StreamSets Pipeline Tutorial

Authors: Mohsen Alam; Randall Lunetto
Instructor: Jongwook Woo
Date: 01/22/2021

Lab Tutorial

Mohsen Alam (malam@calstatela.edu)

Randall Lunetto (rlunett@calstatela.edu)

01/22/2021

Create and Configure a Pipeline on StreamSets using Twitter API

Objectives

In this hands-on lab, you will learn how to:

- Create and configure a pipeline that uses Twitter's API live stream filtered Tweets
- Store the Tweets in real-time to your local storage file system for future analysis

Platform Spec

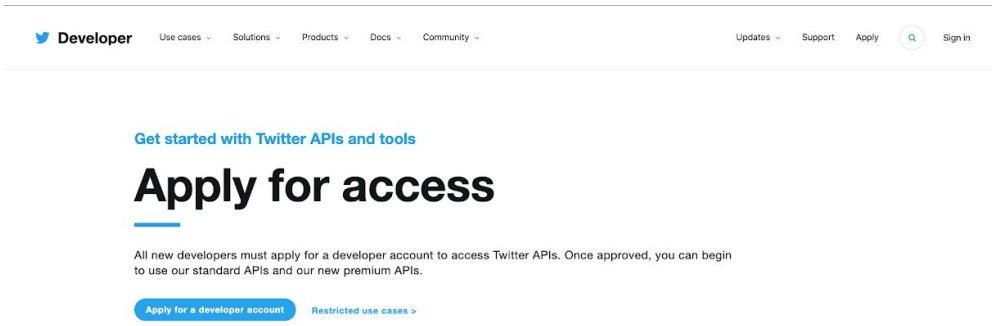
- Windows 10
- CPU Speed: 2.38 GHz
- # of CPU cores: 1

- # of nodes: 1
- Total Memory Size: 16 GB

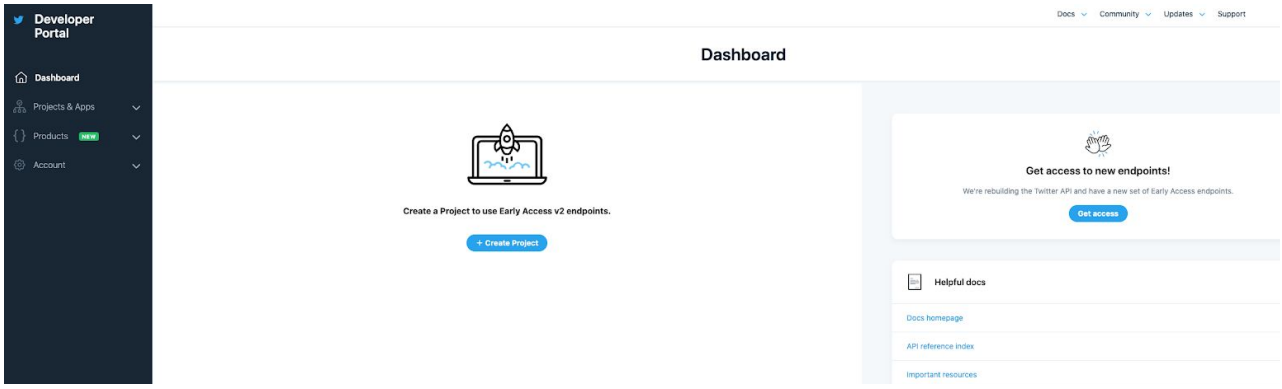
Step 1: Obtain Twitter API Keys

Before you can begin building your pipeline, you'll need to obtain your Twitter API keys first. Be sure to use your school email address and apply as a 'Student' for instant access.

1. Apply for Twitter API keys using your school email at https://developer.twitter.com/en/apply-for-access

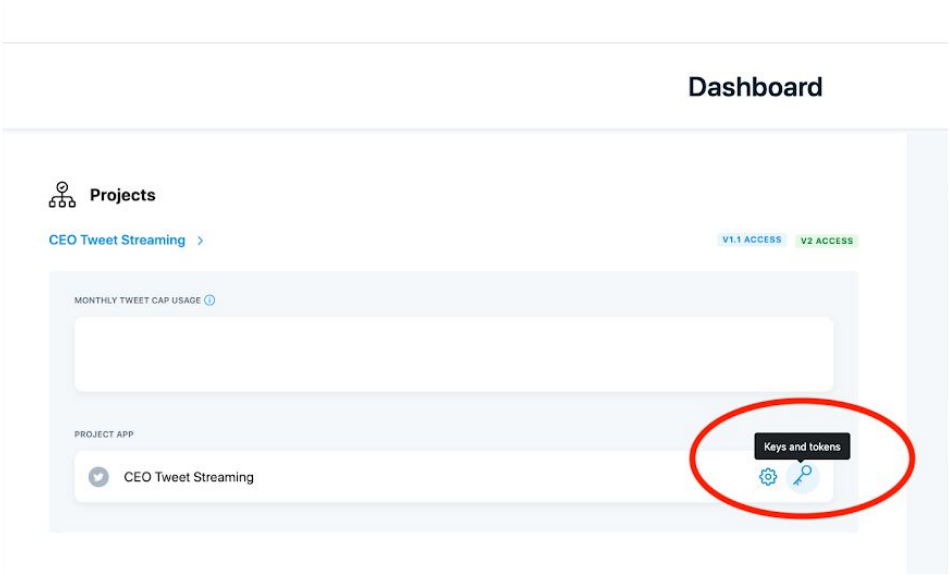


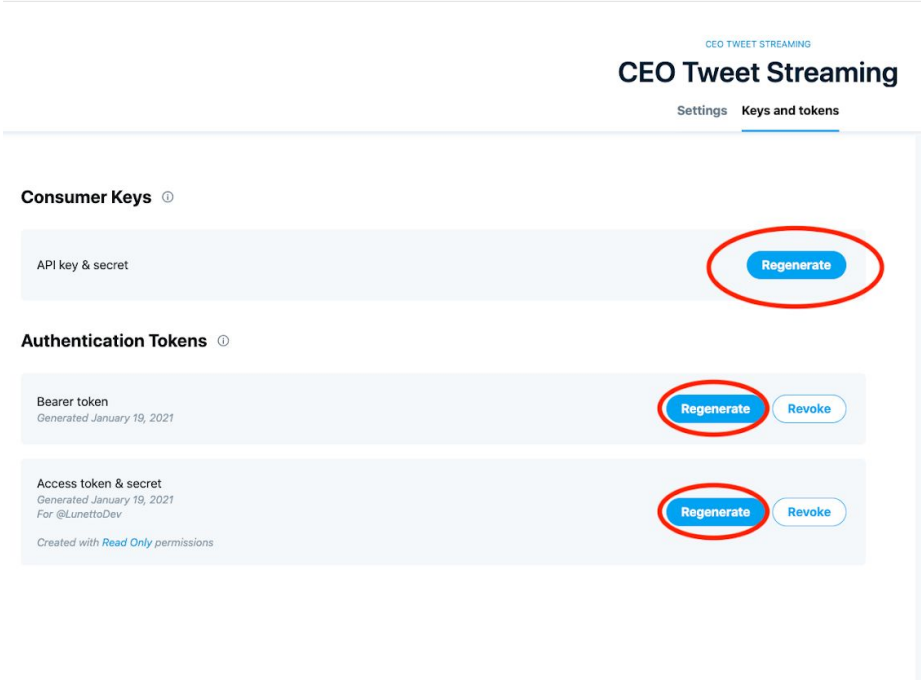
2. Once you're approved, access your Twitter Developer Dashboard at https://developer.twitter.com/en/portal/dashboard



3. Click 'Create Project' and name your project and app (for example: "CEO Tweet Streaming")

4. Click the key icon on your Project App to open your apps Keys and tokens page





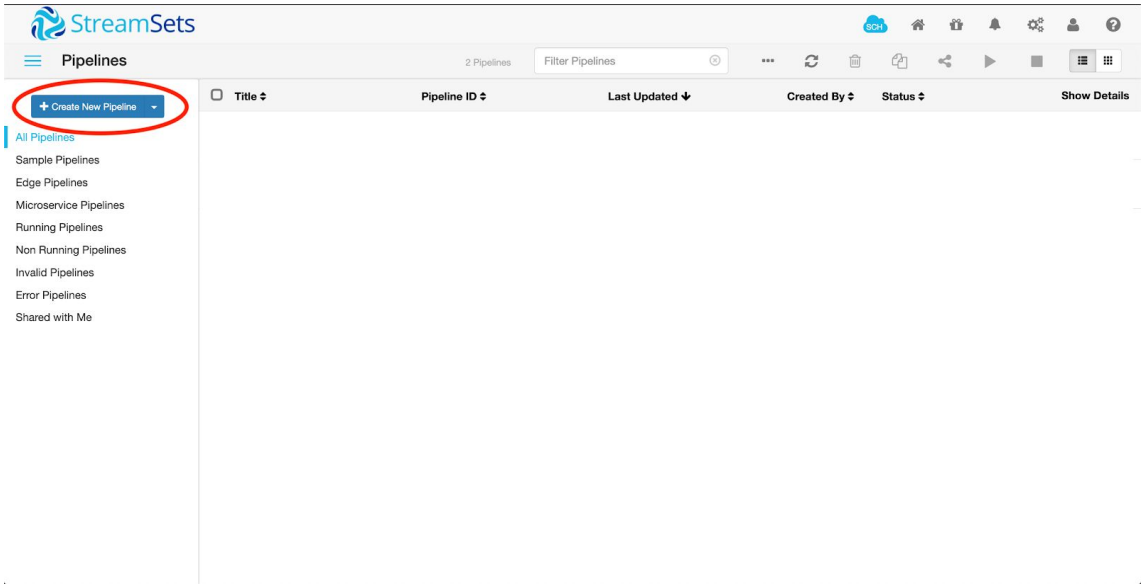
5. Click the ‘Regenerate’ button for your Consumer Keys and immediately copy paste them onto a notepad and save. Repeat this step for your ‘Bearer token’ and ‘Access token & secret’.

   You will need these later in the tutorial.

Step 2: Create New Pipeline and add Origin and Destination

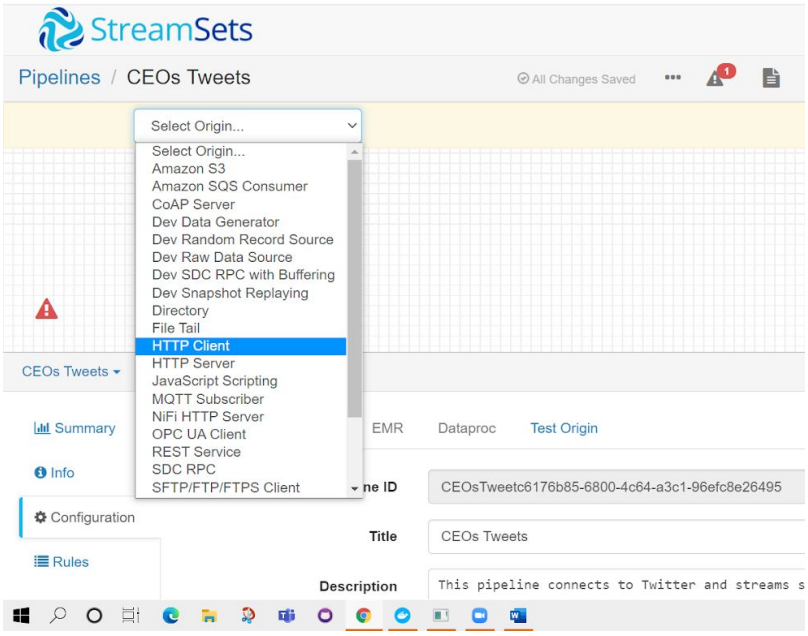
In this step you create your pipeline and start building it by by adding the origin and destination.

1. Click the Create New Pipeline button

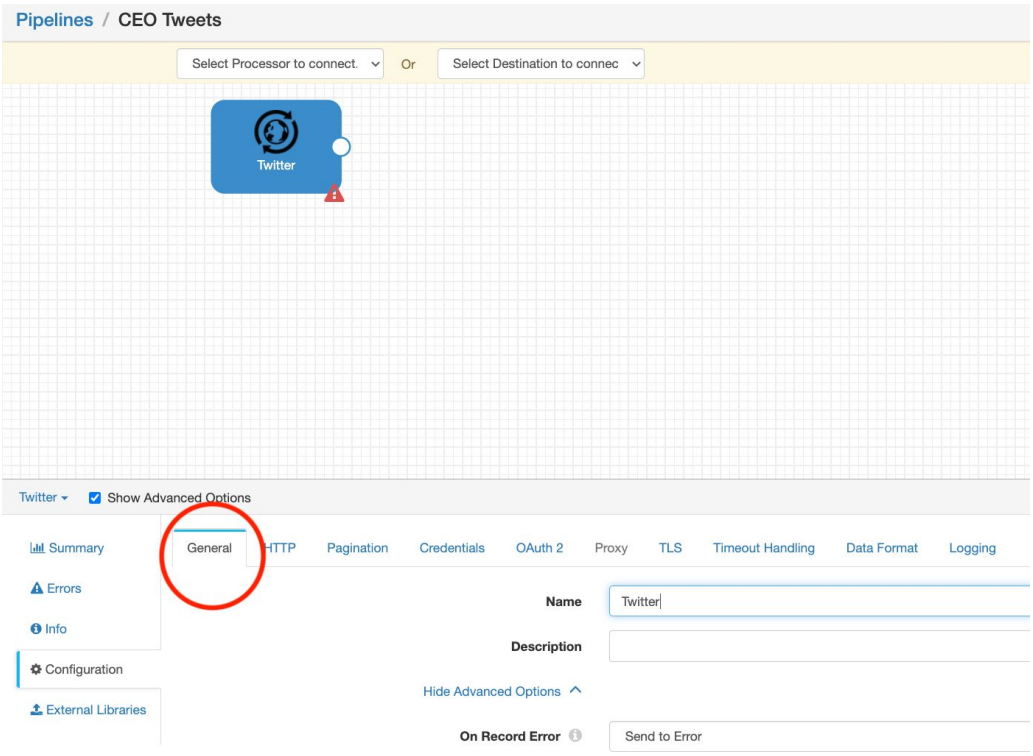


2. Give the pipeline a name and description (for example: CEO Tweets)

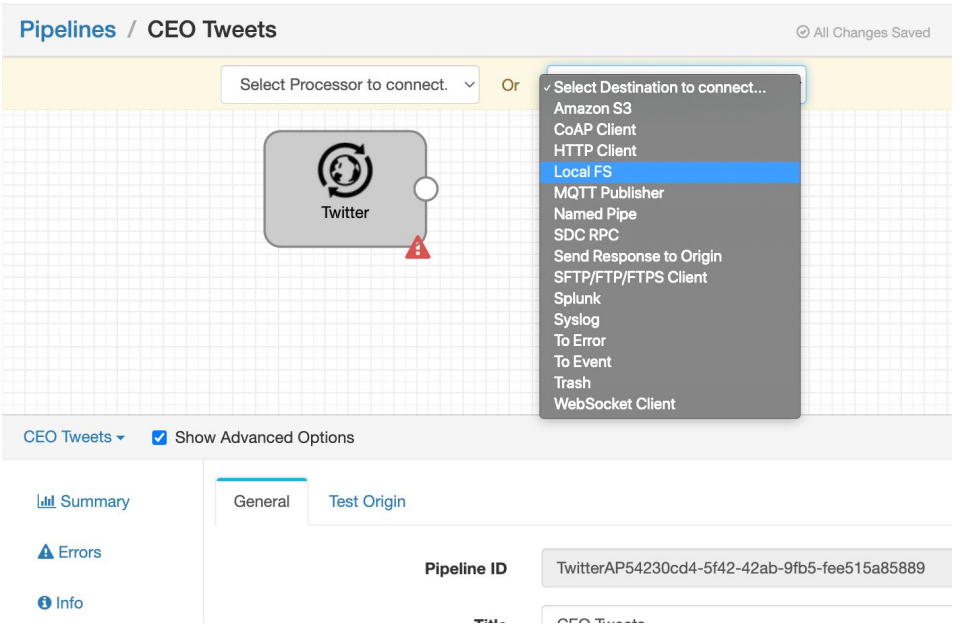
3. Click the Select origin drop down menu and select 'HTTP Client'



4. Click the ‘General’ tab below and name this origin: ‘Twitter’



5. Click the Select Destination to connect drop down menu and select ‘Local FS’



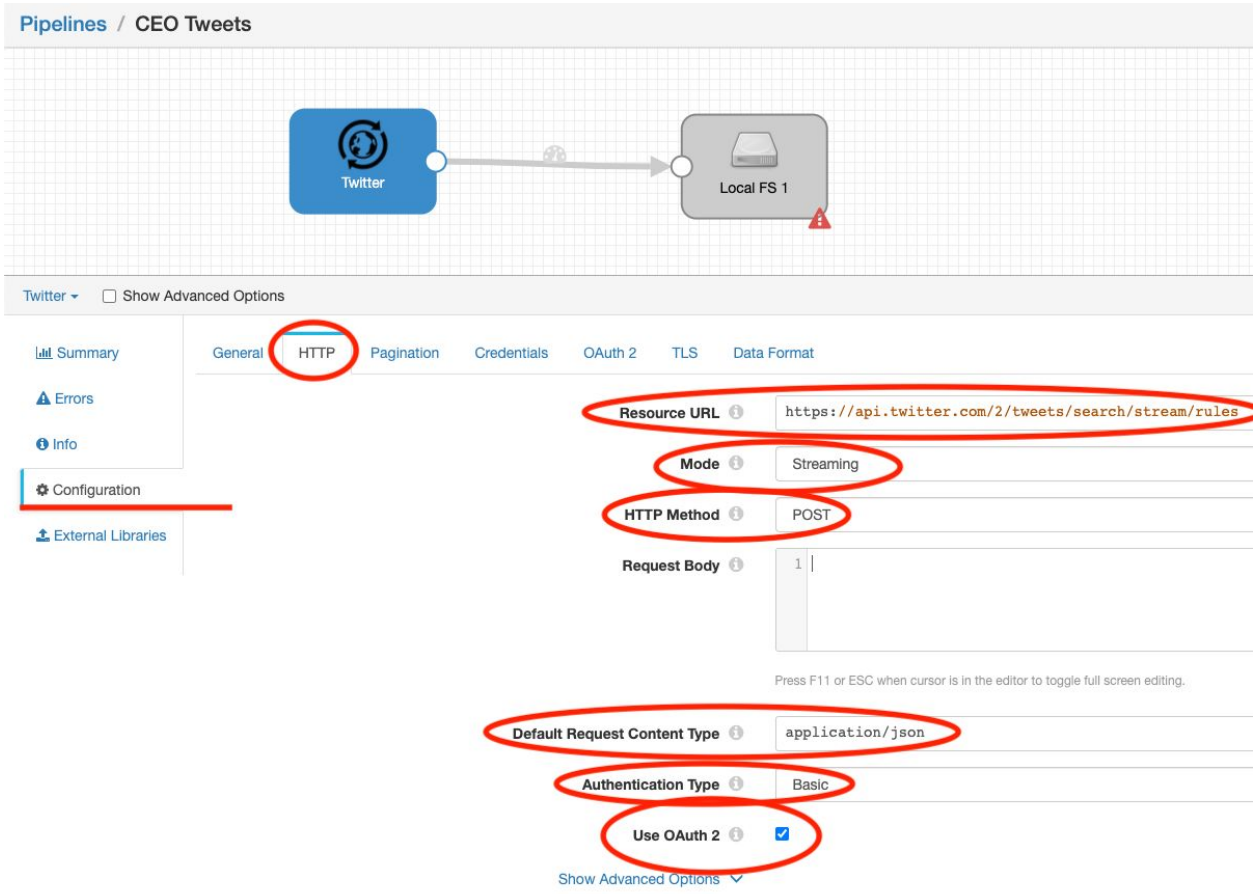
Step 3: Configure Origin & Set Rules

This step configures the origin's ('Twitter') HTTP, Pagination, Credentials, OAuth 2, TLS, and Data Format using your Twitter API keys. Also, you'll set up your Rules to filter your Tweets stream.

1. Click on your origin 'Twitter', then click 'Configuration' on the bottom left menu to display your configuration areas.

HTTP Tab:

2. Input the **Resource URL**: https://api.twitter.com/2/tweets/search/stream/rules
3. Select the **Mode**: 'Streaming'
4. Select the **HTTP Method**: 'POST'
5. Input the **Default Request Content Type**: 'application/json'
6. Select the **Authentication Type**: 'Basic'
7. Enable **'Use OAuth 2'** (this box should stay checked)



8. Add your rules into the **Request Body**:

```
{"add":
      [
```

```
            {"value": "from:YOUR OWN TWITTER HANDLE lang:en", "tag":
"test users in English"},

            {"value": "CEO has:links lang:en", "tag": "tweets about CEO
in english" },

            {"value": "CEO said lang:en", "tag": "tweets about CEO in
english" },

            {"value": "#CEO lang:en", "tag": "tweets about CEO in
english" },


            {"value": "from:benioff from:richardbranson from:bill_gross
from:chrisbrogan  from:briansolis from:jon_ferrara from:mcuban
from:garyvee from:jack from:levie from:stevecase from:rwang0
from:romanstanek from:petercashmore from:steveforbesceo from:timoreilly
from:elonmusk from:bhalligan from:charleneli from:donaldtrump
from:michaeldell from:CEO_INGDIRECT from:andrewgrill from:wendyslea
from:billgates from:shervin from:mickyarison lang:en", "tag": "CEOs
tweet in english" },

            {"value": "from:markfidelman from:marissamayer
from:rupertmurdoch from:aneelb from:mtbert from:kevinrose
from:bryankramer from:dens from:dickc from:davidmorin
from:manpowergroupjj from:michellerhee from:invoker from:davidkarp
from:drewhouston from:jeffweiner from:jeremys from:eldsjal
from:NPRgaryknell from:amfamjack from:peretti from:andrewmason
from:westernunionCEO from:tim_cook lang:en", "tag": "CEOs tweet in
english"}

            ]
      }
```

Note: Each 'value' is a custom rule you create to filter your stream of Tweets. Refer to Twitter API Documentation for rule options and syntax. You will initiate the rules after you've completed Step 4: Configure Destination.

Pagination Tab:

9. Select 'None'

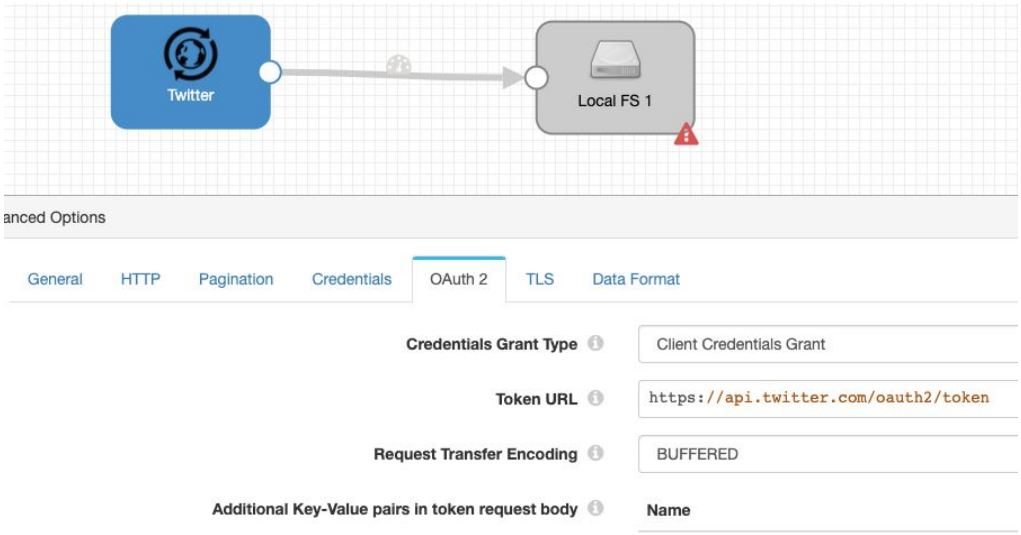
Credentials Tab:

10. Copy and Paste your Twitter **API KEY** into the Username field

11. Copy and Paste your Twitter **API SECRET KEY** into the password field

OAuth 2 Tab:

12. Select **Credentials Grant Type:** ‘Client Credentials Grant’
13. Input **Token URL:** https://api.twitter.com/oauth2/token
14. Select **Request Transfer Encoding: ‘**Buffered’
15. Leave ‘Additional Key-Value pairs in token request body’: **BLANK**



TLS Tab:

16. Leave Box **UNCHECKED**

Data Format Tab:

17. Select **Data Format:** ‘JSON’
18. Select **JSON Content:** ‘Multiple JSON Objects’

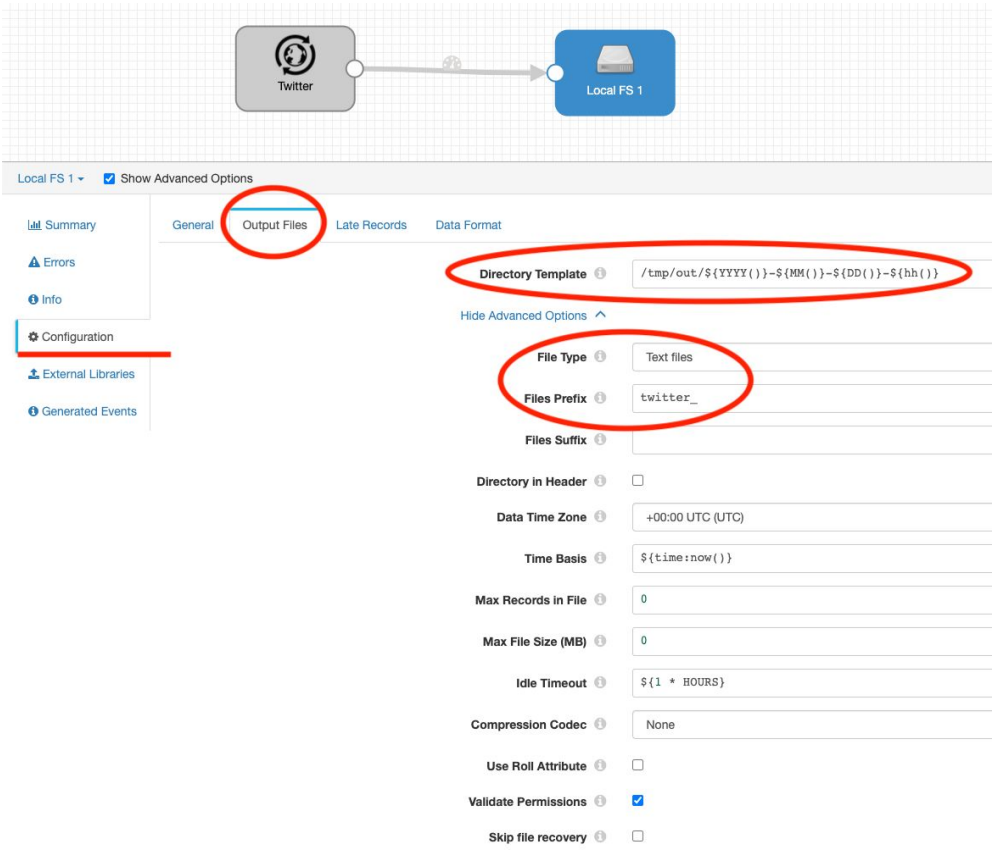
Step 4: Configure Destination and POST Rules

This step configures the destination ('Local FS') Output Files, Late Records, and Data Format. Also, you will initiate/POST your rules to filter your Tweets stream.

1. Click on your destination 'Local FS' then click 'Configuration' on the bottom left menu to display your configuration areas.

Output Files:

2. Input into **Directory Template:** /tmp/out/${YYYY()}-${MM()}-${DD()}-${hh()}
3. Click 'Show Advanced Options'
4. Select **File Type: '**Text Files'
5. Input **Files Prefix:** 'twitter_'
6. Leave all other options as **Default**



Late Records:

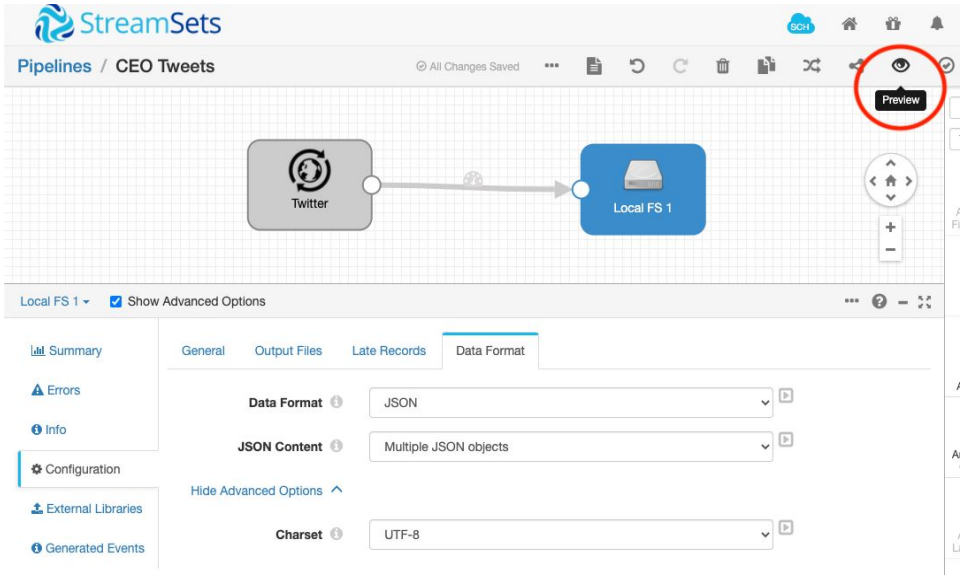
7. Leave all options as **Default**

Data Format:

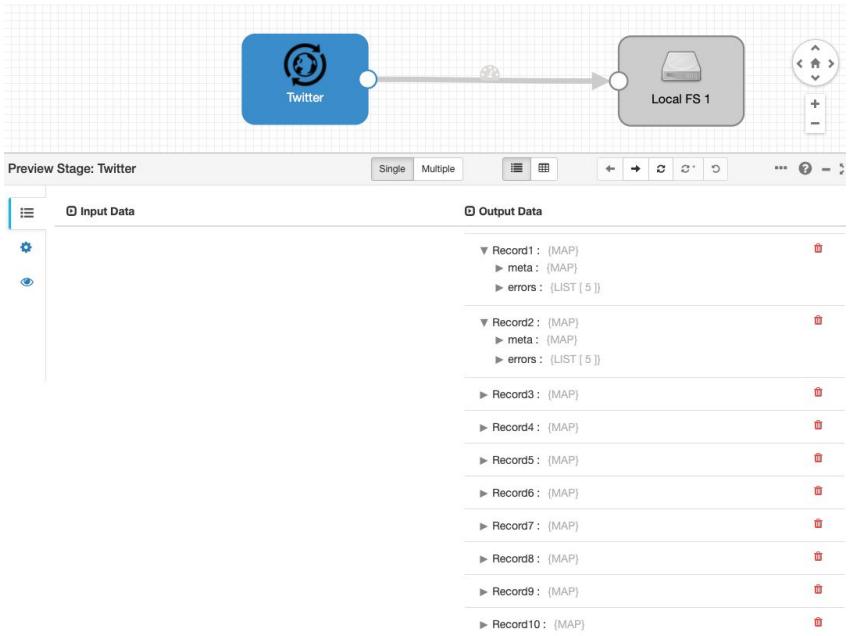
8. Select **Date Format:** 'JSON'
9. Select **JSON Content:** 'Multiple JSON Objects'

Now you're ready to initiate/POST your rules.

10. Click the 'Preview' icon at the top right (it's an icon of a human eye)



You should see a preview of your Output Data as a list of records:



Note: Open the 'meta' menu, then open 'summary' menu and you will see which rule that record is associated with. If you see errors listed, review the error message and troubleshoot.

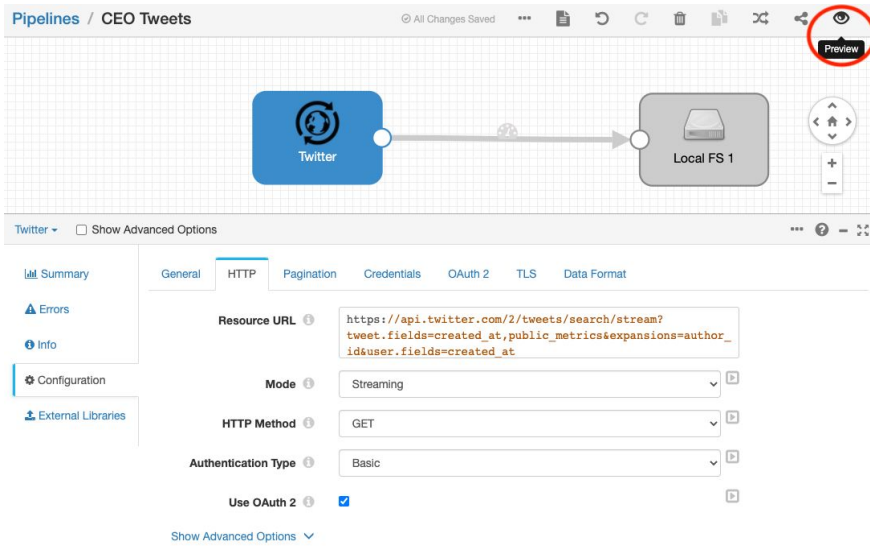
Step 5: GET Responses and Stream Filtered Tweets

This is the final step to completing your pipeline build.

1. Click on your origin ‘Twitter’ then click ‘Configuration’ on the bottom left menu to display your configuration areas.
2. Click the HTTP Tab
3. Insert the **Resource URL:** https://api.twitter.com/2/tweets/search/stream?tweet.fields=created_at,public_metrics&expansions=author_id&user.fields=created_at
4. Select **Mode:** ‘Streaming’
5. Select **HTTP Method:** ‘GET’
6. Select the Authentication Type: ‘Basic’
7. Enable **‘Use OAuth 2’** (this box should stay checked)

Now you’re ready to GET your responses of streaming filtered Tweets.

8. Click the ‘Preview’ icon at the top right (it’s an icon of a human eye)



Review the preview of your data and troubleshoot as needed. Once you have zero errors, you can click the green RUN button at the top right (next to the preview button) and allow your pipe to stream real time filtered tweets into your Local FS.

Your Summary dashboard should look like this once successfully running.

This is the end of the tutorial.

References

1. URL of Data Source: http://www.twitter.com
2. URL of your Github: https://github.com/mohsenualam/Twitter-in-StreamSets
3. URLs of References:

- https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/api-reference/get-tweets-search-stream

- https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/api-reference/post-tweets-search-stream-rules
- https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Processors/HTTPClient.html#concept_s4p_15f_5y
- https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/integrate/build-a-rule