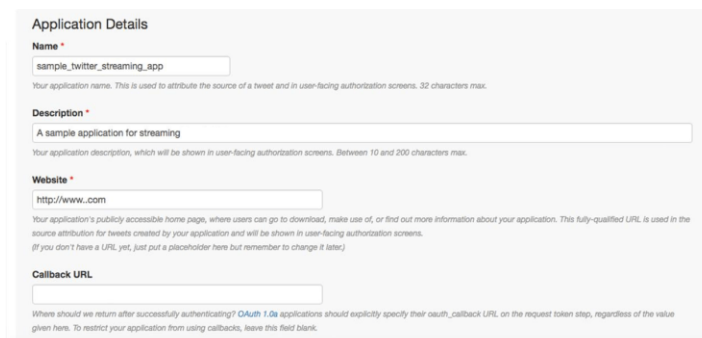
Maharishi International University

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.”

#### Create an application

A screenshot of the 'Create an application' form on the Twitter Developer Portal. The form is titled 'Application Details' and contains four main sections: 'Name', 'Description', 'Website', and 'Callback URL'. Each section has a text input field and a small explanatory note below it. The 'Name' field contains 'sample\_twitter\_streaming\_app'. The 'Description' field contains 'A sample application for streaming'. The 'Website' field contains 'http://www.com'. The 'Callback URL' field is empty. The form is styled with a light gray background and white text for labels and notes.

### 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.

### Your Access Token

*This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.*

Access Token

Access Token Secret

Access Level

Read and write

Owner

Owner ID

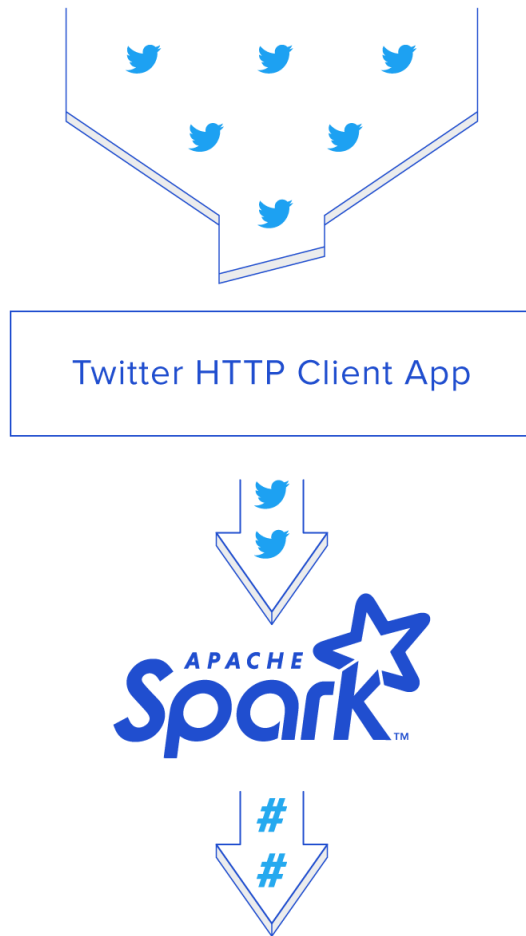
And now I am ready for the next step.

### 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