Using HDP and HDF to Harness the Power of Social Media

This lab will walk through the collection and analysis of Twitter data leveraging the Hortonworks Data Platform 2.6 (HDP) and Hortonworks DataFlow 3.0 (HDF), combining both data in motion and data at rest. While the initial analytics are simple, this pattern enables the building blocks for more advanced use cases.

During this lab, participants will be able to perform Twitter live feed data ingestion based on some keywords, filtering, transformation, attribute extraction, and delivery to HDFS for data storage, as well as Apache Kafka using Apache NiFi.

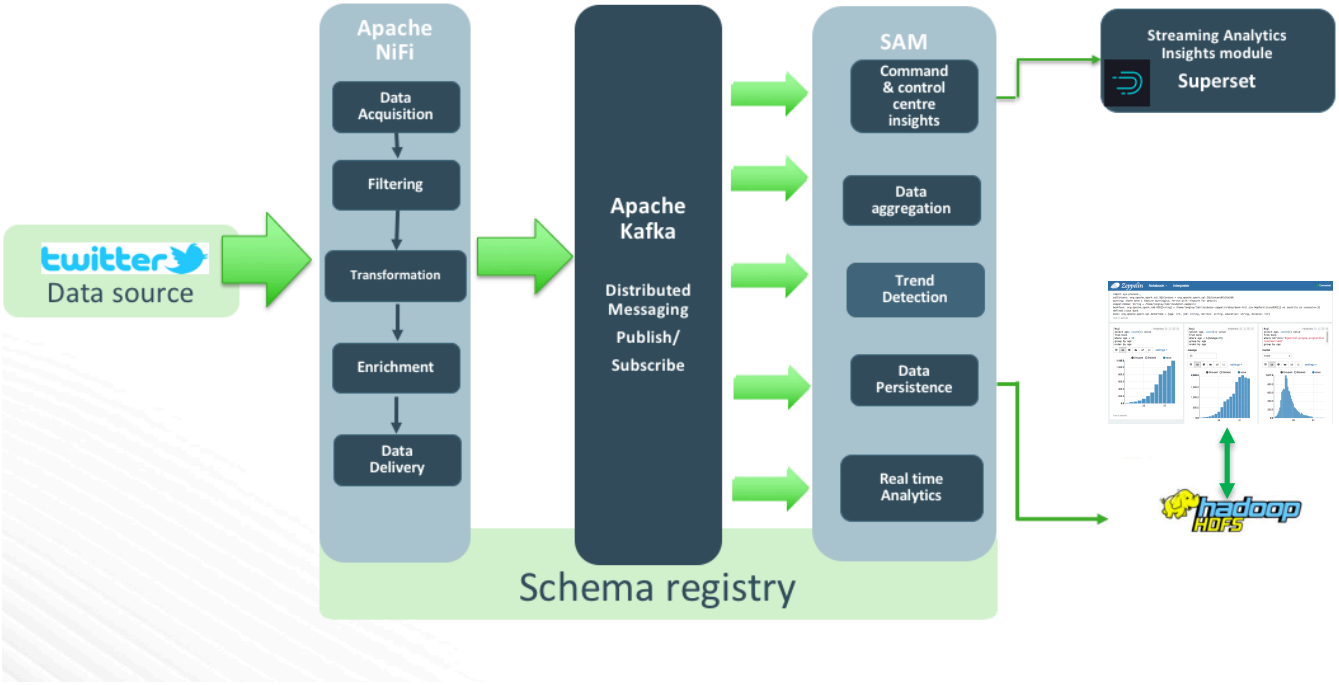
Participants will then be able to perform analytics on data at rest using Hive, SparkSQL as well as Apache Zeppelin for visualization, and data exploration.

Once the data is ingested into Kafka, participants will be able to develop quickly a Streaming Analytics Manager application to further analyze data and derive real-time insights from the data. Data will also be visualized in near real-time using Apache Druid and Superset.

Overview 3
Pre-requisites 4
Create your Twitter API keys 5
Create A Twitter Application 5
Verify your Hortonworks Sandbox web access 8
Using Apache NiFi for ingesting Twitter Data 9
Download the Twitter dataflow template 12
Import the Template 12
Instantiate the Template 14
Configure the Twitter Grab Garden Hose Processor 15
Configure PUTHDFS 16
Review the remaining Processors 18
Start the Flow 19
Check the data on HDFS 20
Data Analysis with Hive 22
Create the Hive Table 22
Query Data using SQL on Hive 23
Data Visualization with Zeppelin 24
Real-Time Data Visualization with Apache Druid and Superset 28
Locate the Druid Twitter datasource 28
Create a visualization and a dashboard 30

Overview

Below is a conceptual diagram of the implementation of Twitter Analytics performed in this lab using Hortonworks Data Flow and Hortonworks Data Platform.



Pre-requisites

Where possible, the following pre-requisites should be completed prior to start of the lab. Please follow the instructions in the Lab prerequisites document. Prerequisites steps included in the documents are:

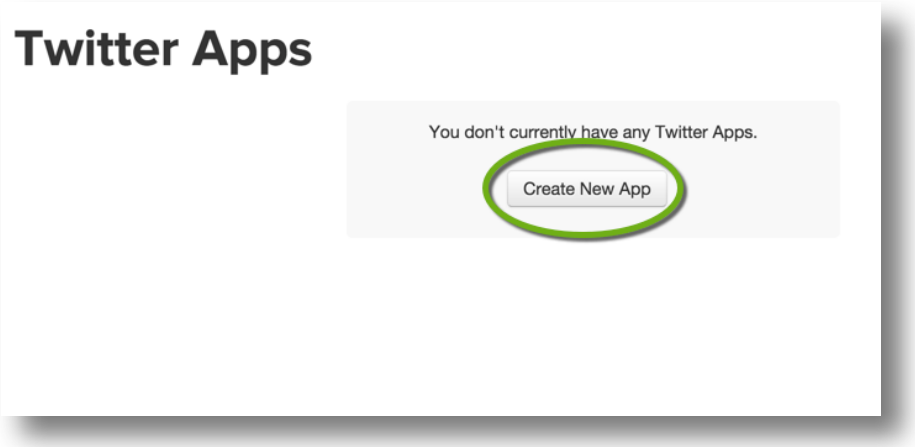
- **Create your Twitter API keys.**
- **Verify your Hortonworks Sandbox web access.**

Create your Twitter API keys

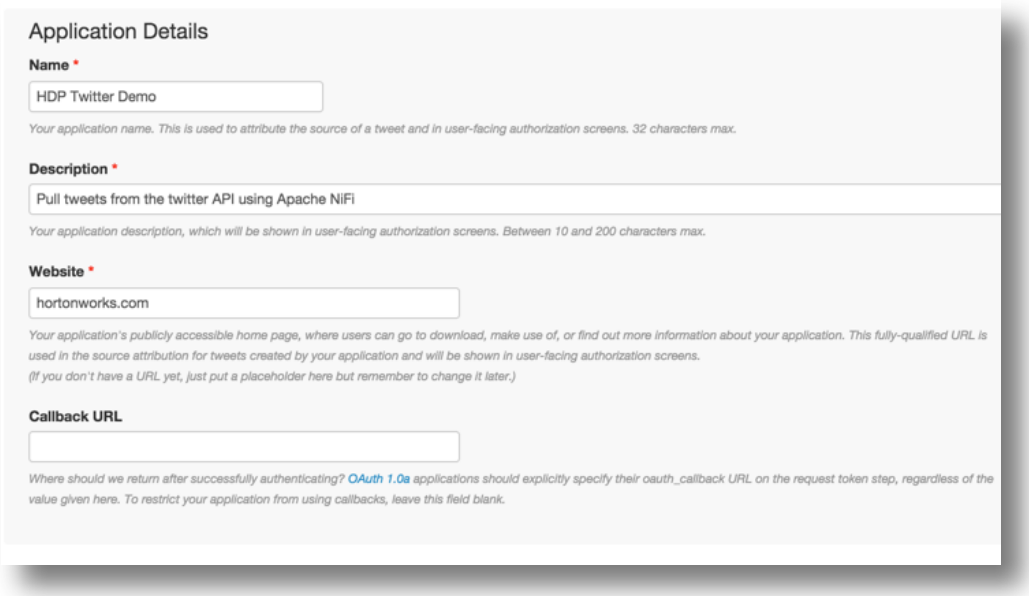
You'll need to register your own Twitter application. It's quite simple and only takes a few short steps

Create A Twitter Application

First head over to the Twitter Apps Website ([https://apps.twitter.com/)](https://apps.twitter.com/) and Sign In using your Twitter account (or make one if you don't have one yet!). Then click Create a New App.



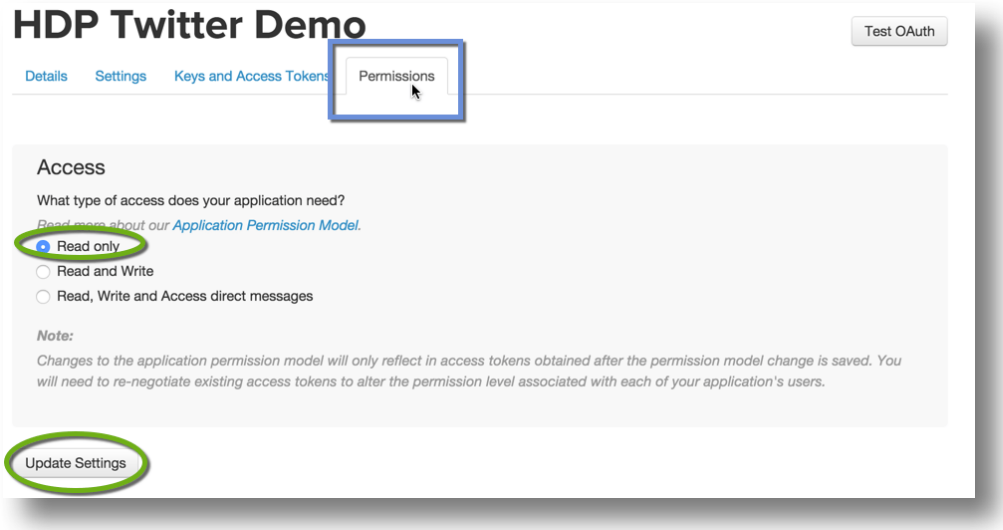
After you've clicked that you'll need to fill in some details about your application. Feel free to put whatever you want.



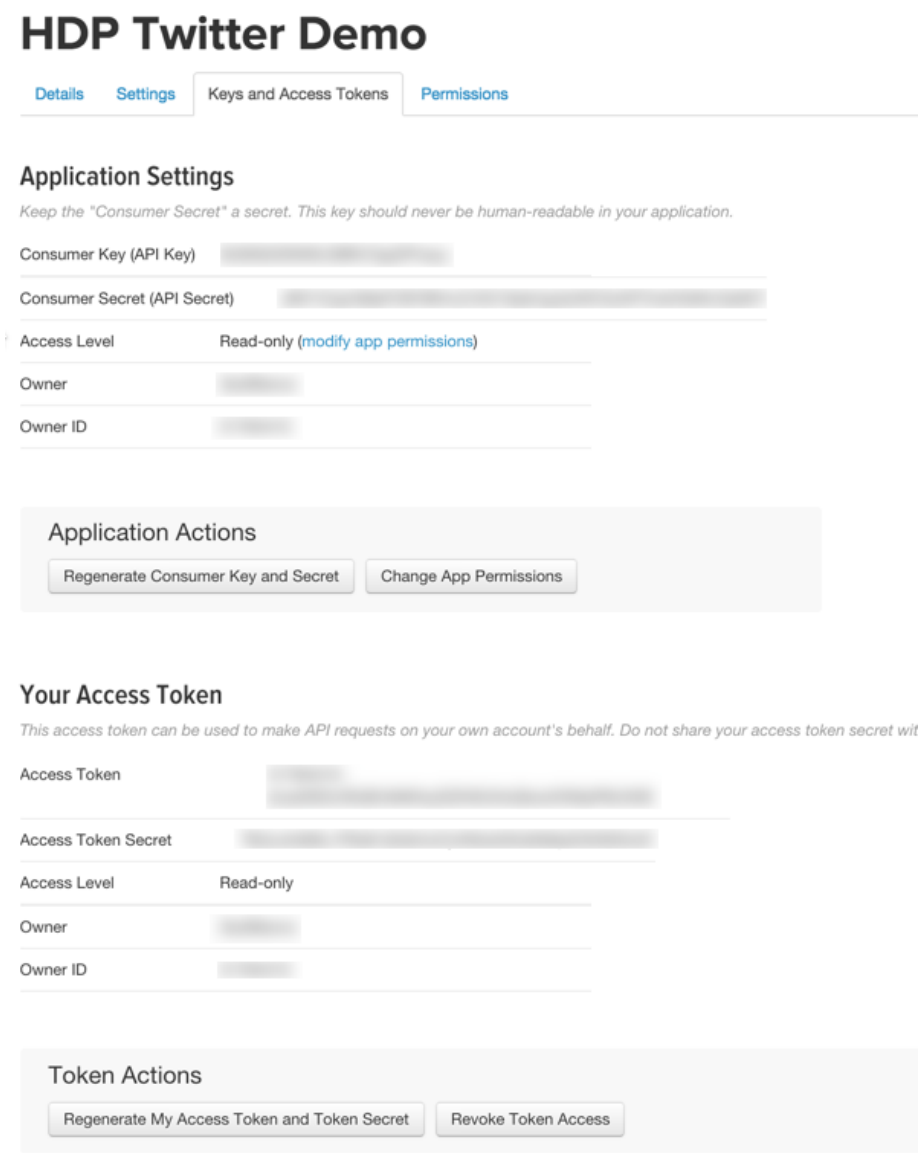
Then click **Create Your Twitter Application** at the bottom of the screen after reading the developer agreement.

Note that you might need to add your mobile phone to your Twitter account before creating your application

Once you've done that you should be greeted by a dashboard for your Twitter application. Head over to the permissions tab and select the **Read Only** Option and **Update** your application.



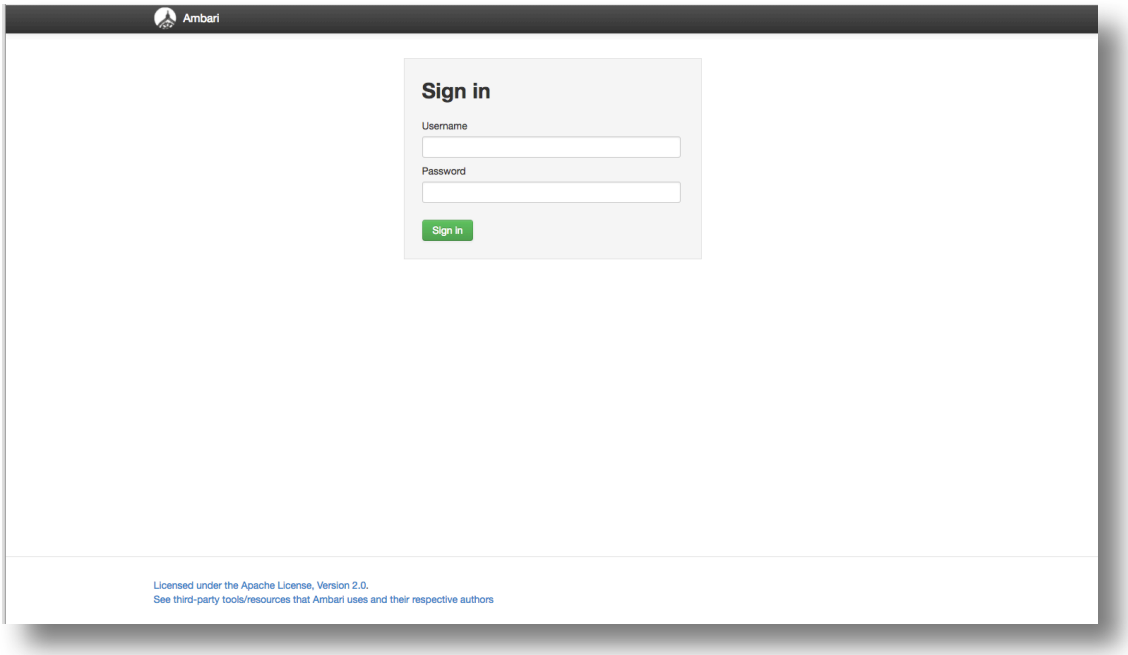
Finally you need to generate your OAuth key. You can do this by clicking **Test OAuth** on the top of the permissions page, or by heading to **Keys and Access Tokens** and then finding the option that allows you to generate your OAuth tokens. Your **keys** and **access tokens** should look similar to the following:



Please make note of your **Consumer Key**, **Consumer Secret**, **Access Token**, and Access Token Secret. You will need these to create the data flow in NiFi.

Verify your Hortonworks Sandbox web access

Once these prerequisites steps are done, please make sure your VM is started on Azure by accessing it with an Internet Browser at http://< sandbox-ip> :8080 .
You should expect the following web page :



If the VM is down, please contact the lab instructor to be provided with a new Hortonworks Sandbox.

Using Apache NiFi for ingesting Twitter Data

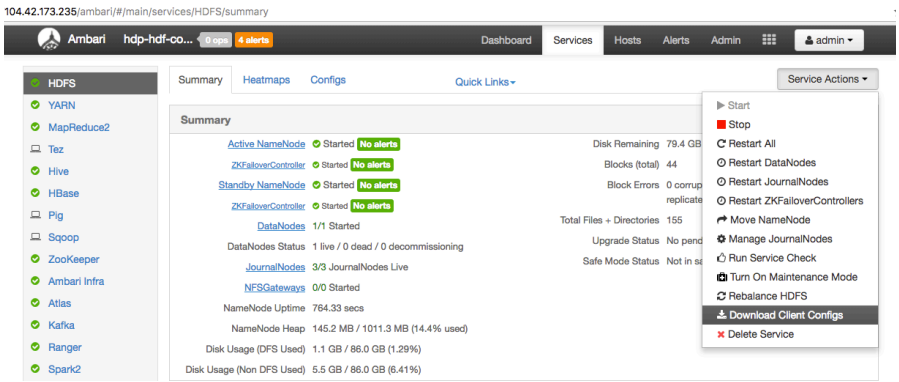
HDF (NiFi) has a powerful templating system to allow for easy sharing and consumption of data flows. The following section focuses on importing a prepared Twitter dataflow template, customize parameters according to your environment (Twitter credentials, Twitter keywords, IP addresses) and starts the collection of Tweets into HDFS.

Prerequisites for Hive –HDFS integration

1. Download the following configuration files Copy them on the Nifi instance under /home/nifi/ and position the directory ownerships for nifi/hadoop (*chown – R nifi:hadoop /home/nifi/*) :

    - core-site.xml ,
    - hdfs-site.xml
    - hive-site.xml

    This files can be obtained in the HDP ambari admin UI as follow :

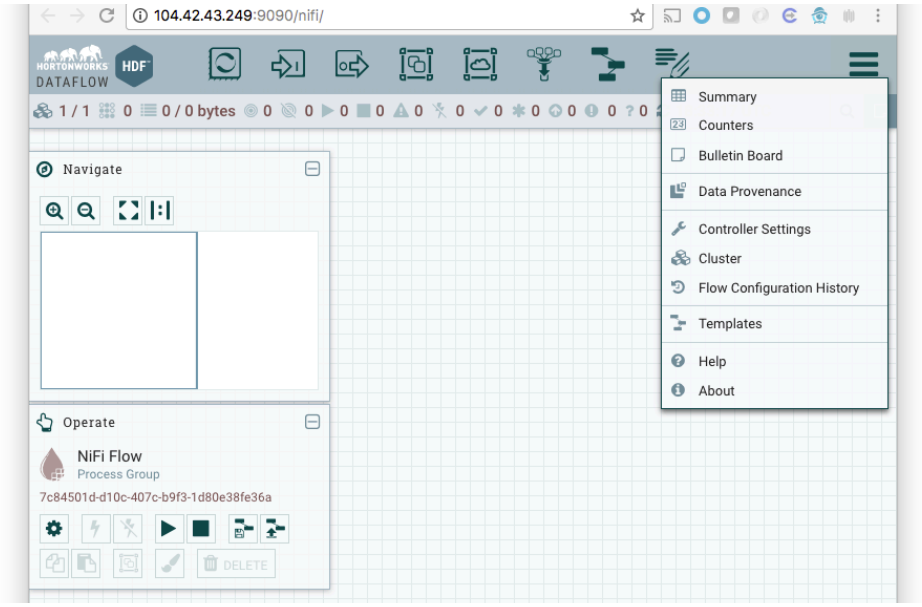


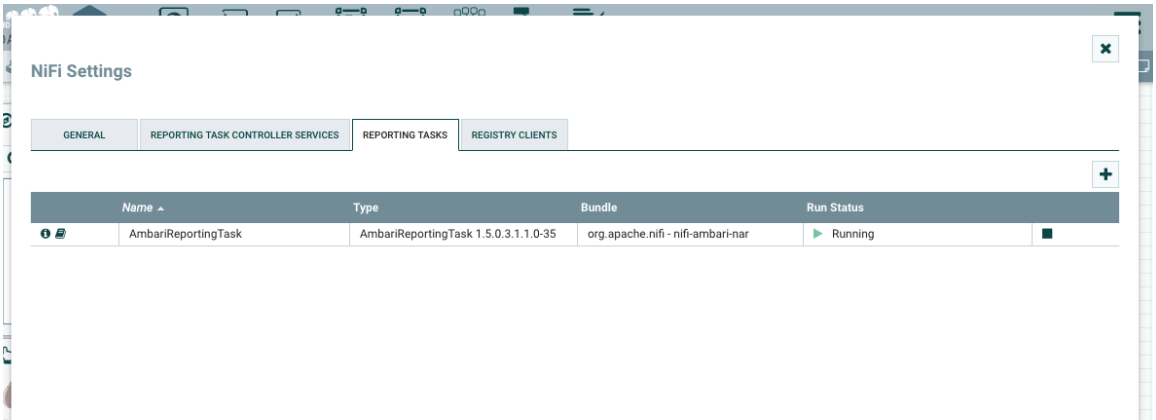
Make sure Nifi instance is can reach the required (hive, hdfs, altas, kafka broker) HDP services through the firewall or network security groups.

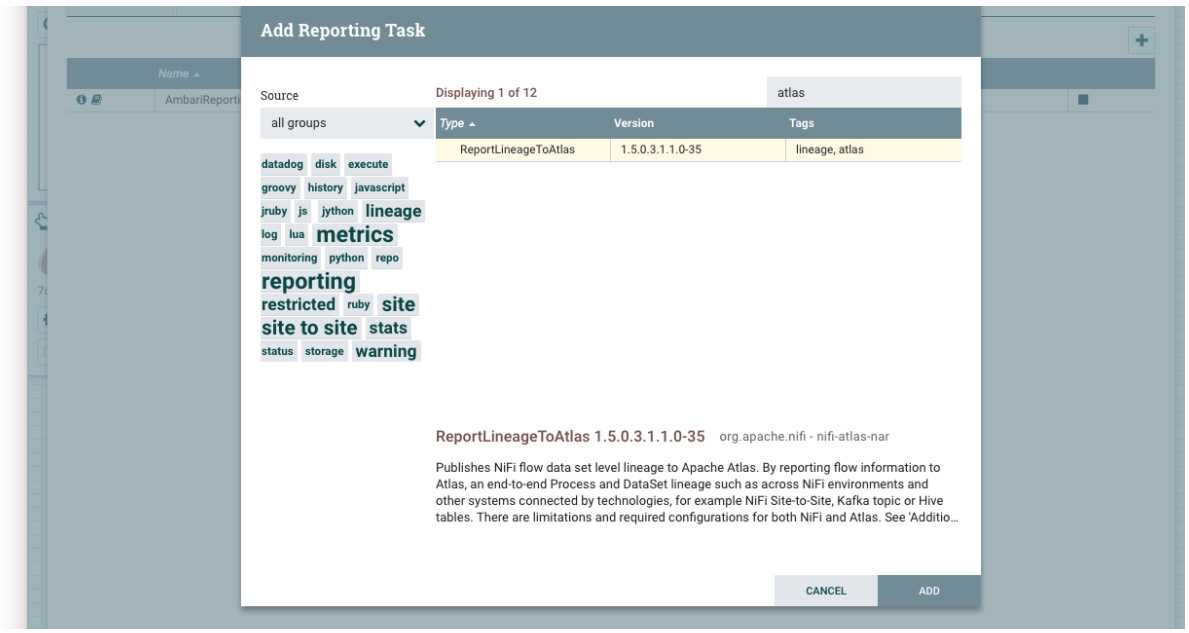
Prerequisites for ATLAS integration

- URL for Atlas REST API: BASIC with username / password
- Kafka broker IP address and port (Ex: some-IP:6667)

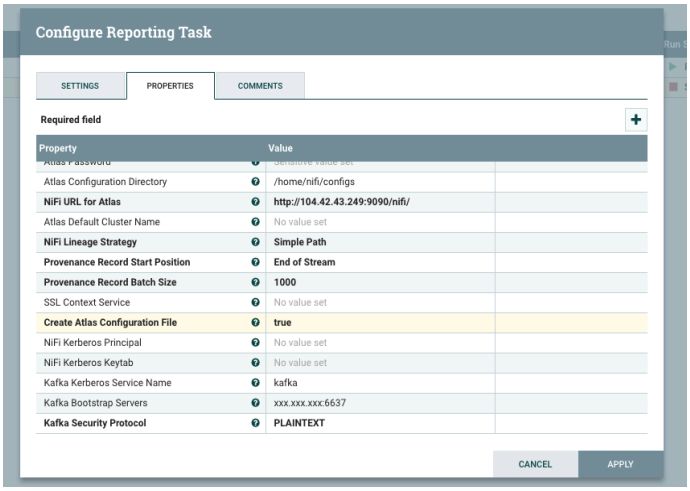
1. Add ReportLineageToAtlas as a controller service :







2. Configure ReportLineageToAtlas



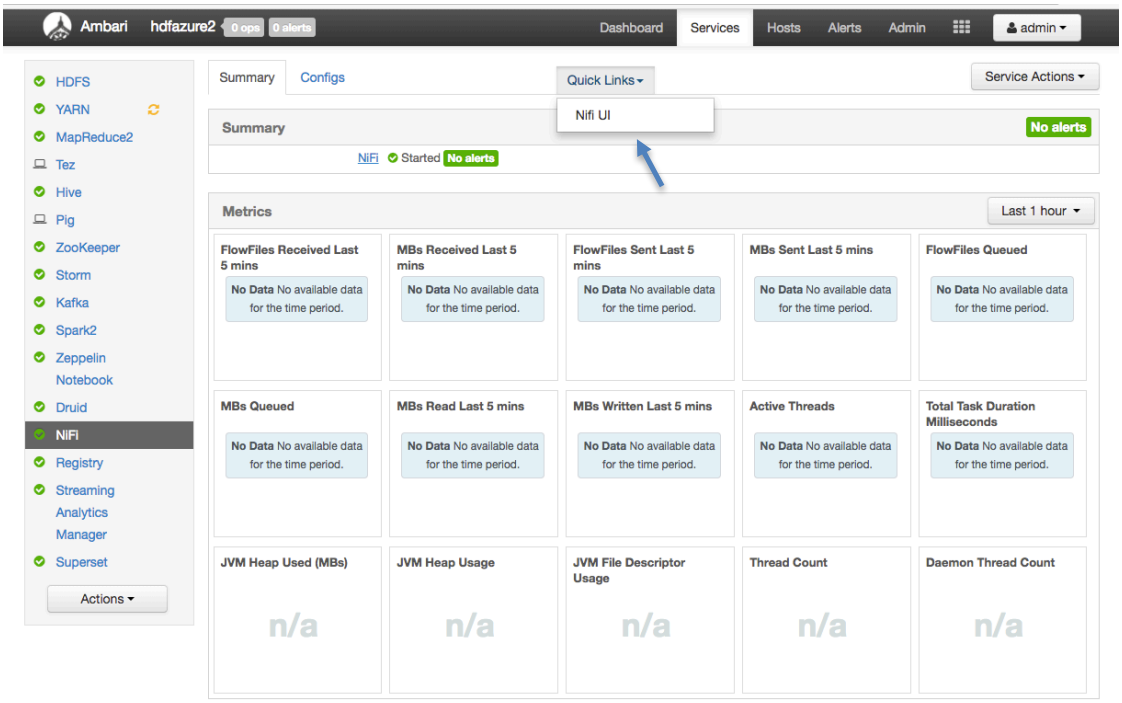
Download the Twitter dataflow template

The Twitter dataflow template should be downloaded to the workstation accessing NiFi at the following link.

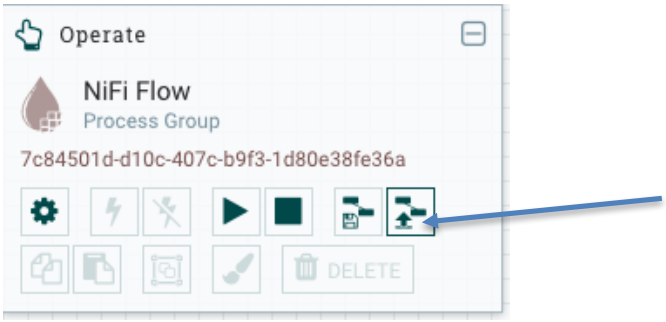
https://github.com/oascofare/poc-cbreak-hdp-hdf/blob/master/HDF/nifi-template/Twitter-to-hdsf.xml

Import the Template

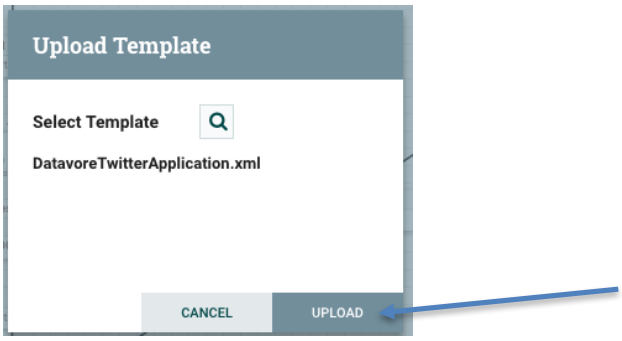
Click on Nifi Service Quick link or navigate to the NiFi web interface at *http://<public ip>: 9999/nifi/* via a web browser.



Import the template by selecting the Import Template Icon found in the NiFi Flow Operate tile on the left-hand side of the canvas.

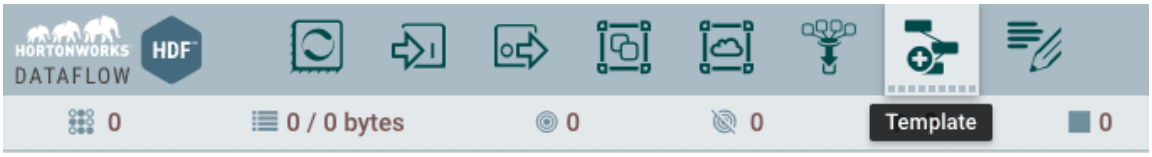


Click the Browse icon ( ) and navigate to where the DatavoreTwitterApplication.xml file was downloaded. Click on Upload.

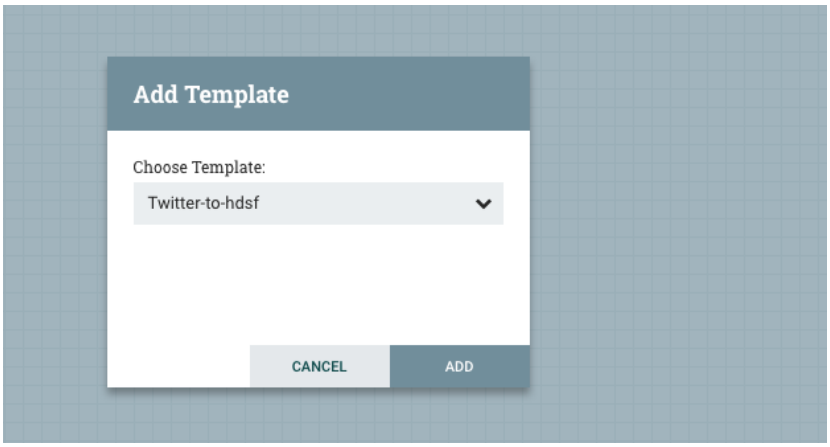


Instantiate the Template

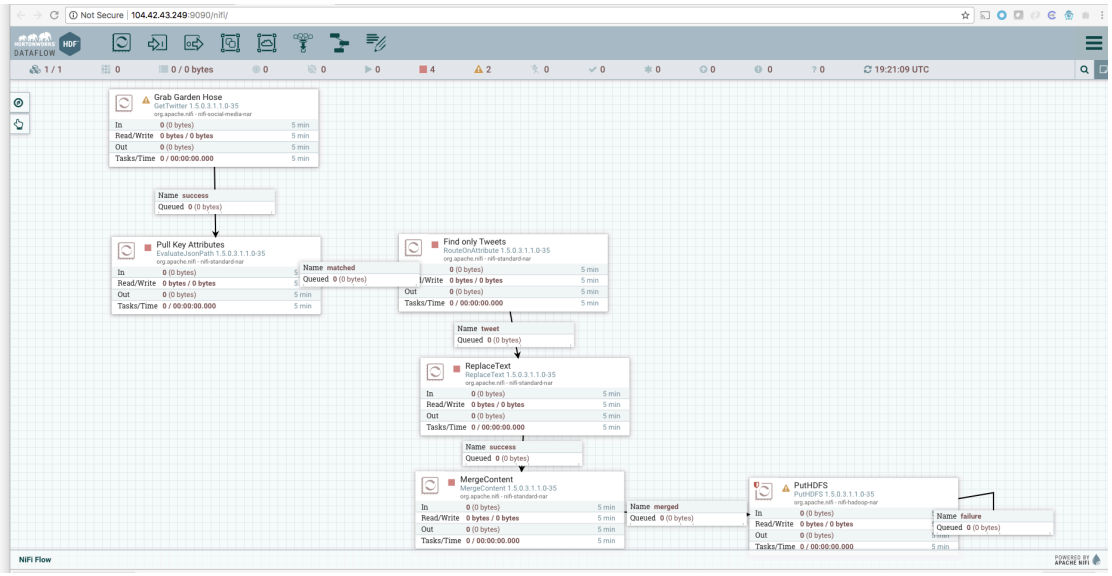
Drag and drop the template artifact onto the canvas.



Select ***Twitter-to-hdsf*** and click *Add*

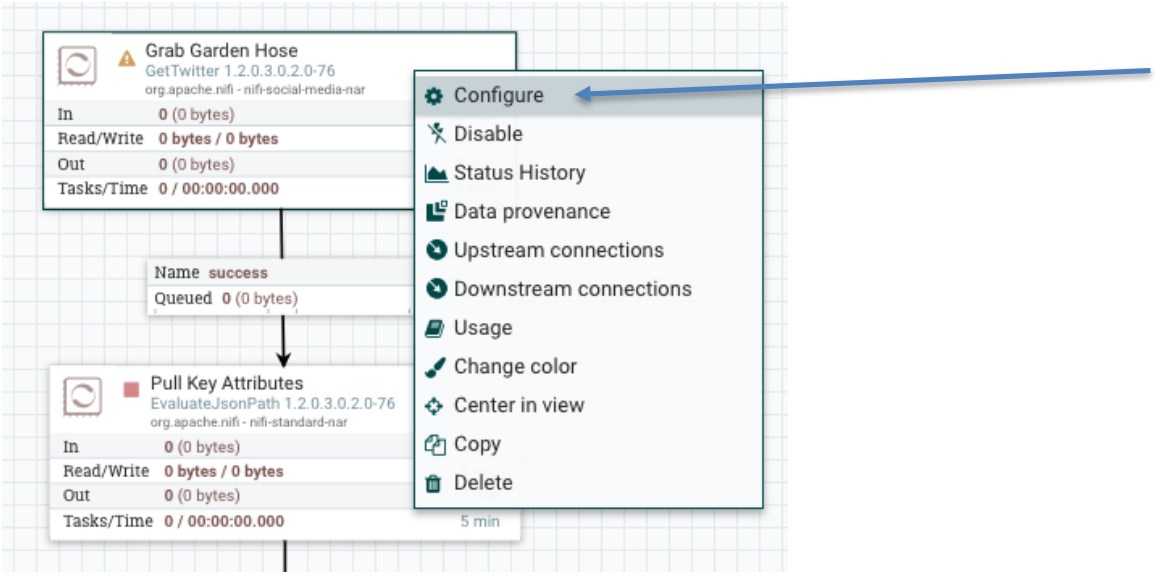


This will import the Data Flow onto your NiFi canvas.



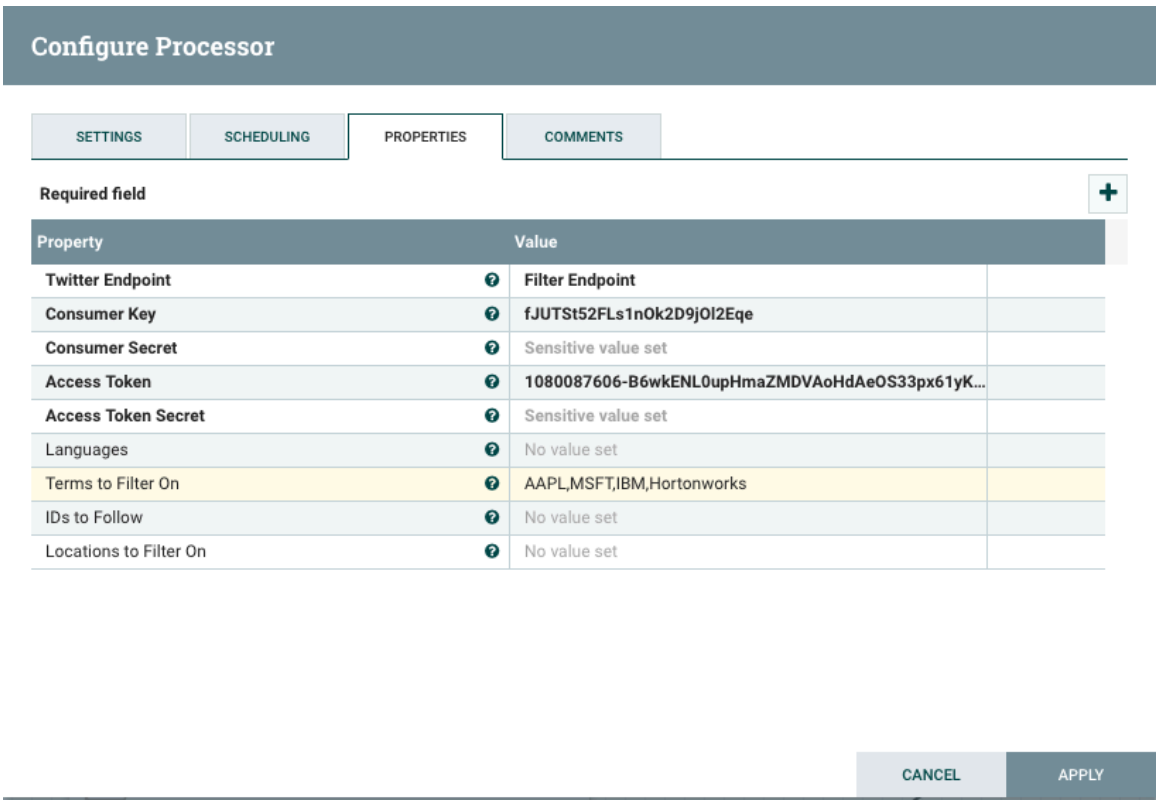
Configure the Twitter Grab Garden Hose Processor

Next, configure the Twitter Grab Garden Hose Processer with the Twitter API information. Right click the processor and select Configure. Then, select the PROPERTIES tab.

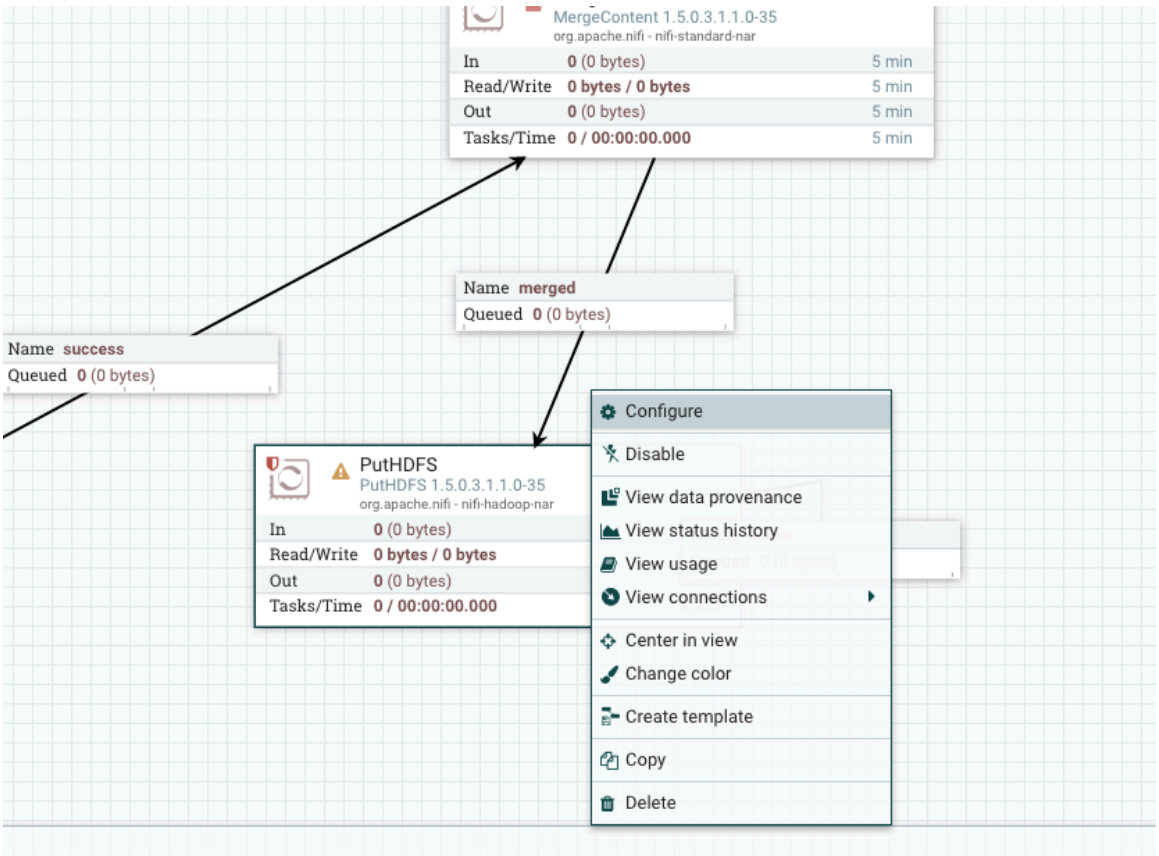


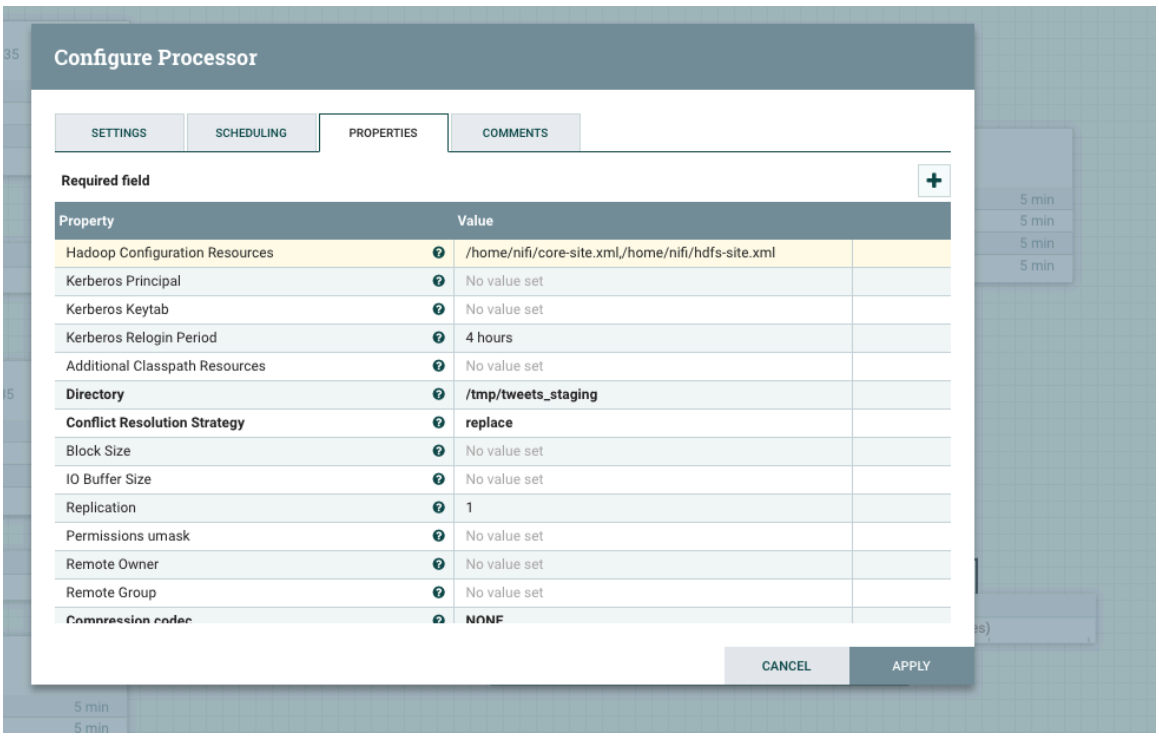
Fill in the following mandatory properties (in bold) matching to your Twitter application:

- Consumer Key
- Consumer Secret
- Access Token
- Access Token Secret



Configure PUTHDFS


MergeContent 1.5.0.3.1.1.0-35
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name merged
Queued 0 (0 bytes)
Name success
Queued 0 (0 bytes)
PutHDFS
PutHDFS 1.5.0.3.1.1.0-35
org.apache.nifi - nifi-hadoop-nar
In 0 (0 bytes)
Read/Write 0 bytes / 0 bytes
Out 0 (0 bytes)
Tasks/Time 0 / 00:00:00.000
Configure
Disable
View data provenance
View status history
View usage
View connections
Center in view
Change color
Create template
Copy
Delete



Review the remaining Processors

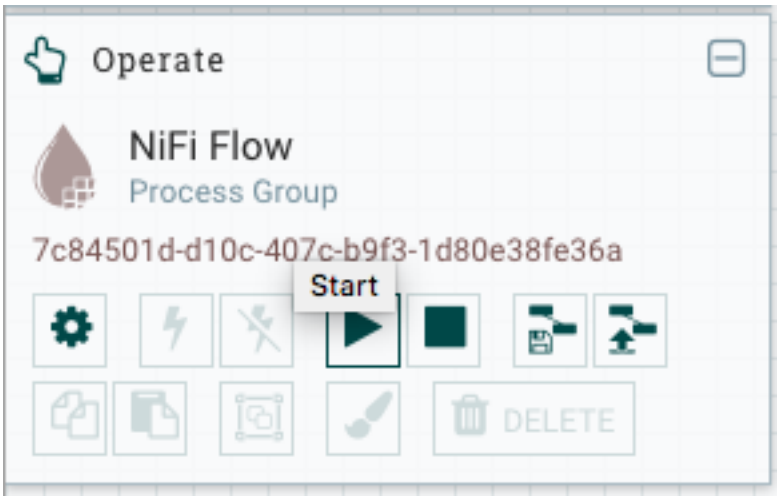
Review the other processors and modify properties as needed:

- **EvaluateJsonPath**: Pulls out attributes of tweets
- **RouteonAttribute**: Ensures only tweets with non-empty messages are processed
- **ReplaceText**: Formats each tweet as pipe (|) delimited line entry e.g. tweet_id|tweet_unixtime|tweet_stringtime}|tweet_timezone|tweet_screenname|tweet_username|tweet_followers|tweet_hashtags
- **MergeContent**: Merges tweets into a single file (either 20 tweets or 120s, whichever comes first) to avoid having a large number of small files in HDFS. These values can be configured.
- **PutHDFS**: writes tweets to HDFS under /tmp/tweets_staging
- **UpdateAttribute**: Add additional attributes to the flow files, which are the schema name pointing to the schema name in the schema registry, as well as the kafka topic name where the tweets are being written.

Start the Flow

If the processors are correctly configured, all processors should have a red stop symbol in the upper left (except green for PublishKafkaRecord, as this particular processor was started as a referencing component of the controller services). If a yellow caution symbol is present, this indicates required configuration is missing.

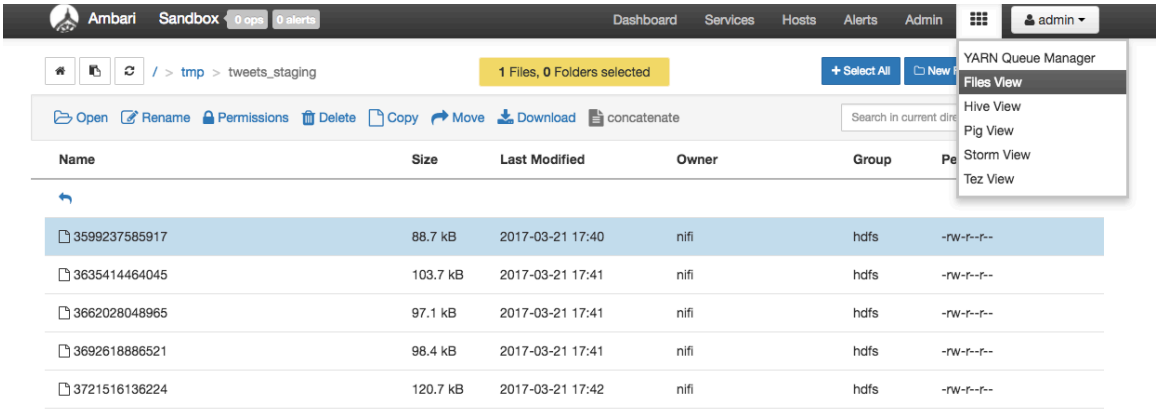
Start the flow by selecting an empty area of the canvas and clicking the start icon.



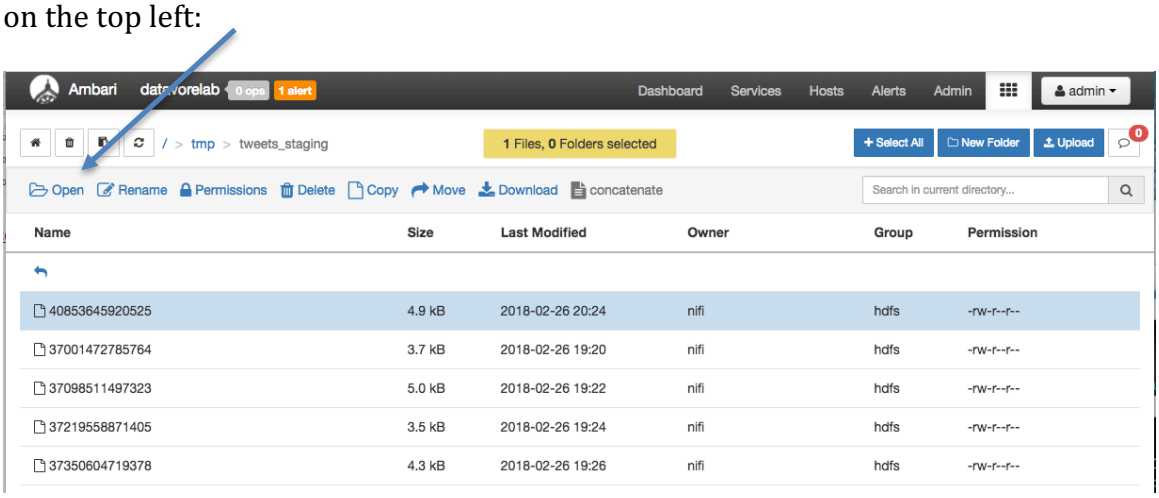
Check the data on HDFS

Now that the Twitter dataflow is running, the matching tweets will be persisted to HDFS and Kafka. Validate that the tweets are successfully written to HDFS for now.

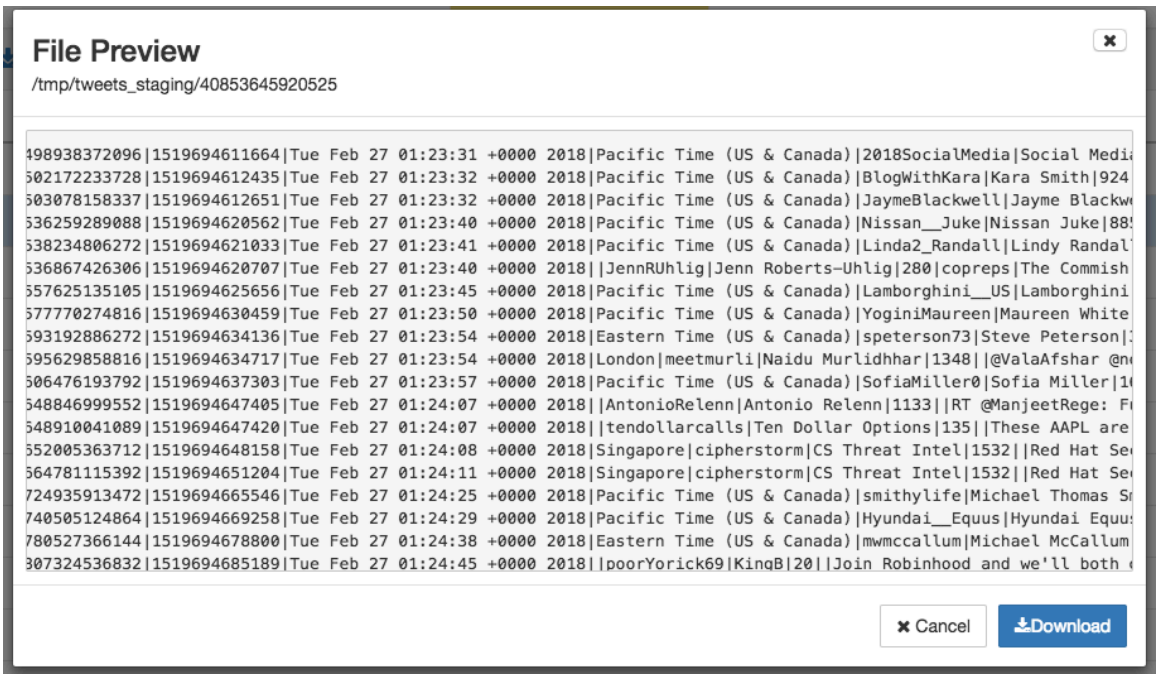
Tweets appear under /tmp/tweets_staging dir in HDFS. You can see this via Files view in Ambari:



You can open the file to visualize its content. Select one of the file, and click on Open on the top left:



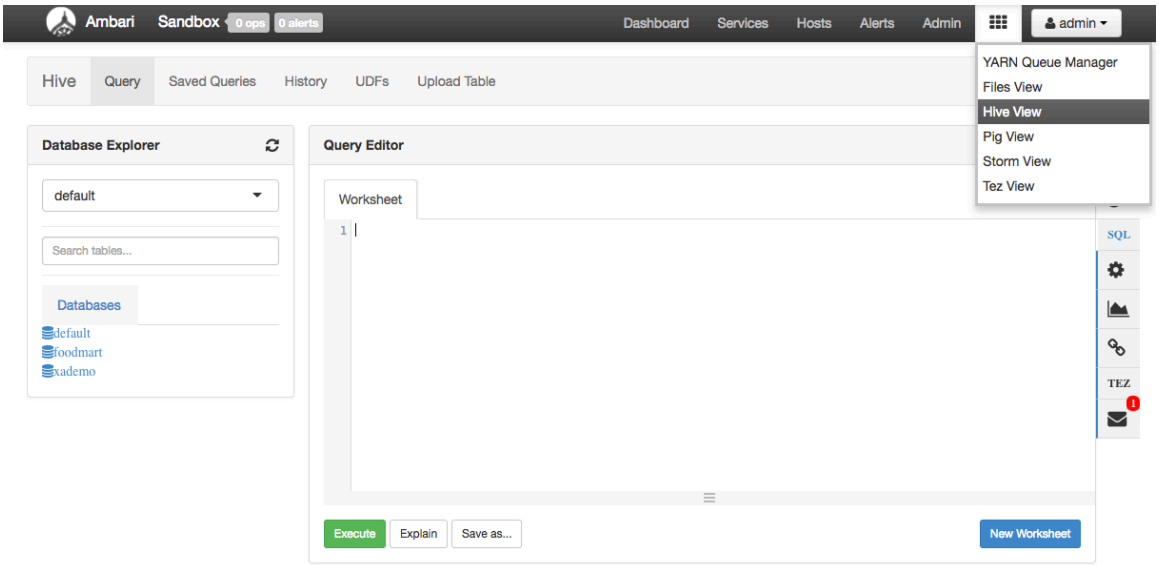
You can now preview the contents of the file. If needed, you can download the file on your local workstation for further review of its contents:


File Preview
/tmp/tweets_staging/40853645920525
498938372096|1519694611664|Tue Feb 27 01:23:31 +0000 2018|Pacific Time (US & Canada)|2018SocialMedia|Social Media
502172233728|1519694612435|Tue Feb 27 01:23:32 +0000 2018|Pacific Time (US & Canada)|BlogWithKara|Kara Smith|924
503078158337|1519694612651|Tue Feb 27 01:23:32 +0000 2018|Pacific Time (US & Canada)|JaymeBlackwell|Jayme Blackw
536259289088|1519694620562|Tue Feb 27 01:23:40 +0000 2018|Pacific Time (US & Canada)|Nissan__Juke|Nissan Juke|88
538234806272|1519694621033|Tue Feb 27 01:23:41 +0000 2018|Pacific Time (US & Canada)|Linda2_Randall|Lindy Randal
536867426306|1519694620707|Tue Feb 27 01:23:40 +0000 2018||JennRUhlig|Jenn Roberts-Uhlig|280|copreps|The Commish
557625135105|1519694625656|Tue Feb 27 01:23:45 +0000 2018|Pacific Time (US & Canada)|Lamborghini__US|Lamborghini
577770274816|1519694630459|Tue Feb 27 01:23:50 +0000 2018|Pacific Time (US & Canada)|YoginiMaureen|Maureen White
593192886272|1519694634136|Tue Feb 27 01:23:54 +0000 2018|Eastern Time (US & Canada)|speterson73|Steve Peterson|
595629858816|1519694634717|Tue Feb 27 01:23:54 +0000 2018|London|meetmurli|Naidu Murlidhhar|1348||@ValaAfshar @n
506476193792|1519694637303|Tue Feb 27 01:23:57 +0000 2018|Pacific Time (US & Canada)|SofiaMiller0|Sofia Miller|1
548846999552|1519694647405|Tue Feb 27 01:24:07 +0000 2018||AntonioRelenn|Antonio Relenn|1133||RT @ManjeetRege: F
548910041089|1519694647420|Tue Feb 27 01:24:07 +0000 2018||tendollarcalls|Ten Dollar Options|135||These AAPL are
552005363712|1519694648158|Tue Feb 27 01:24:08 +0000 2018|Singapore|cipherstorm|CS Threat Intel|1532||Red Hat Se
564781115392|1519694651204|Tue Feb 27 01:24:11 +0000 2018|Singapore|cipherstorm|CS Threat Intel|1532||Red Hat Se
724935913472|1519694665546|Tue Feb 27 01:24:25 +0000 2018|Pacific Time (US & Canada)|smithylife|Michael Thomas S
740505124864|1519694669258|Tue Feb 27 01:24:29 +0000 2018|Pacific Time (US & Canada)|Hyundai__Equus|Hyundai Equu
780527366144|1519694678800|Tue Feb 27 01:24:38 +0000 2018|Eastern Time (US & Canada)|mwmccallum|Michael McCallum
807324536832|1519694685189|Tue Feb 27 01:24:45 +0000 2018||poorYorick69|KingB|20||Join Robinhood and we'll both
Cancel
Download

Data Analysis with Hive

After persisting the tweets in HDFS, the opportunity to perform analytics against the data is nearly unlimited. Leveraging Hive, SQL can be used to analyze the Tweets. The following sections will create the Hive table and run some basic queries against the tweets.

You can access Hive via the Hive view available on Ambari:

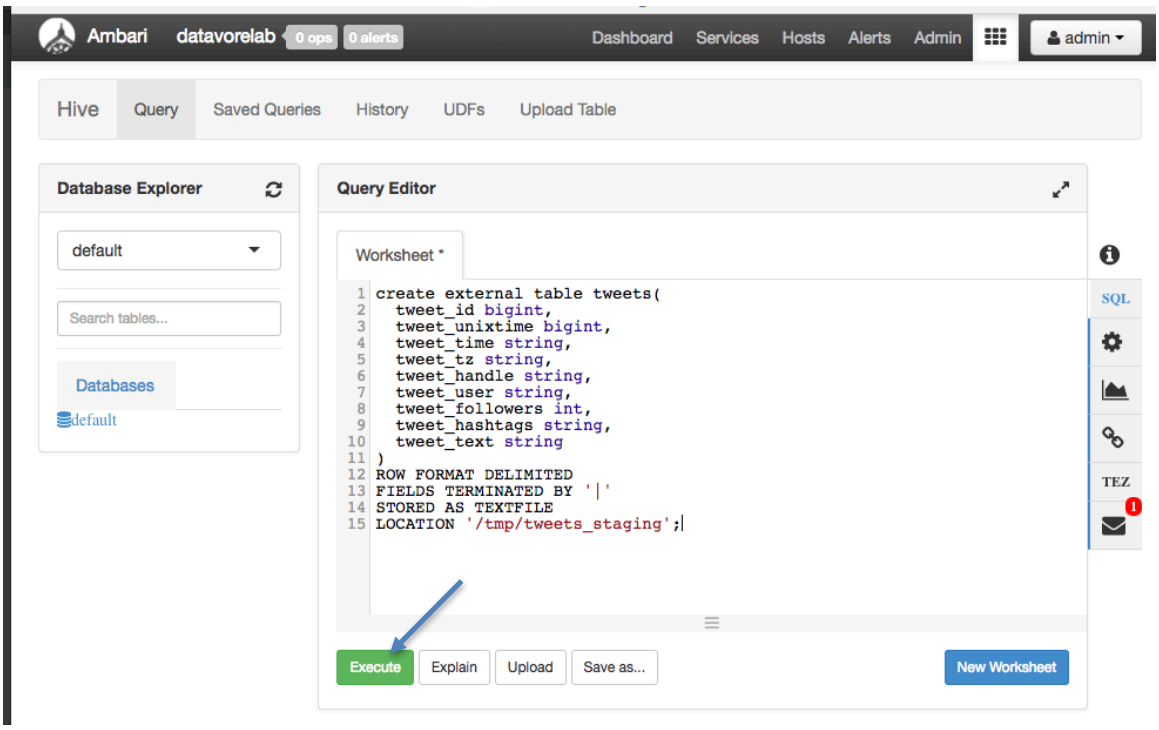


Create the Hive Table

Create the hive table to allow querying the tweets by pasting query available at the following link onto the worksheet:

https://github.com/oascofare/Datavore2018/blob/master/tweets.sql

Click on Execute button:

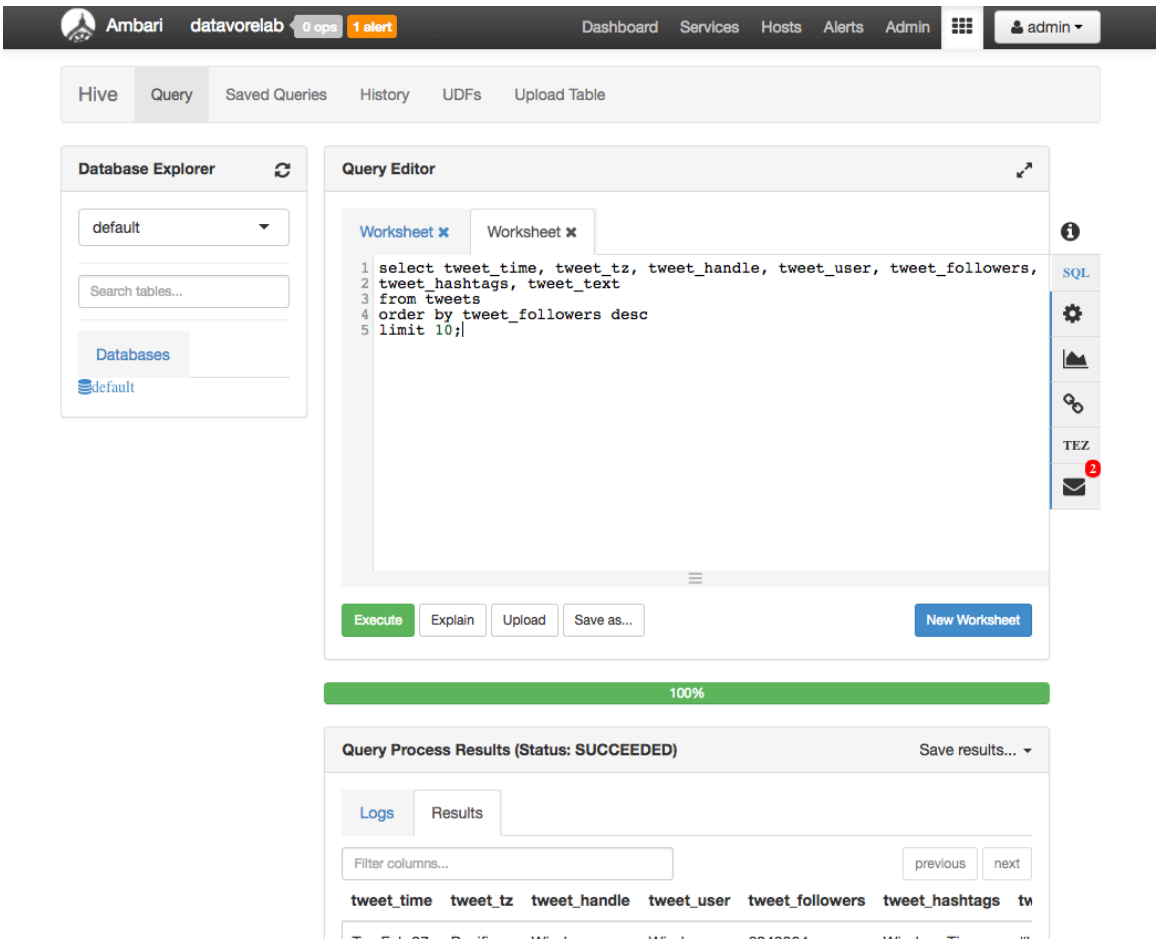


Query Data using SQL on Hive

Click on New Worksheet (blue button on the right-hand side of the screen). Copy and paste the query found at the following link onto the new worksheet:

https://github.com/oascofare/Datavore2018/blob/master/top_tweets.sql

This query will find the top 10 influential tweets (assuming it is based on the number of followers). Click on Execute.

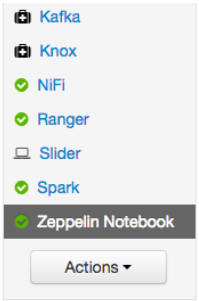


Data Visualization with Zeppelin

In this section, we are going to visualize the twitter data using Zeppelin. SparkSQL will be used to perform data analysis.

All the queries in this section are available on the following link:
https://github.com/oascofare/Datavore2018/blob/master/zeppelin-queries.sql

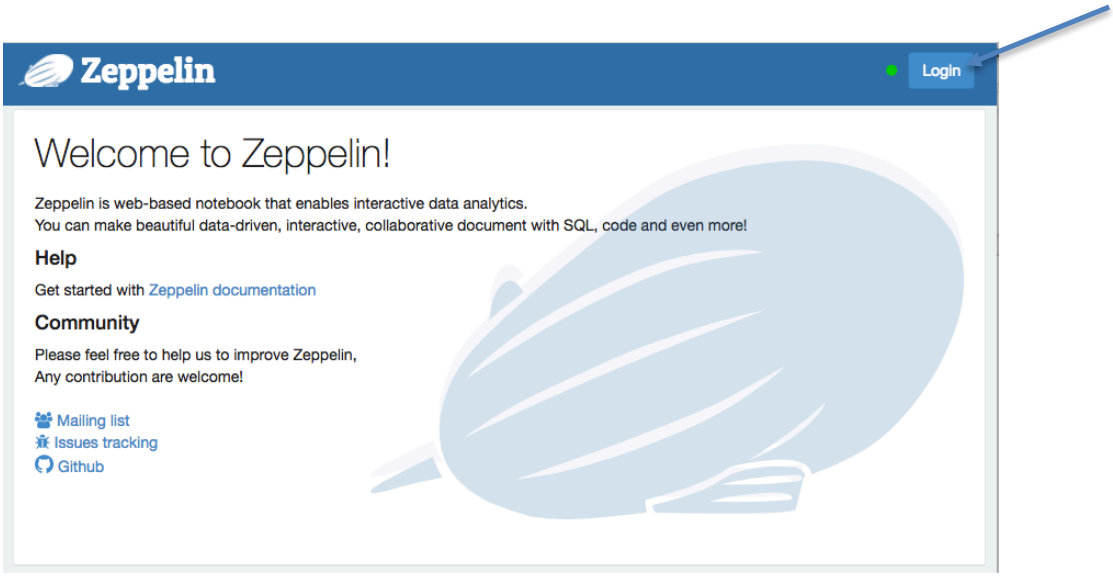
Make sure the Zeppelin service is started in Ambari:



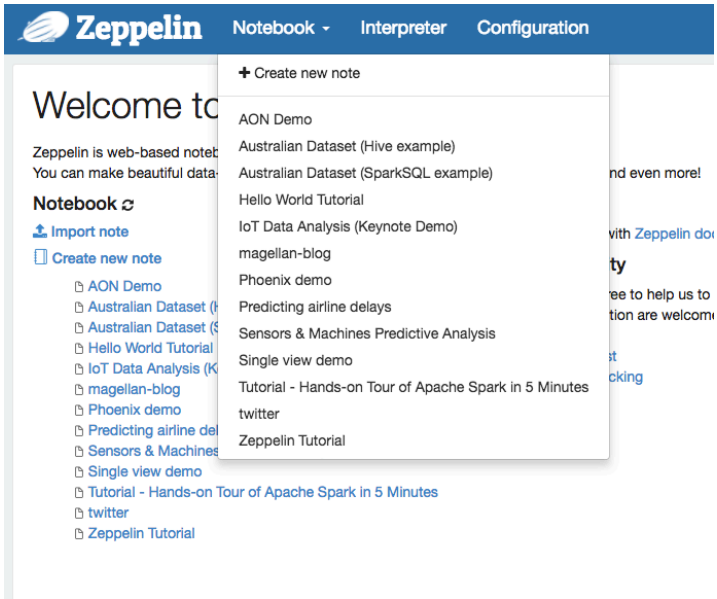
Navigate to Zeppelin using the following URL: http://<public-ip>:9995/#/

Click on Login on the top right-hand side of the screen and use the following credentials:

Username: admin
Password: admin



Use the Notebook drop down menu to create a new note called "Twitter Analysis"



After creating the note, open it up to the blank Notebook screen and type the following command. `%sql` is the SparkSQL interpreter.

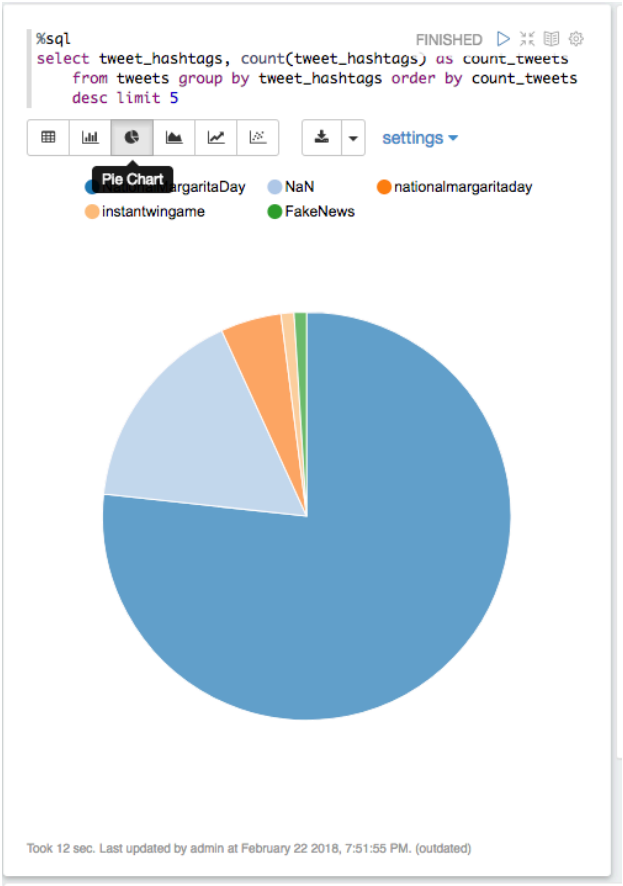
```
%sql
--Top 10 tweets
select tweet_handle, tweet_user, tweet_followers, tweet_hashtags,
tweet_text from tweets order by tweet_followers desc limit 10
```

The query above is similar to the query submitted from Hive view and will give the top 10 influential tweets (based on the number of followers). In this case, Hortonworks is leveraging the Spark processing engine, through the SparkSQL API to perform analytics on the data.

On a new paragraph, the query below will count the top 5 trending hashtags (based on the number of tweets)

```
%sql
--Top 5 trending hashtags
select tweet_hashtags, count(tweet_hashtags) as count_tweets from tweets
group by tweet_hashtags order by count_tweets desc limit 5
```

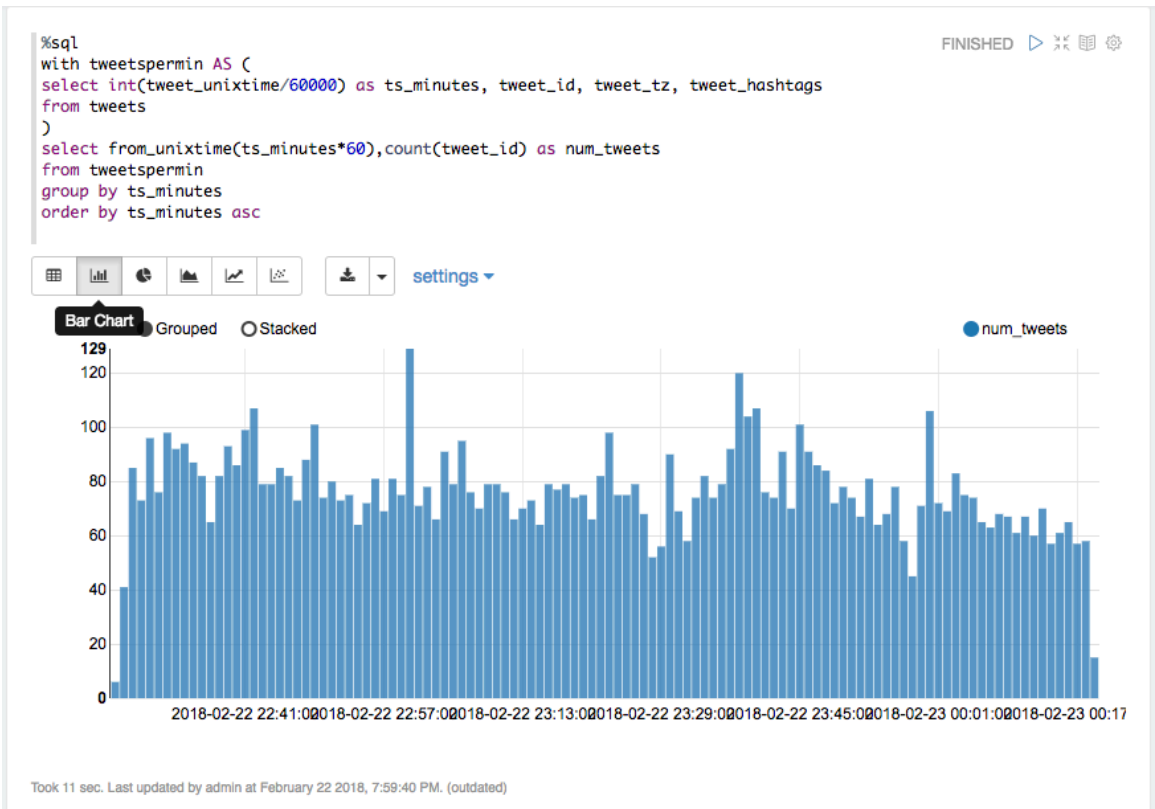
Click on the pie chart icon to visualize the data:



In a next paragraph, we can try to visualize the tweet throughput per minute for the time frame the data has been collected using the following query:

```
%sql
--Tweet Flow per minute
with tweetspermin AS (
select int(tweet_unixtime/60000) as ts_minutes, tweet_id, tweet_tz,
tweet_hashtags
from tweets
)
select from_unixtime(ts_minutes*60),count(tweet_id) as num_tweets
from tweetspermin
group by ts_minutes
order by ts_minutes asc
```

Click on the bar chart icon to visualize the data as a bar chart:



Time allowing, you can optionally add additional paragraphs to this Zeppelin notebook to further explore the data set.

Real-Time Data Visualization with Apache Druid and Superset

Superset is a Business Intelligence tool packaged with many features for designing, maintaining and enabling the storytelling of data through meaningful data visualizations for real-time data.

In this section, we will explore data visualization with Superset for the Twitter dataset, that has been ingested into Druid, through the SAM application.

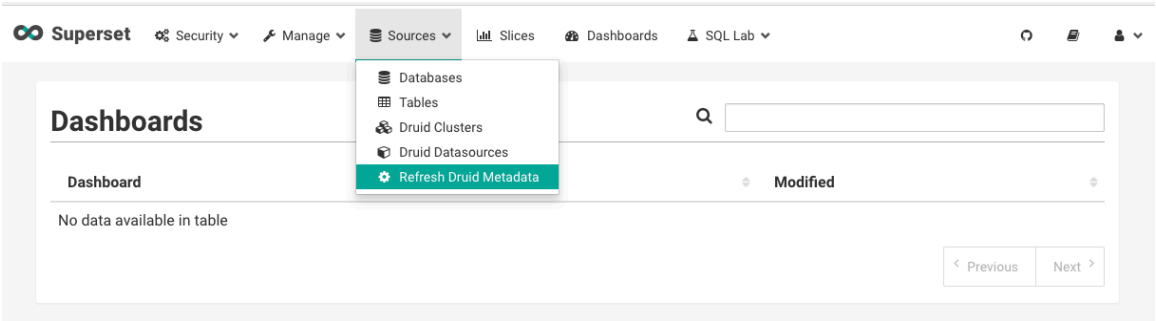
In order to access Superset, use the following URL:
http://<vm-public-ip>:9087/login

In order to login, use the following credentials:

- Username: **admin**
- Password: **admin**

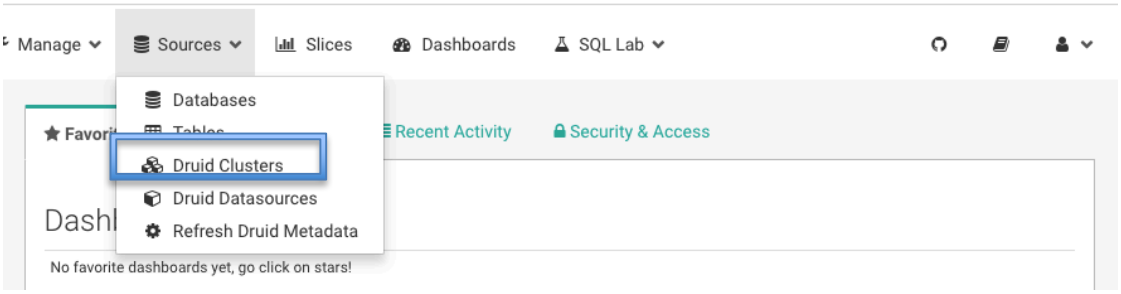
Locate the Druid Twitter datasource

Once on the main menu for Superset, please click on Sources from the menu and select 'Refresh Druid Metadata':

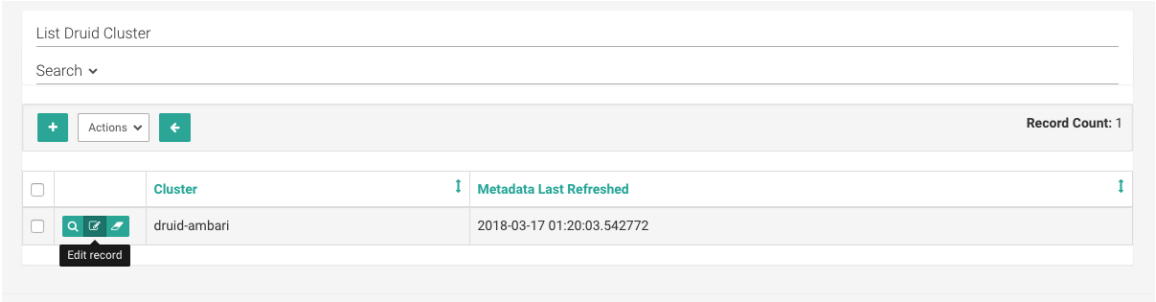


The twitter-dsn data source created by the SAM application should now appear. It takes around 30 minutes for the initialization of the new datasource under Druid:

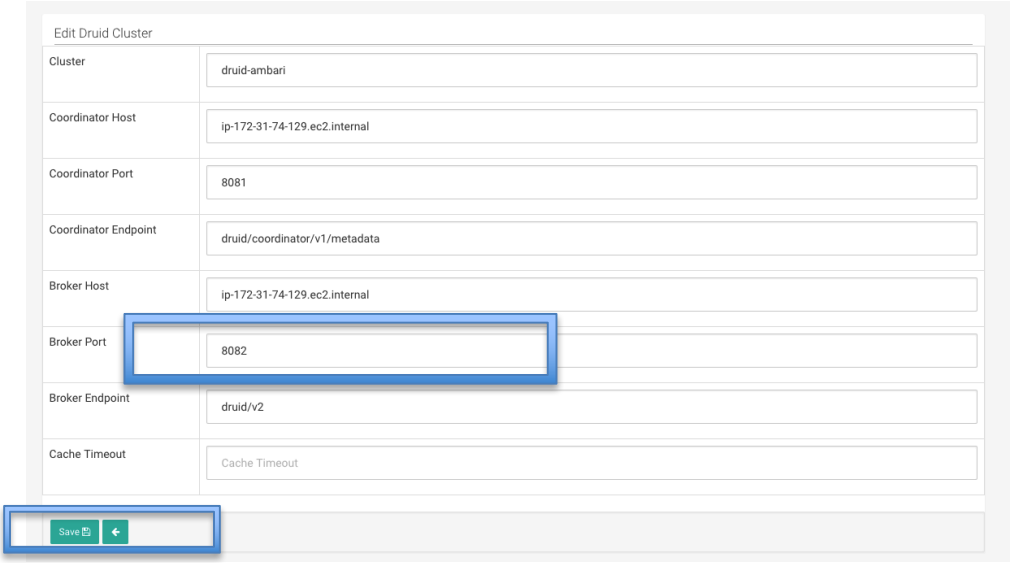
If you can't your data source or get an error when you click on the on twitter-dsn hyperlink ; go on the source menu on the top, then click on Druid Clusters.



Then click on "edit record"



Update the Broker Port to 8082 , then click on save .



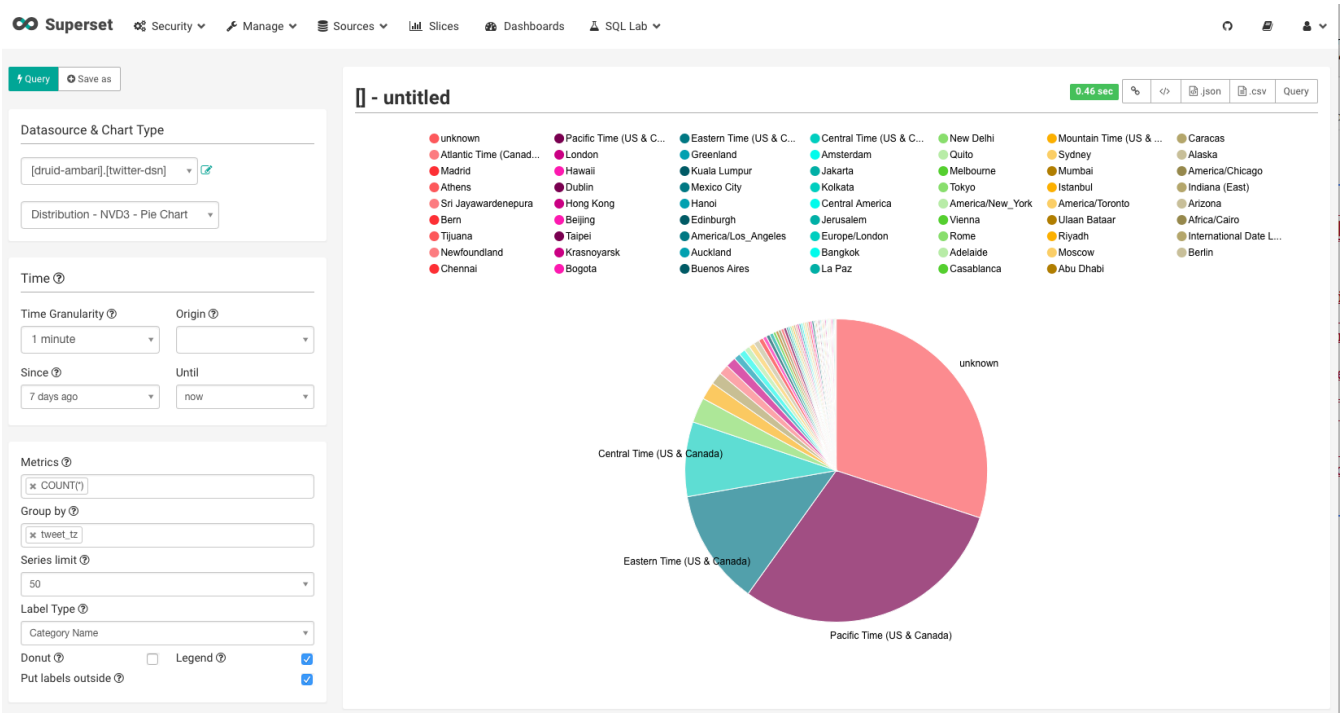
If you still can't see your data source you may want to go back to the previous step to monitor the Storm application and ensure that tuples are written onto Druid. Once confirmed, you should wait for more time before refreshing the Druid Metadata again.

Create a visualization and a dashboard

We can now generate at least one visualization to check the timezone region most active in terms of tweet counts. Click on twitter-dsn hyperlink. A new window will open (See below). Enter the following fields:

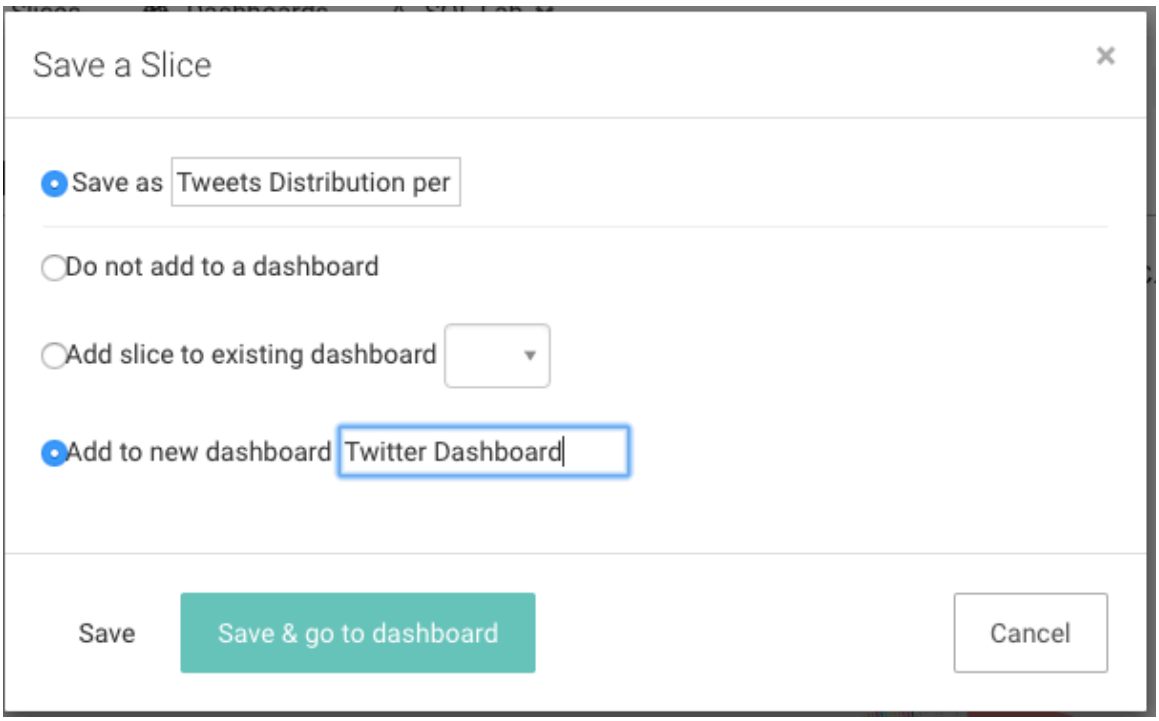
- Chart Type: **Distribution – NVD3 – Pie Chart**
- Time Granularity: **1 minute**
- Since: **7 days ago**
- Metrics: **COUNT(*)**
- Group by: **tweet_tz**

Once complete, click on Query (green button) on the top left of the screen. You should see the visualization appear.

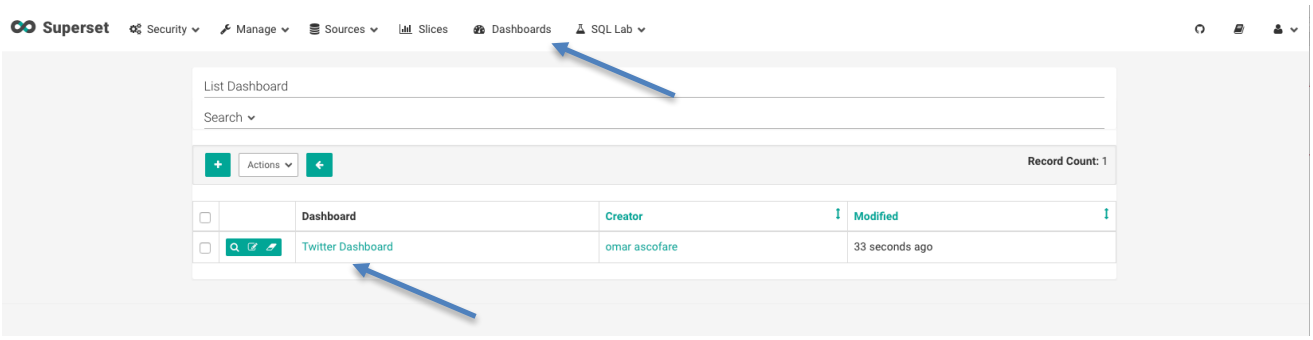


Click on Save as:

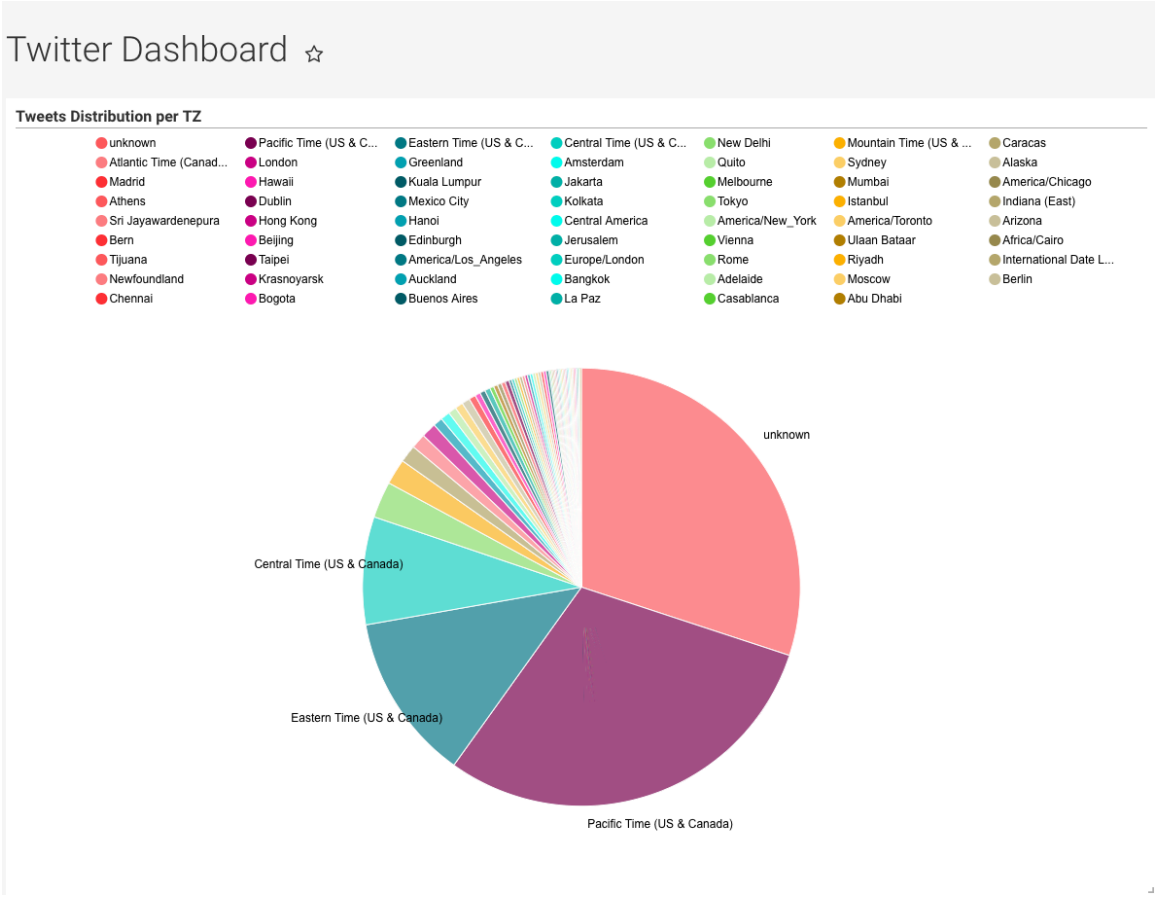
- Save as: **Tweets Distribution per TZ**
- Add to new dashboard: **Twitter Dashboard**



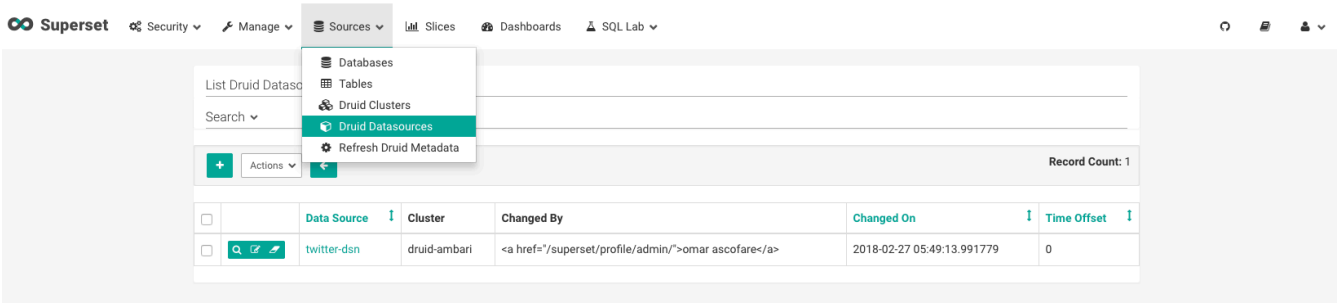
Click on Save. Then, click on Dashboards on the top menu, then click on Twitter Dashboard hyperlink:



You should now see your dashboard with a single visualization:



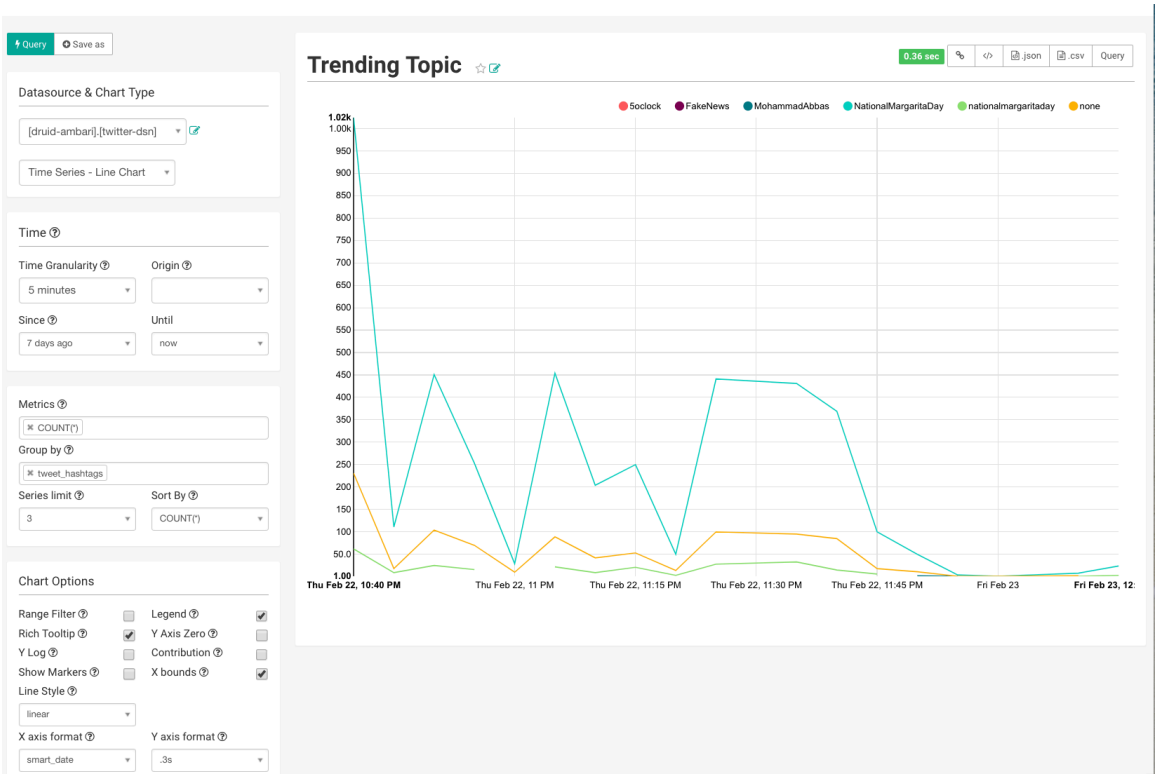
We can add another visualization to the dashboard. Select Sources from the top menu and click on the twitter-dsn hyperlink:



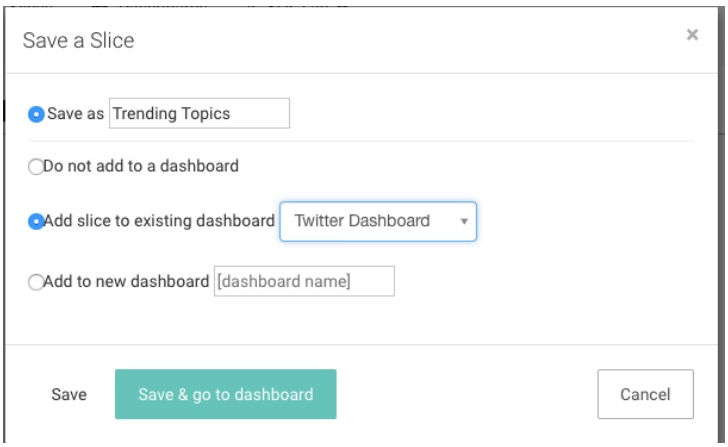
Enter the following fields:

- Chart Type: **Time-Series – Line Chart**
- Time Granularity: **5 minutes**
- Since: **7 days ago**
- Metrics: **COUNT(*)**
- Group by: **tweet_hashtags**
- Series limit: **3**
- Sort By: **COUNT(*)**

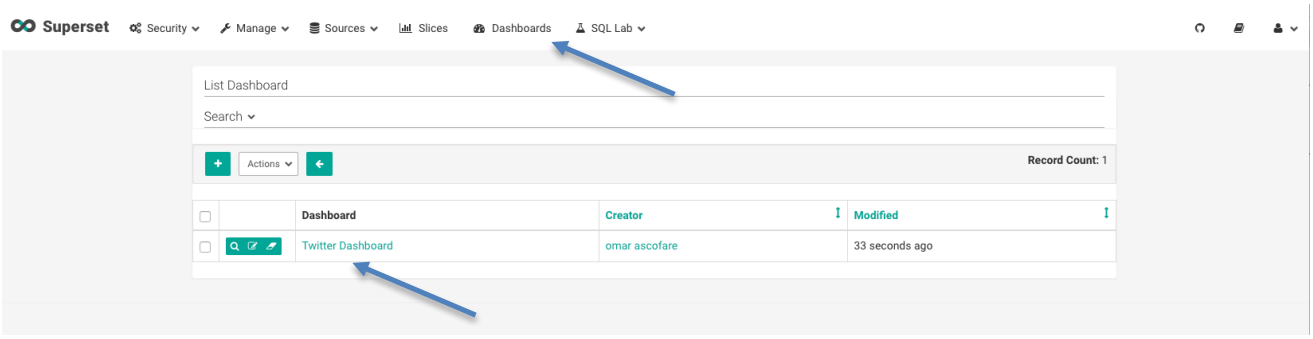
Once complete, click on Query (green button) on the top left of the screen. You should see the visualization appear.



You can now save this slide and add it as part of the Twitter Dashboard. Click the 'Save as' button:



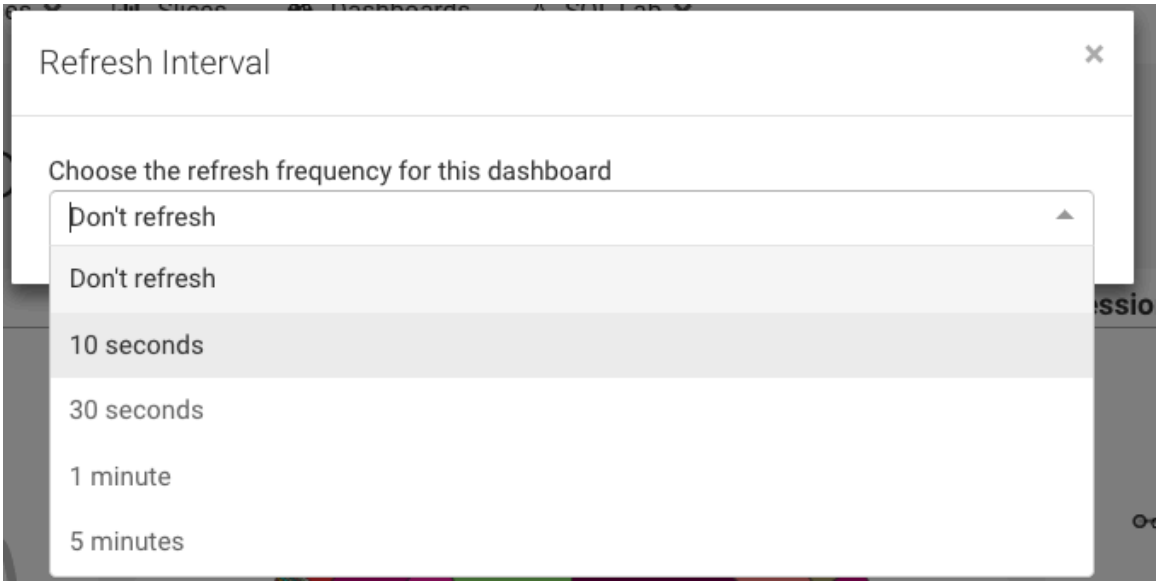
Click on Save. Then, click on Dashboards on the top menu, then click on Twitter Dashboard hyperlink:



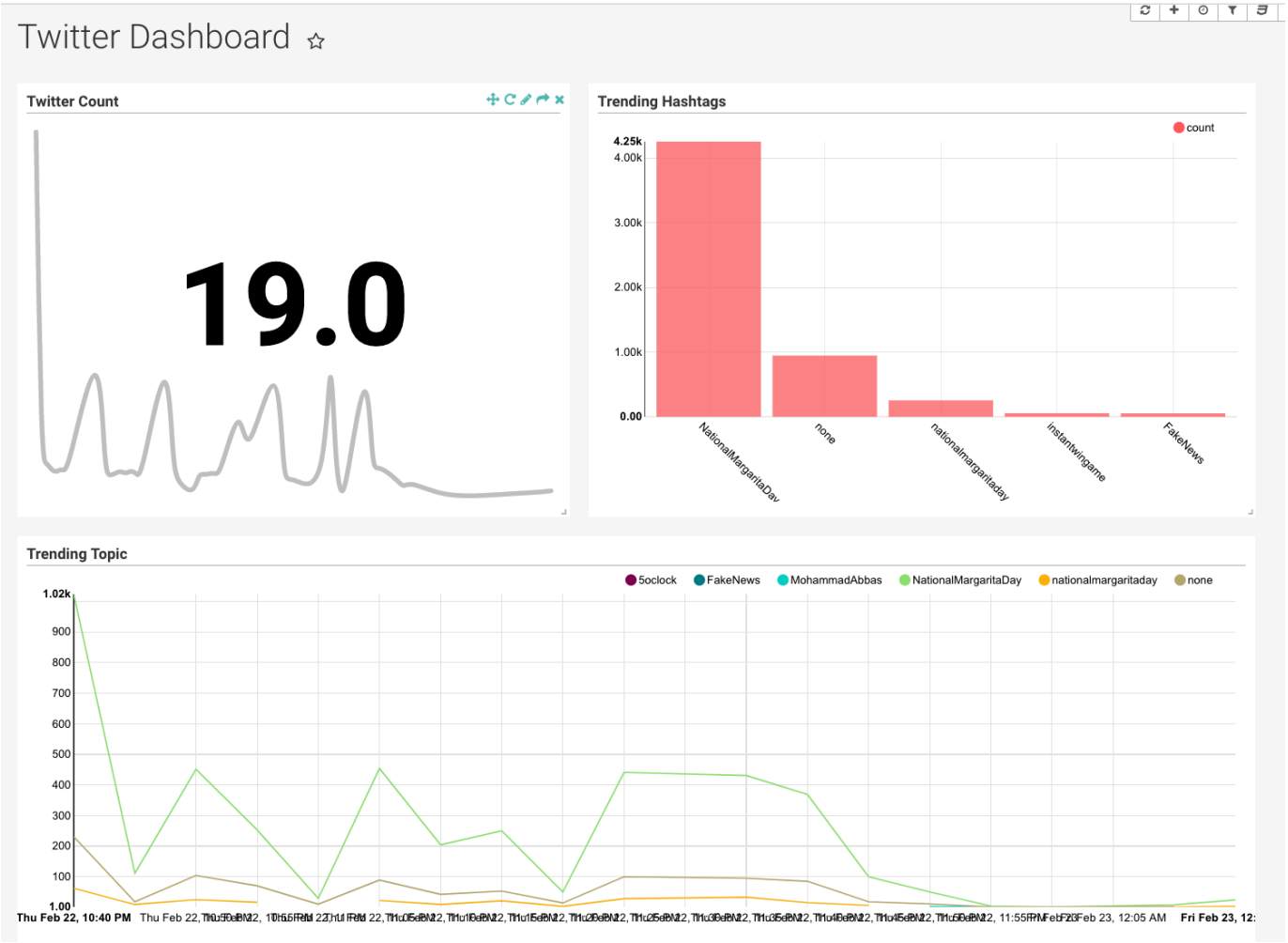
It may take a few minutes to execute the initial queries required for your visualization, and see the new visualization added on your dashboard.

Note that these queries can be refreshed automatically at a set interval to get a dynamic dashboard. This interval can be set if you click on the clock icon on the top right of the dashboard:

You can then optionally choose the refresh interval:



If time permits, please consider adding additional visualization such as tweet count, or a histogram showing the trending hashtags. Below are additional examples of visualizations:


Twitter Dashboard
Twitter Count
19.0
Trending Hashtags
count
4.25k
4.00k
3.00k
2.00k
1.00k
0.00
NationalMargaritaDay
none
nationalmargaritaday
instantwingame
FakeNews
Trending Topic
5oclock
FakeNews
MohammadAbbas
NationalMargaritaDay
nationalmargaritaday
none
1.02k
900
800
700
600
500
400
300
200
100
1.00
Thu Feb 22, 10:40 PM
Fri Feb 23, 12:05 AM