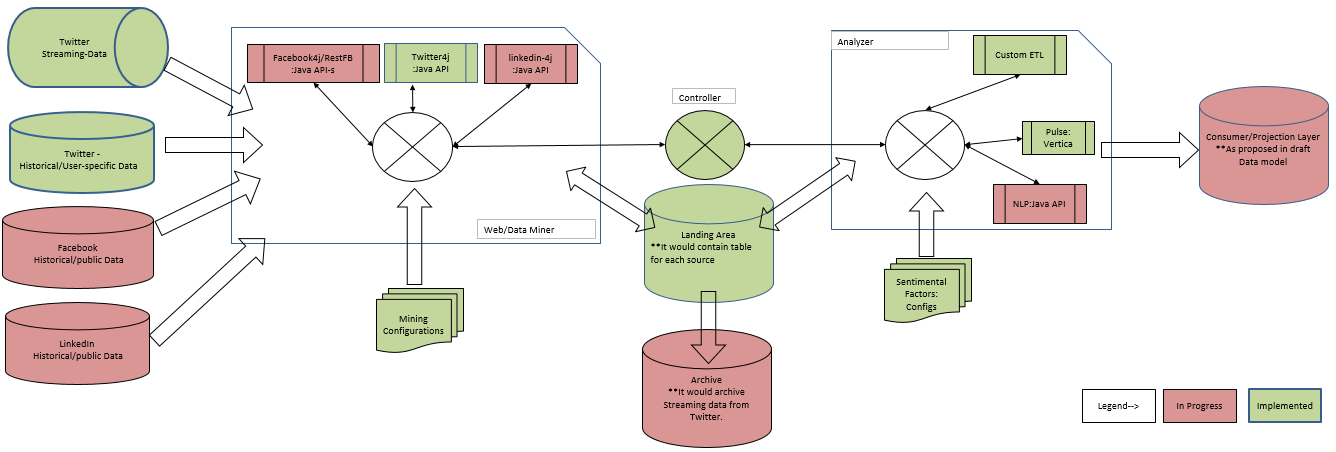
Social Media Data Mining and Sentimental Analysis Framework

## Architecture



## How do I start it?

>cd /home/dbadmin/DataMiner/dist

> java -jar DataMiner.jar –Default

[or]

>java -jar DataMiner.jar –Auth <Configuration File with full path> -Log <directory to store the log files>

## How do I restart it?

The restart mechanism will be same as initiation. After finding the root cause of failure execute the following command to restart it.

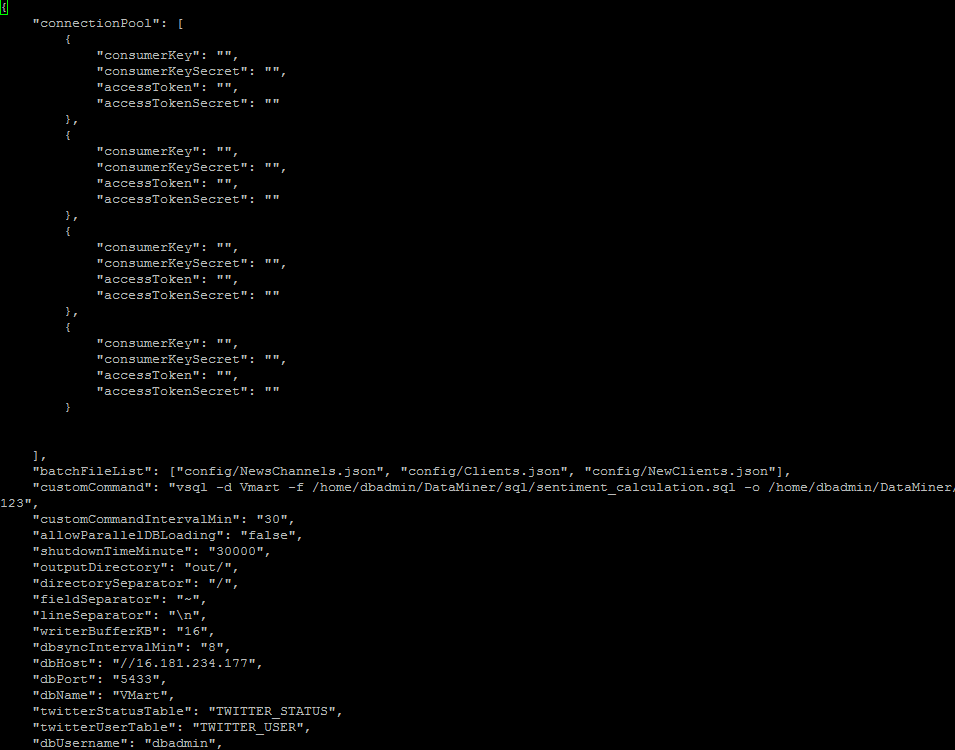
>cd /home/dbadmin/DataMiner/dist

> java -jar DataMiner.jar –Default

## What are Application specific configuration ?:

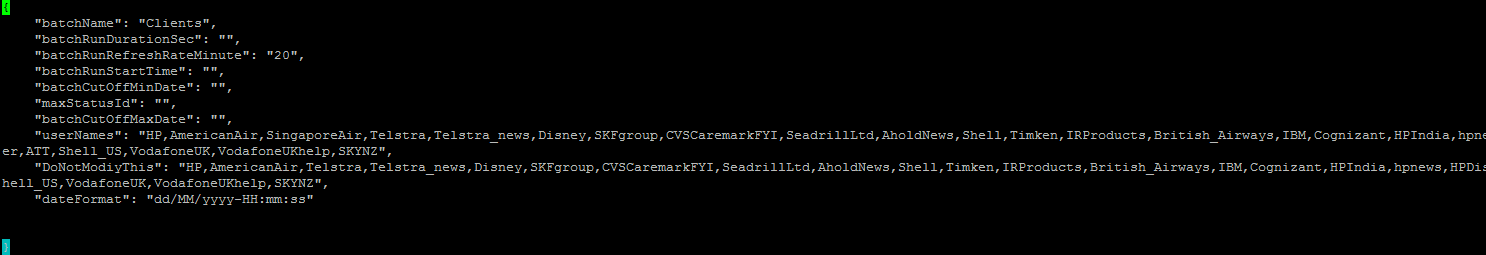
>vi config/AppConfig.json

Following file contains the application specific configurations:



## What is batch File and what are batch specific configuration?

The batch is nothing but the set of configuration for particular users that has similar configuration like (schedule time, refresh interval, cut off dates for data etc.)



## How do I know if it is functioning properly?

>vi log/DataMiner\_2014xyz.log

--This log file logs each debug information as well as error details (if any) for monitoring purpose.

## Where do I check intermediate data files generated before loading to vertica? How do I manually load these files if required?

>ls –ltr log/

--The tables of vertica are loaded from the temp output files created in this directory.

--This application logs the COPY command that it tries to execute on target DB in log file. So if there are some DB failures one can go to log file, copy the “COPY command for vertica” and execute it on vsql.

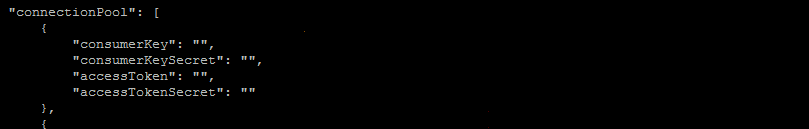
--In case the these commands are not loaded one can construct the COPY command for files in output directory with delimiters from configuration file and load execute them on vsql

## What is Rate-limit for twitter, how is it being handled in this module and how do I enhance it?

-- Twitter restricts the number of connection for a user to avoid maintain server hit. Fortunately twitter  has designed this restriction algorithm for [specific operations](https://dev.twitter.com/docs/rate-limiting/1.1/limits)(click to read more). For instance, If try to create more than 180 connection to search users per 15 minute I will be blocked for next 15 minutes, but at the same time I will be allowed to use the same connection for other operations like searching status (that again has limit of 180 connection per 15 minute).

-- We have created an algorithm to efficiently handle this limitation at some extent. For that we include multiple twitter connection with API authentication details, and application will be continuously monitoring the rate limit of status each API connection for required operation and switches to un-blocked connection if required.

--To increase the efficiency, we can make an agreement with twitter or we can keep on registering the developer account on Twitter and use generated API key-s as shown in below section:



## How do I schedule a new batch?

--A new batch cannot be included while running the application. A new batch configuration file has to be created with schedule start time and interval, this batch file needs to be added to application configuration file and Application can be restarted.

--As of now this application has two batch files.

NewsChannels: This captures the data for famous news channels to get the current trend

* CNN
* CNN-breaking news
* BBC Breaking News
* BBCWorld
* Times now
* New York times
* Ndtv
* Washington post
* Fox News
* Fox news Headlines

Clients:

* HP
* FlyFrontier
* AmericanAir
* VirginAmerica
* SingaporeAir
* united
* Telstra
* Telstra\_news
* Disney
* SKFgroup
* CVSCaremarkFYI
* SeadrillLtd
* AholdNews
* Shell
* Timken
* IRProducts
* British\_Airways
* IBM
* Cognizant
* HPIndia
* hpnews
* HPDiscover
* ATT
* Shell\_US
* VodafoneUK
* VodafoneUKhelp
* SKYNZ
* sbm\_bank
* BoM\_Mauritius
* BofA\_News
* BofA\_Help
* bankofengland
* bankofcanada
* Barodabank
* TheOfficialSBI
* BOI\_Tweets
* ICICIBank
* ICICIBank\_Care
* HDFC\_Bank

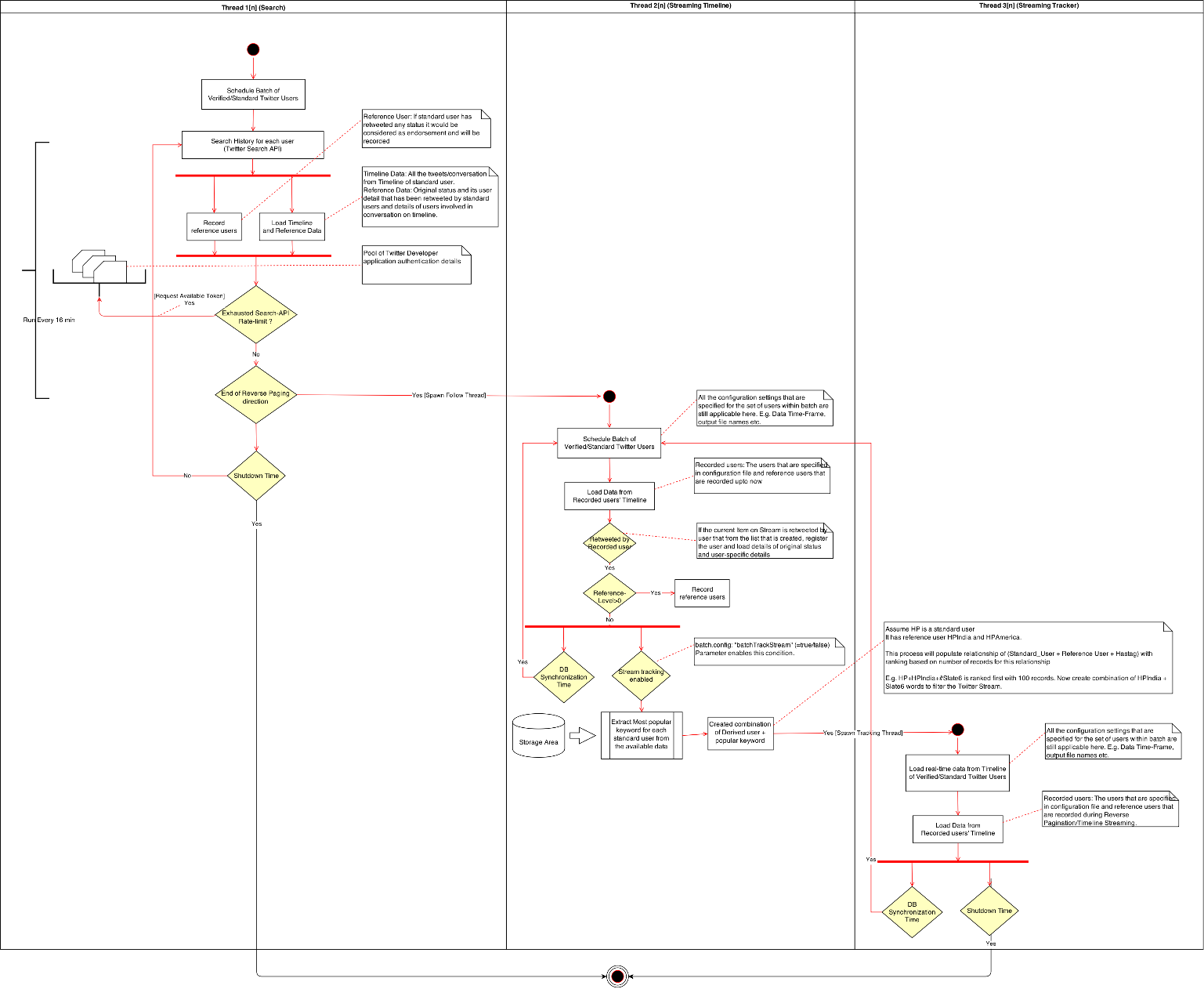
## What approach are we using to get this details?

--Assume I am given a  huge book to be read and told catch with page ‘x’. Since it is not wise to read it from start as I can never catch up within given time, but I also need to get some idea on what is going on in page ‘x’ so I would read  only few pages before page ‘x’ and then will continue reading all the pages after that.

--In our case the starting point of our data capture is page ‘x’ (This will be indicated with INITIAL PAGING direction in log file for each user in each batch )but I would like to get some data from history that is 1 week (This will be indicated with REVERSE PAGING direction in log file for each user) and then finally application will keep on capturing data in FORWARD PAGING direction.

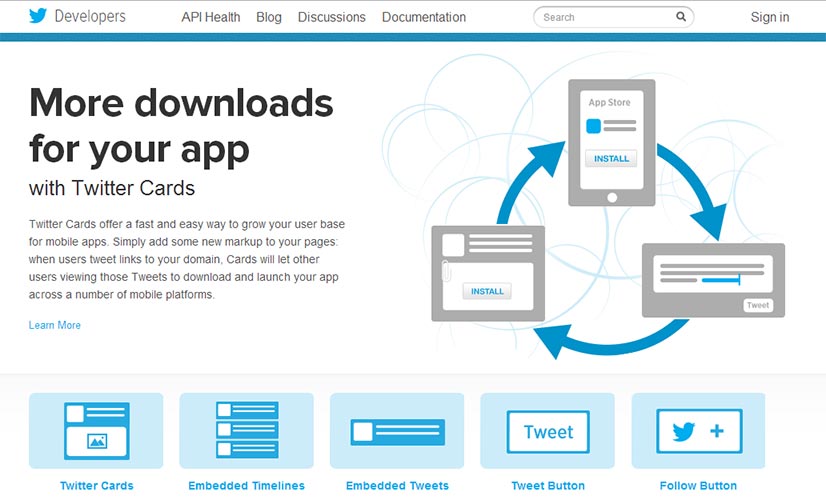
--To achieve each of these paging methodology I am using status Id as an  indicator of Twitter data set (page number in case of book example). To more get details on how each paging direction works in application go through the link [Twitter Paging algorithm](https://dev.twitter.com/docs/working-with-timelines)(click to read more).

## What is the algorithm that we are using?



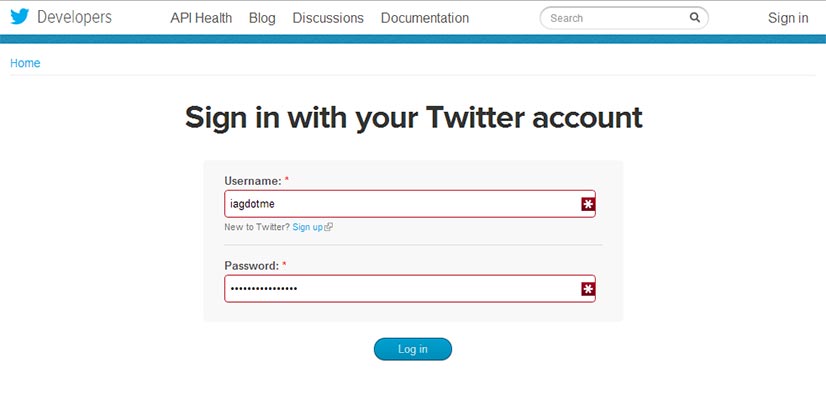
## How do I get Twitter Developer Account?

**Step 1** Visit the Twitter Developers’ Site



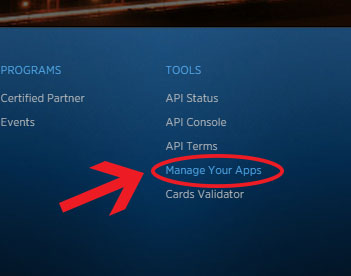
The first thing you need to do is head on down to [dev.twitter.com](http://dev.twitter.com/). In order to create an account, all you need to do is click on the “Sign In” link at the top right.

**Step 2** Sign in with your Twitter Account



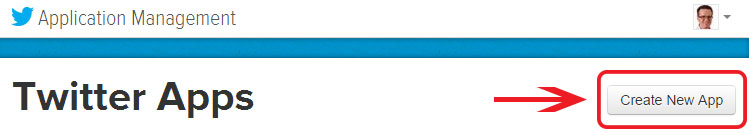
Next, sign in with the Twitter account you want to associate with your app. You do have a Twitter account don’t you?!

**Step 3** Go to apps.twitter.com



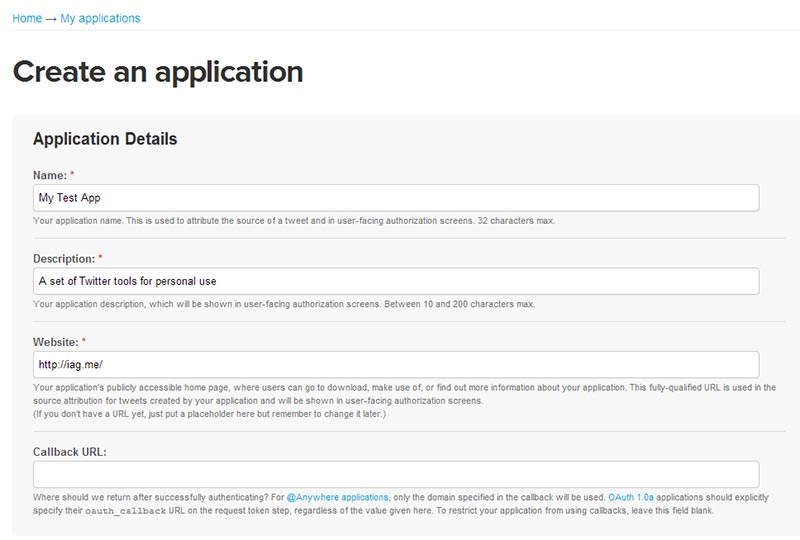
Once you’re logged in, you will need to visit Twitter’s app website. This can be found at [apps.twitter.com](http://apps.twitter.com/) (it can also be found by clicking “manage your apps” in the footer of the Twitter Developers site.

**Step 4** Create a New Application



If you are new to the Developers site you won’t see any applications registered. Either way, it’s time to create our first application. To do this, click on the big “Create a new application” button.

**Step 5** Fill in your Application Details



It’s been easy so far, but this screen often makes people confused. Don’t worry, it’s a lot easier than you think…

**Name:**  
First of all you need to give your app a **unique name** (one that no one else has used for their Twitter app). Since we’re going to be an creating app for personal use and not one that other people can register and use, just put your domain name in or perhaps even your name.

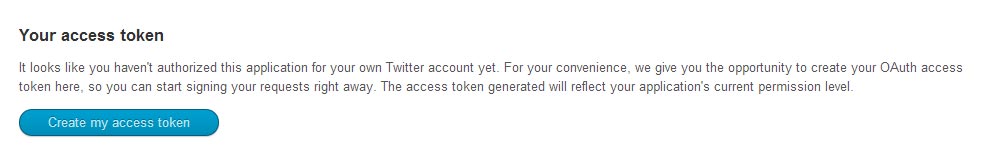
**Description:**  
You don’t have to worry much about the description- you can change this later, but it can help to distinguish this app from others that you might create in the future. I’ve put “a set of Twitter tools for personal use”.

**Website:**

Put **your website** in the website field. It’s supposed to be your application’s publicly accessible home page. However, you can provide [www.hp.com](http://www.hp.com) here .

**Callback URL:**  
 I would **ignore the Callback URL field**.

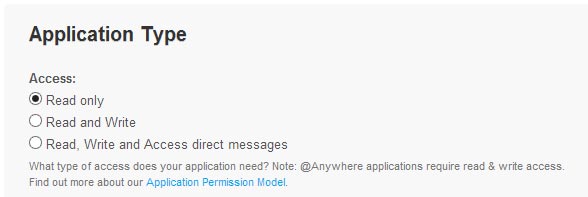
**Step 6** Create Your Access Token



You’ll then be presented with lots of information, but we’re not quite done yet. We now need to authorise the Twitter app for your Twitter account. To do this, click the “Create my access token” button. This takes a few seconds, so if you don’t see the access tokens on the next screen, you may have to refresh the page a few times.

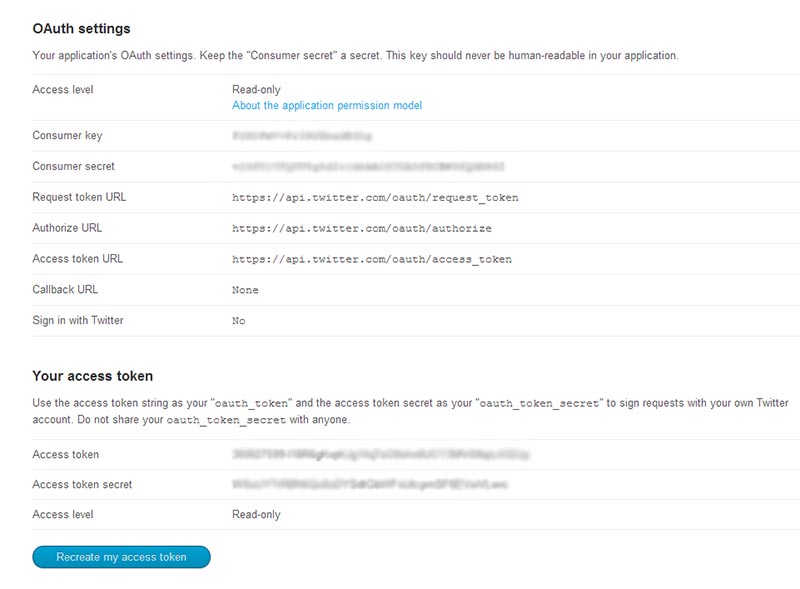
This access token will allow your Twitter application to read Twitter information. You’ll be able to get data of your tweets, mentions, lists and more. However, if you want to do more advanced stuff like sending tweets or deleting, you’ll need to change your access type.

**Step 7** Choose what Access Type You Need



As I mentioned, the default access type when you create a Twitter app is read only. My recommendation is to keep it like this unless you really want to be able to have the app delete or write data (for example sending tweets or deleting direct messages). For our first Twitter app we will only need read only access, but so that you know for future, in order to change the access type, you need to click on the settings tab at the top of the screen. Once you do this, scroll down to where it says “Application Type”. Change the access type to “Read and Write” to be able to read Twitter data and send tweets and select “Read, Write and access direct messages” if you want to also have access to your direct messages.

**Step 8** Make a note of your OAuth Settings



Once you’ve done this, make a note of your OAuth settings. You can copy following configuration to appConfig.json file:

* Consumer Key
* Consumer Secret
* OAuth Access Token
* OAuth Access Token Secret

It goes without saying that you should keep these secret. If anyone was to get these keys, they could effectively access your Twitter account.