Using HDP and HDF to Harness the Power of Social Media

Ce laboratoire vous permettra d' implementer la collecte et l'analyse des données Twitter en exploitant les solutions Hortonworks Data Platform 2.6 (HDP) et Hortonworks DataFlow 3.0 (HDF), en combinant les données en mouvement et les données au repos. Bien que les analyses initiales soient simples, ce modèle active les blocs de construction pour des cas d'utilisation plus avancés.

Au cours de cet atelier, les participants seront en mesure d'effectuer l'ingestion de données en direct sur Twitter en fonction de mots clés, filtrage, transformation, extraction d'attributs et transmission à HDFS pour le stockage de données, ainsi qu'Apache Kafka utilisant Apache NiFi.

Les participants pourront ensuite effectuer des analyses sur des données au repos à l'aide de Hive, de SparkSQL ainsi que d'Apache Zeppelin à des fins de visualisation et d'exploration de données.

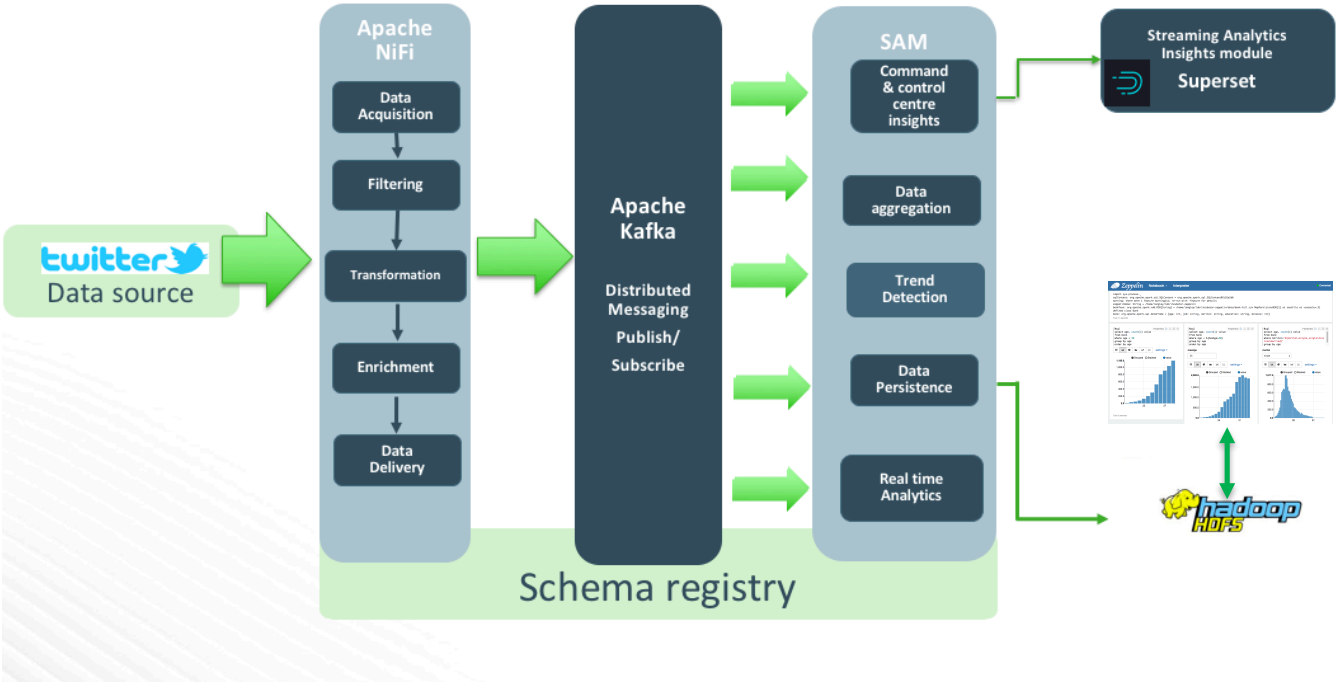
Une fois les données ingérées dans Kafka, les participants seront en mesure de développer rapidement une application Streaming Analytics Manager afin d'analyser davantage les données et d'en tirer des informations en temps réel.

Les données seront également visualisées en temps quasi réel avec Apache Druid et Superset.

Overview ..... **Error! Bookmark not defined.**
Pre-requisites ..... **Error! Bookmark not defined.**
Create your Twitter API keys ..... **Error! Bookmark not defined.**
Create A Twitter Application ..... **Error! Bookmark not defined.**
Verify your Hortonworks Sandbox web access .. **Error! Bookmark not defined.**
Using Apache NiFi for ingesting Twitter Data 9
Download the Twitter dataflow template ..... **Error! Bookmark not defined.**
Import the Template ..... **Error! Bookmark not defined.**
Instantiate the Template 12
Configure the Twitter Grab Garden Hose Processor 13
Configure PUTHDFS 14
Review the remaining Processors 16
Start the Flow ..... **Error! Bookmark not defined.**
Check the data on HDFS ..... **Error! Bookmark not defined.**
Data Analysis with Hive ..... **Error! Bookmark not defined.**
Create the Hive Table ..... **Error! Bookmark not defined.**
Query Data using SQL on Hive ..... **Error! Bookmark not defined.**
Data Visualization with Zeppelin ..... **Error! Bookmark not defined.**
Real-Time Data Visualization with Apache Druid and Superset 27
Locate the Druid Twitter datasource 27
Create a visualization and a dashboard 29

Vue d'ensemble

Vous trouverez ci-dessous un diagramme conceptuel de la mise en œuvre de Twitter Analytics réalisée dans ce laboratoire à l'aide de Hortonworks Data Flow et de Hortonworks Data Platform.



Conditions préalables

Dans la mesure du possible, les conditions préalables suivantes doivent être remplies avant le début du laboratoire. Veuillez suivre les instructions du document sur les conditions préalables du laboratoire. Les étapes préalables incluses dans les documents sont les suivantes:

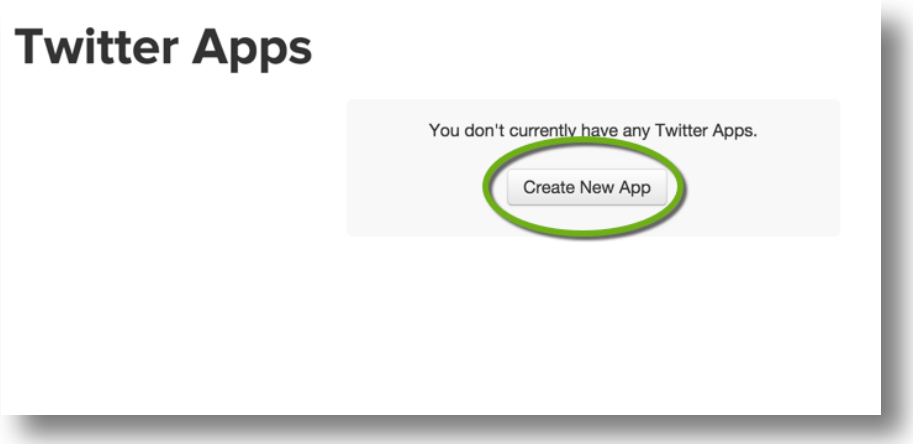
- **La création d'un compte d'abonnement Azure Pass.**
- **Créez vos clés API Twitter.**
- **Vérifiez votre accès Web Hortonworks Sandbox.**

Créez vos clés API Twitter

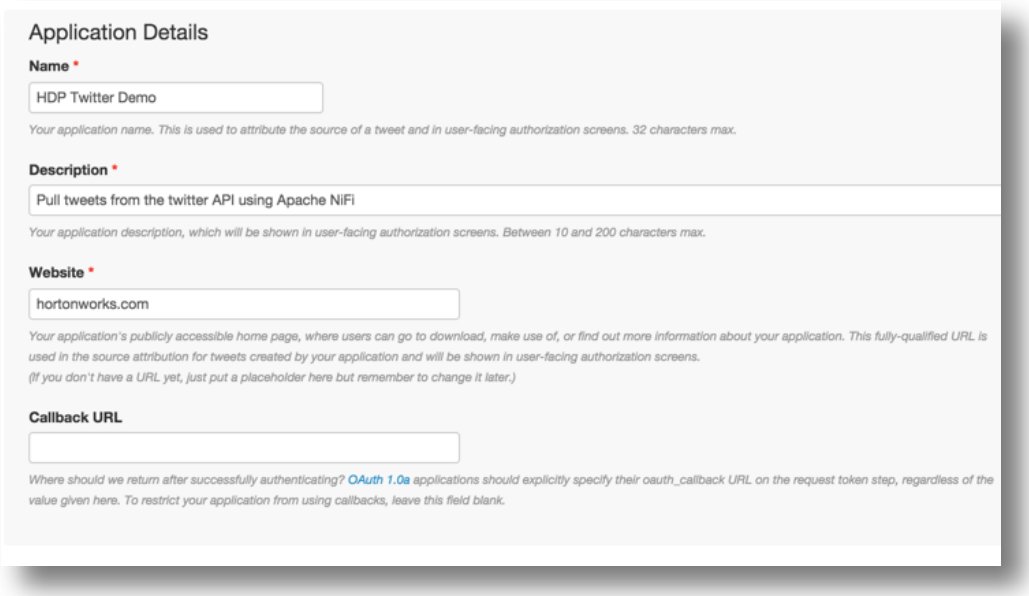
Vous devrez enregistrer votre propre application Twitter. C'est assez simple et ne prend que quelques étapes

Créer une application Twitter

Rendez-vous d'abord sur le site Web des applications Twitter (https://apps.twitter.com/
) et connectez-vous à l'aide de votre compte Twitter (ou créez-en un si vous n'en avez pas encore!). Puis cliquez sur Créer une nouvelle application.



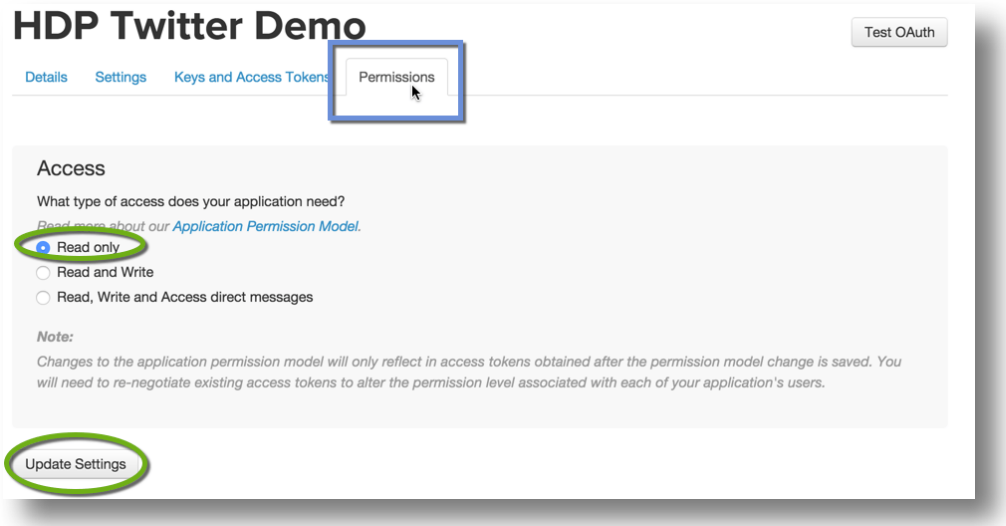
Une fois que vous avez cliqué, vous devrez fournir des détails sur votre candidature. N'hésitez pas à mettre ce que vous voulez.



Cliquez ensuite sur Créer votre application Twitter au bas de l'écran après avoir lu l'accord de développeur.

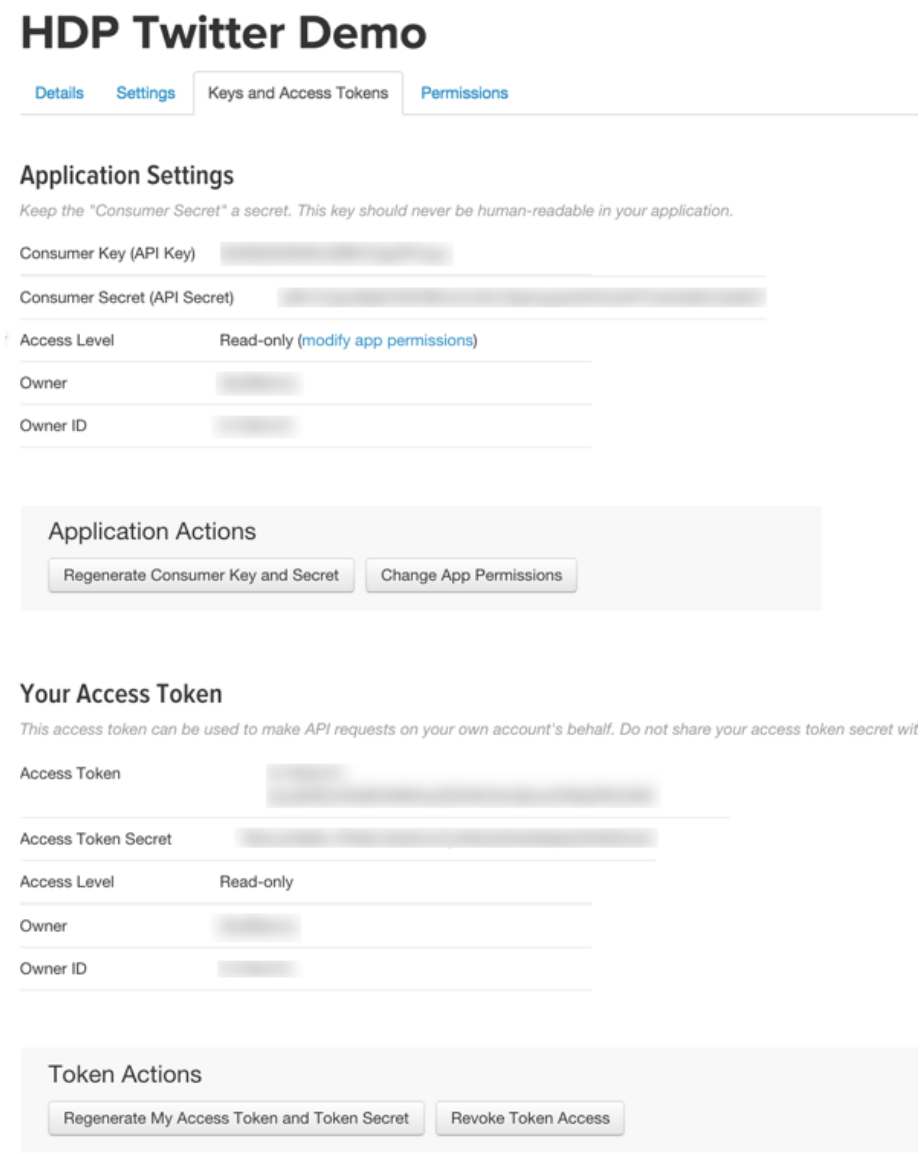
Notez que vous devrez peut-être ajouter votre téléphone mobile à votre compte Twitter avant de créer votre application.

Une fois que vous avez terminé, vous devriez être accueilli par un tableau de bord pour votre application Twitter. Rendez-vous sur l'onglet Autorisations, sélectionnez l'option Lecture seule et mettez à jour votre application.



Enfin, vous devez générer votre clé OAuth. Vous pouvez le faire en cliquant sur Test OAuth en haut de la page des autorisations, ou en sélectionnant Clés et jetons d'accès, puis en recherchant l'option permettant de générer vos jetons OAuth.

Vos clés et vos jetons d'accès doivent ressembler à ce qui suit:

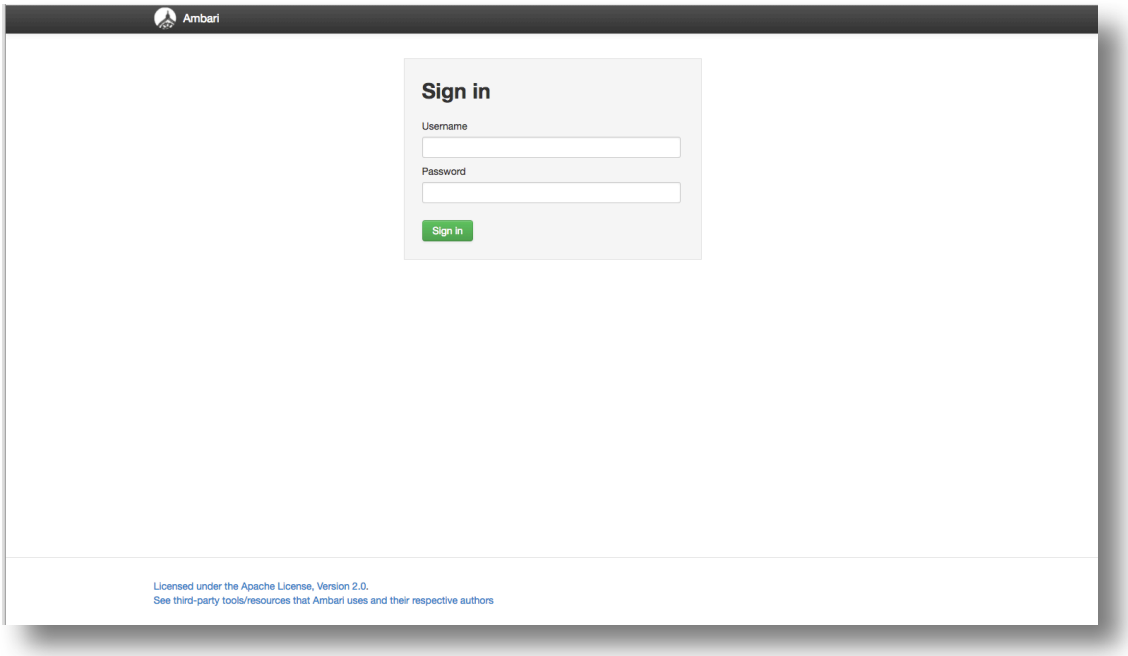


Veuillez noter votre clé de consommateur, votre **Consumer Key**, **Consumer Secret**, **Access Token**. Vous en aurez besoin pour créer le flux de données dans NiFi.

Vérifiez votre accès Web a la machine Hortonworks

Une fois ces étapes préalables effectuées, assurez-vous que votre machine virtuelle est démarrée sur Azure en y accédant via un navigateur Internet à l'adresse.
Http: // <sandbox-ip>: 8080.

Vous devez vous attendre à la page Web suivante:



Si la machine virtuelle est en panne, veuillez contacter l'instructeur de laboratoire pour obtenir une nouvelle machine.

Using Apache NiFi for ingesting Twitter Data

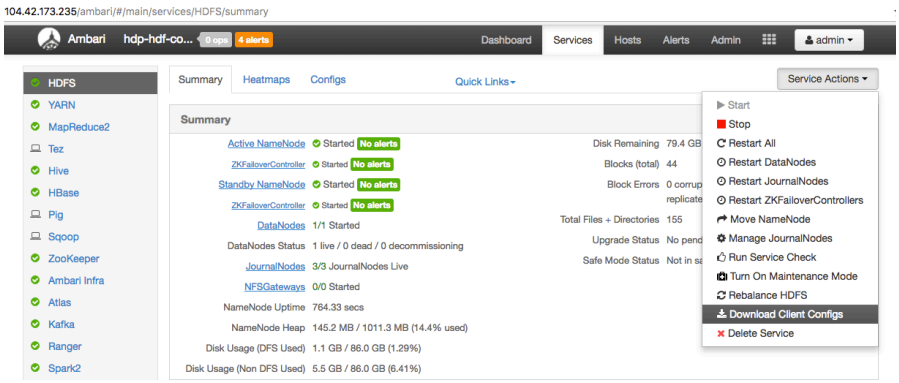
HDF (NiFi) dispose d'un système de modèles puissant permettant un partage et une consommation aisés des flux de données. La section suivante porte sur l'importation d'un modèle de flux de données Twitter préparé, sur la personnalisation des paramètres en fonction de votre environnement (informations d'identification Twitter, mots clés Twitter, adresses IP) et démarre la collecte de Tweets dans HDFS.

Conditions préalables

1. Téléchargez les fichiers de configuration suivants Copiez-les sur la Nifi exemple sous / home / nifi / et positionnez les propriétaires de répertoires pour 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 :



Assurez-vous que l'instance Nifi peut atteindre les services HDP requis (hive, hdfs, altas, kafka broker) via le pare-feu ou les groupes de sécurité réseau.

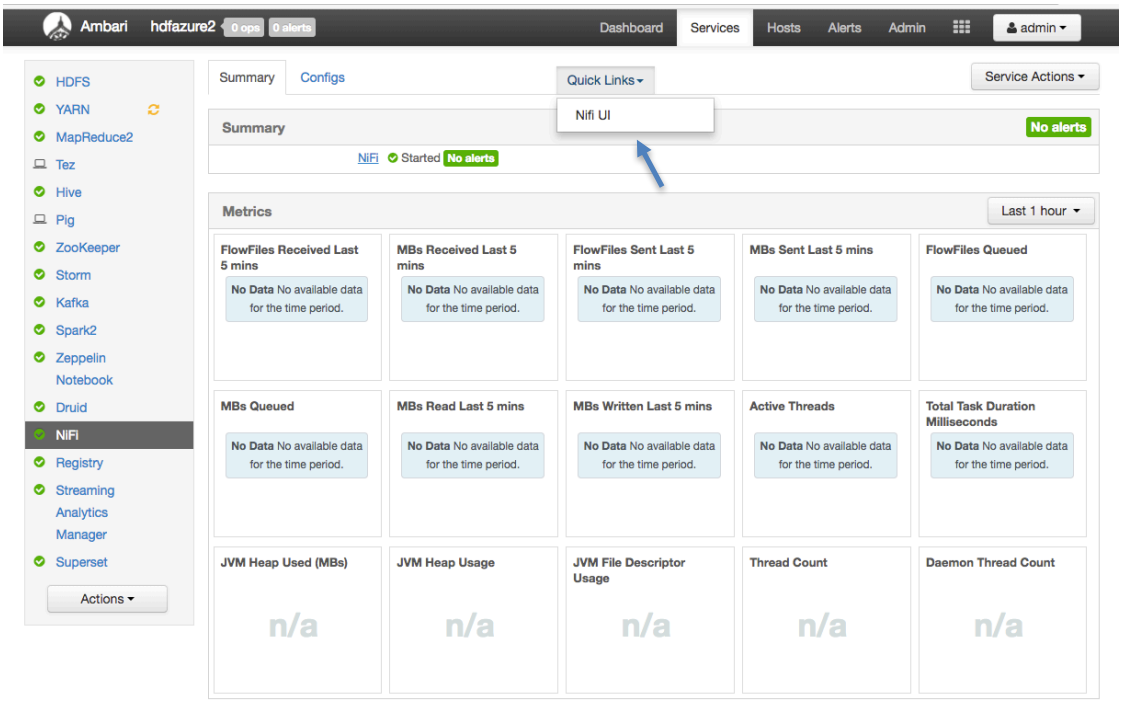
Téléchargez le modèle de flux de données Twitter

Le template de flux de données Twitter doit être téléchargé sur le poste de travail accédant à NiFi au lien suivant.

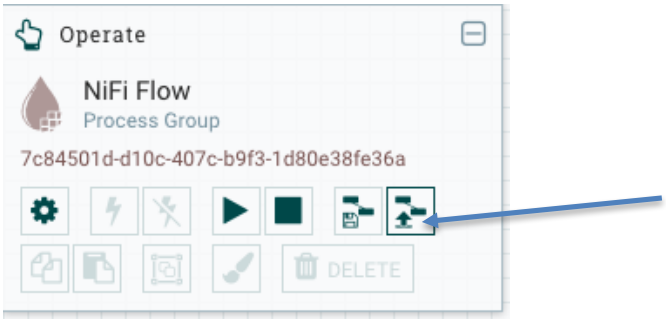
https://github.com/oascofare/Datavore2018/raw/master/DatavoreTwitterApplication.xml

Importer le modèle

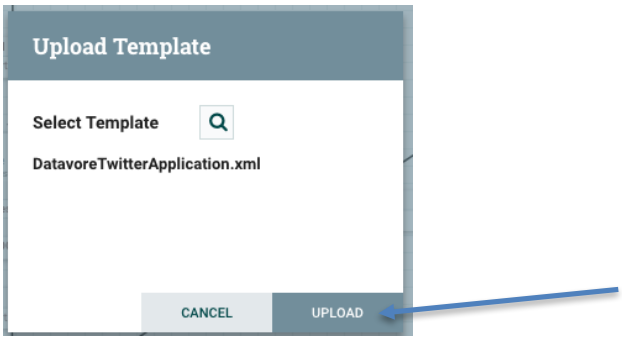
Cliquez sur le lien du service Nifi Ne naviguez pas non plus vers l'interface Web NiFi à l'adresse : *http://<public ip>: 9999/nifi/* .



Importez le modèle en sélectionnant l'icône Importer un Template située dans la mosaïque NiFi Flow Operate sur le côté gauche du canevas.

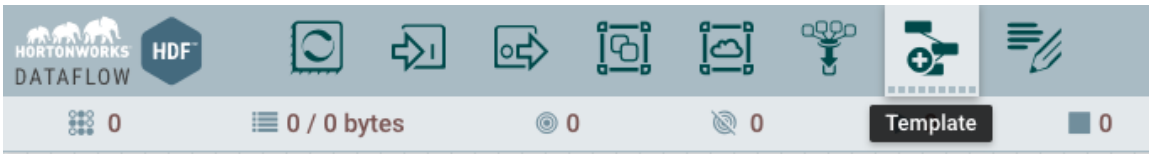


Cliquez sur l'icône Parcourir ( ) et accédez au site où le fichier DatavoreTwitterApplication.xml a été téléchargé. Cliquez sur Upload.

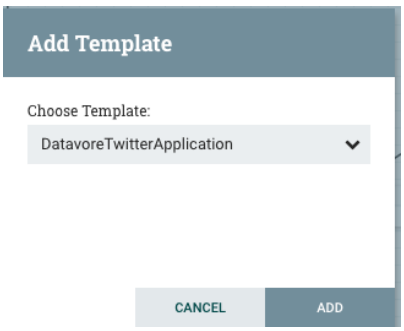


Instancier le modèle de Template

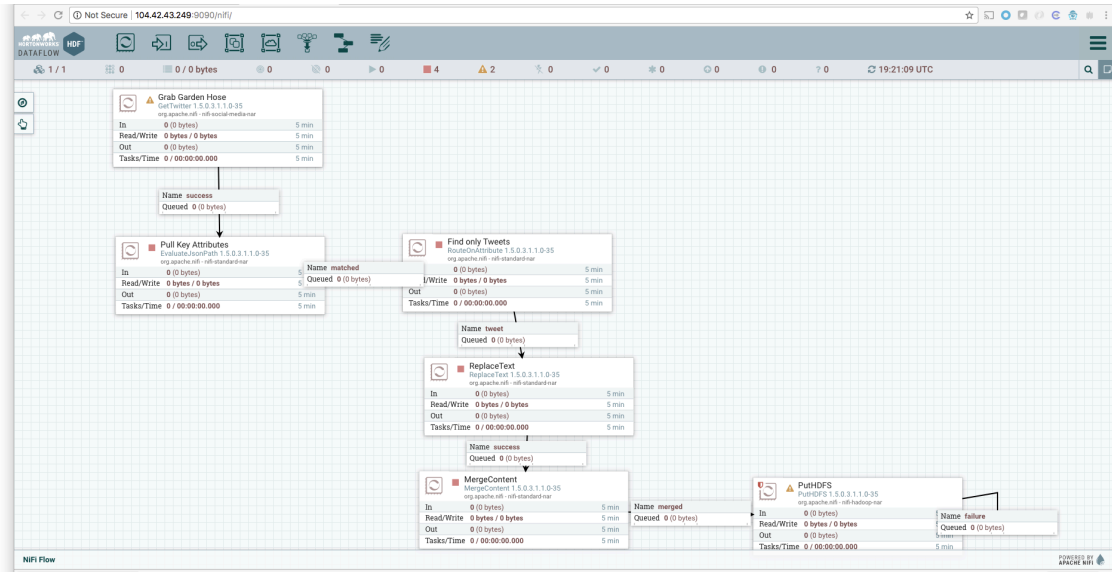
Faites glisser et déposez l'artefact de template sur la toile.



Sélectionnez DatavoreTwitterApplication et cliquez sur Ajouter (*Add)*

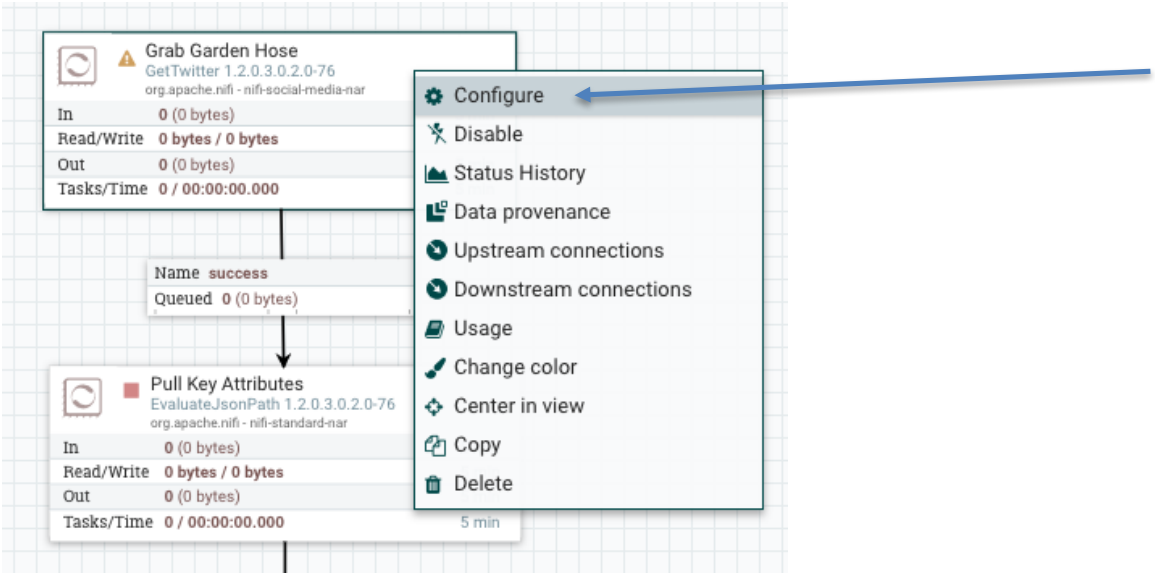


Cela importera le template du flux de données sur votre canevas NiFi.



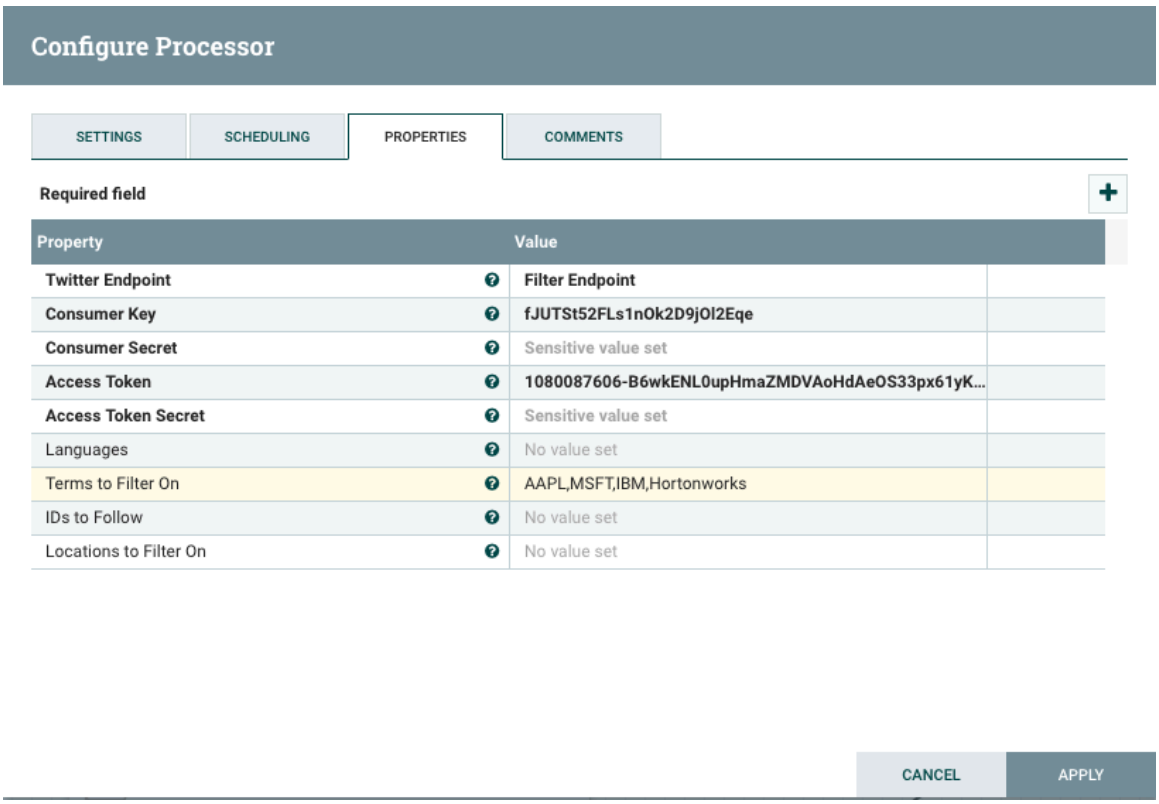
Configurer le Twitter Grab Garden Hose Processor

Ensuite, configurez Twitter Grab Garden Hose Processer avec les informations de l'API Twitter. Cliquez avec le bouton droit sur le processeur et sélectionnez Configurer. Ensuite, sélectionnez l'onglet PROPERTIES

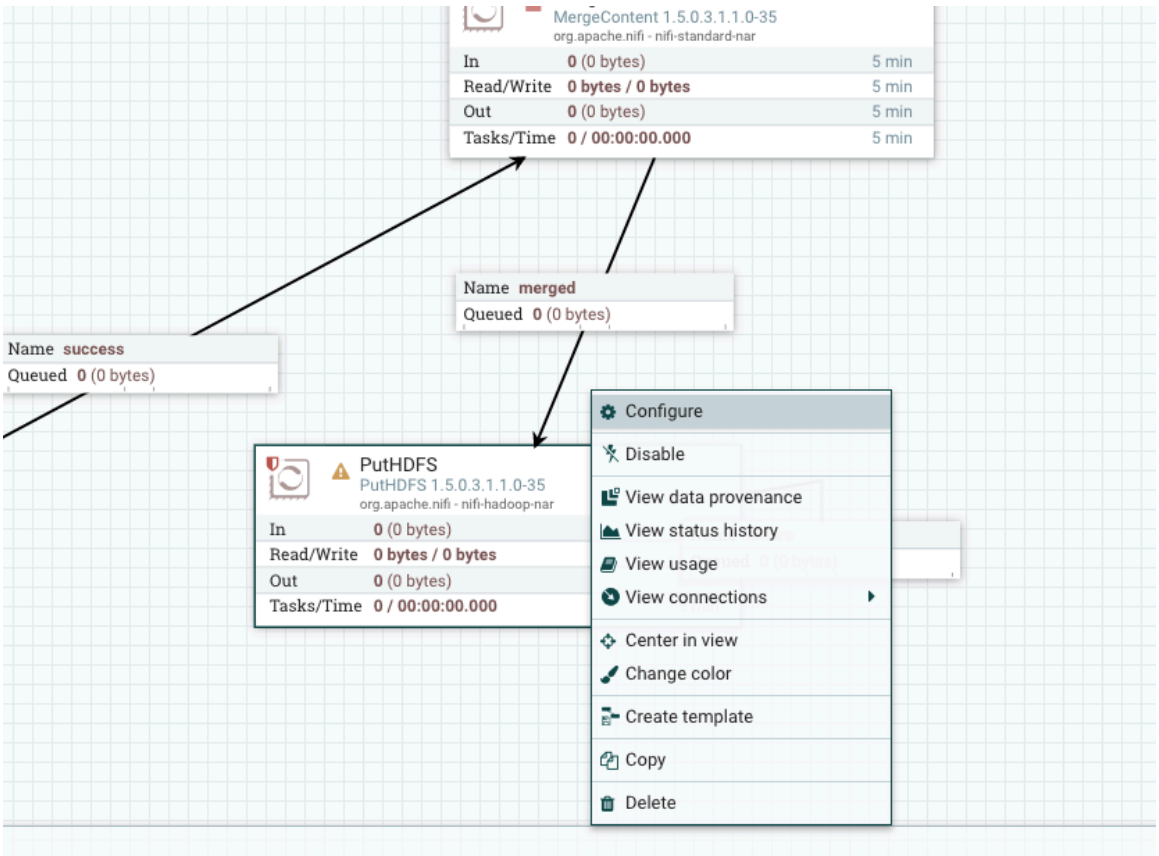


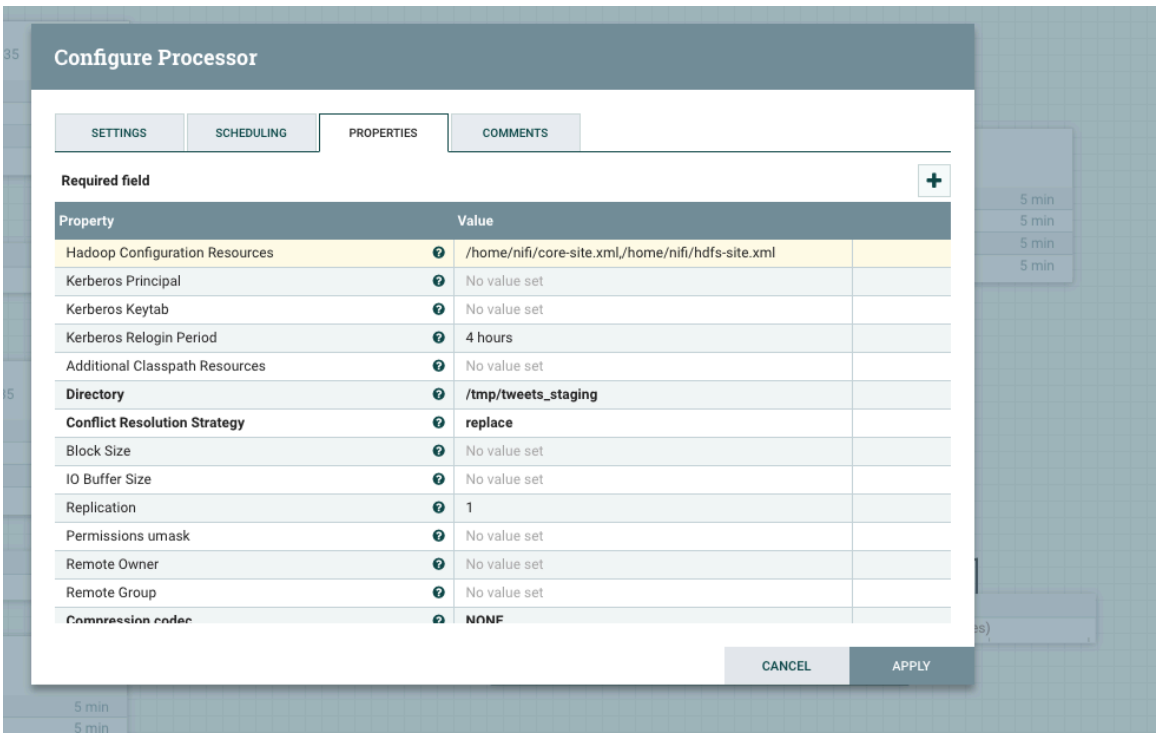
Fill in the following mandatory properties (in bold) matching to your Twitter application:

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



Configurer 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



Revoir le reste Processors

Passez en revue les autres processeurs et modifiez les propriétés si nécessaire:

- **EvaluateJsonPath**: Extrait les attributs des tweets
- **RouteonAttribute**: Assurer que seuls les tweets contenant des messages non vides sont traités
- **ReplaceText**: Formate chaque tweet en tant qu'entrée de ligne délimitée par des tubes (|), par exemple. tweet_id | tweet_unixtime | tweet_stringtime} | tweet_timezone | tweet_screenname | tweet_username | tweet_followers | tweet_hashtags

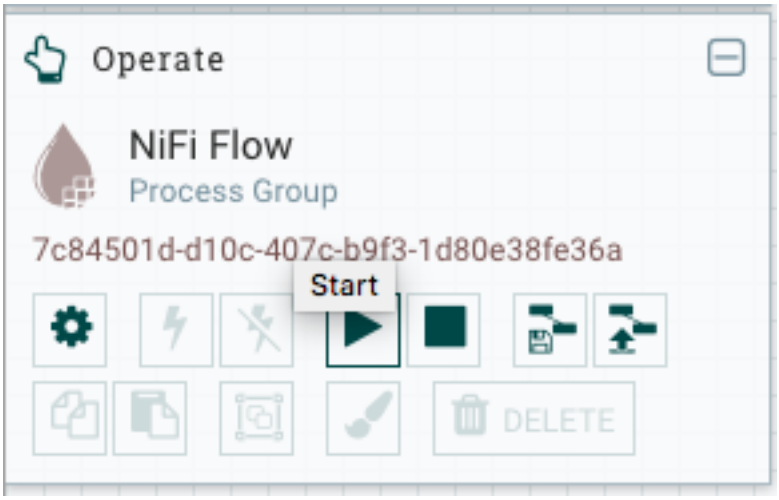
- **MergeContent**: Fusionne les tweets dans un seul fichier (20 ou 120, selon la première éventualité) pour éviter d'avoir un grand nombre de petits fichiers dans HDFS. Ces valeurs peuvent être configurées.

- **PutHDFS**: écrit des tweets sur HDFS sous **/tmp/tweets_staging**

- **UpdateAttribute**: Ajoutez des attributs supplémentaires aux fichiers de flux, qui sont le nom du schéma pointant vers le nom du schéma dans le registre de schéma(schema registry), ainsi que le nom du topic kafka où les tweets sont écrits.

Démarrer le flux

Si les processeurs sont correctement configurés, tous les processeurs doivent avoir un symbole d'arrêt rouge en haut à gauche (sauf le vert pour PublishKafkaRecord, car ce processeur particulier a été démarré en tant que composant de référence des services du contrôleur). Si un symbole d'avertissement jaune est présent, cela signifie que la configuration requise est manquante. Lancez le flux en sélectionnant une zone vide du canevas et en cliquant sur l'icône de démarrage.

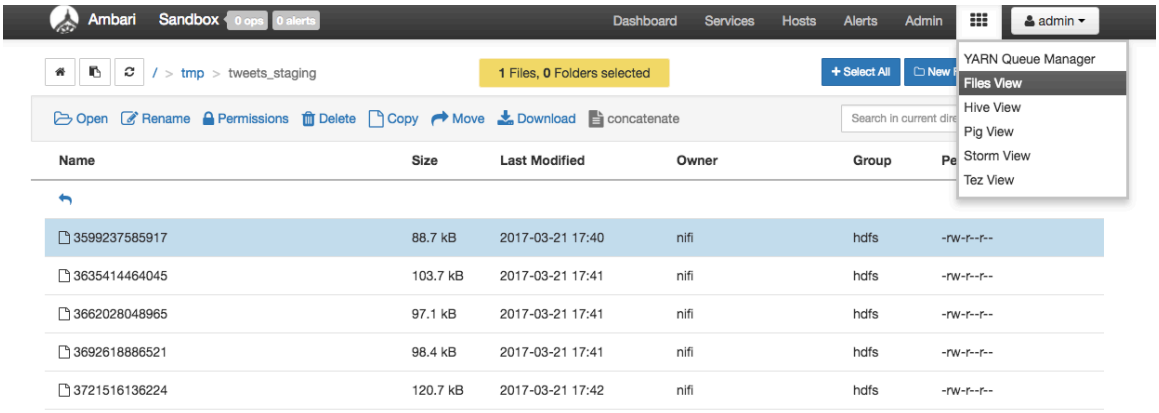


Vérifiez les données sur HDFS

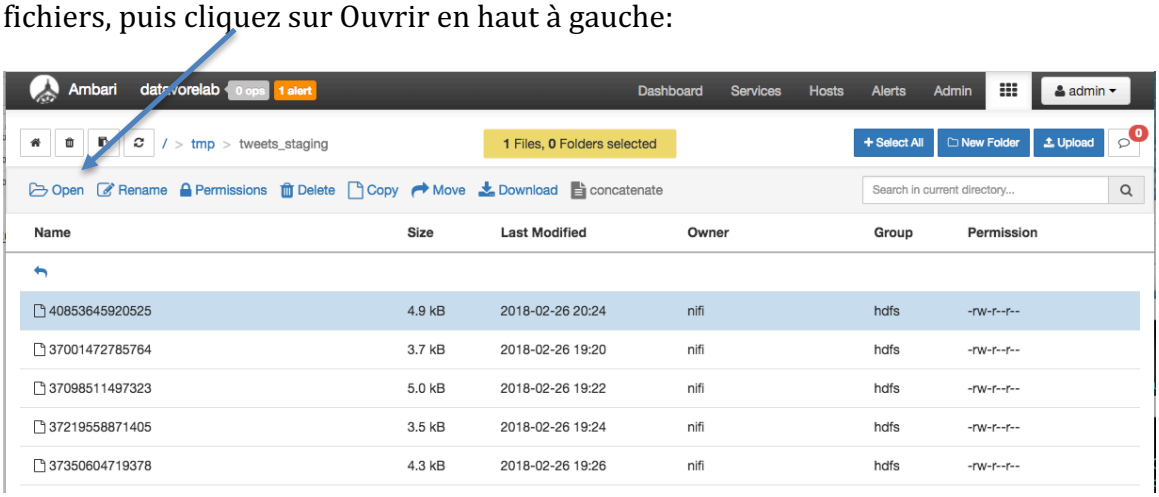
Maintenant que le flux de données Twitter est en cours d'exécution, les tweets correspondants seront conservés vers HDFS et Kafka. Validez que les tweets ont bien été écrits sur HDFS pour le moment.

Les tweets apparaissent sous le repertoire **/tmp /tweets_staging** dans HDFS.

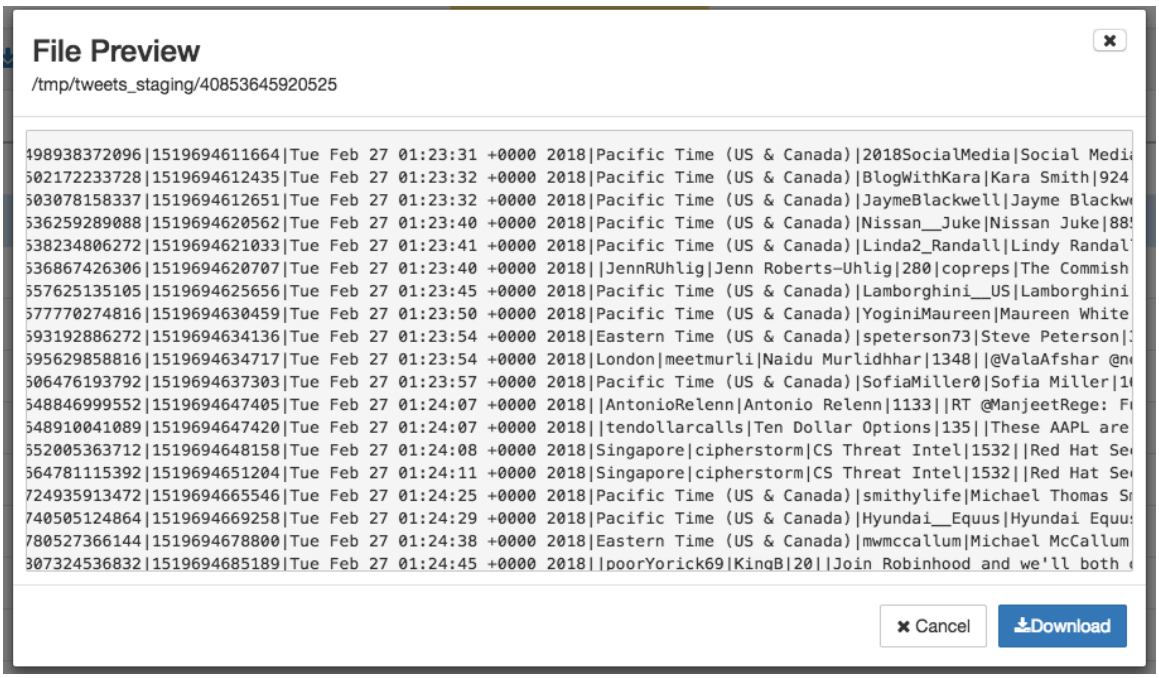
Vous pouvez le voir via la vue Fichiers dans Ambari:



Vous pouvez ouvrir le fichier pour visualiser son contenu. Sélectionnez l'un des fichiers, puis cliquez sur Ouvrir en haut à gauche:



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

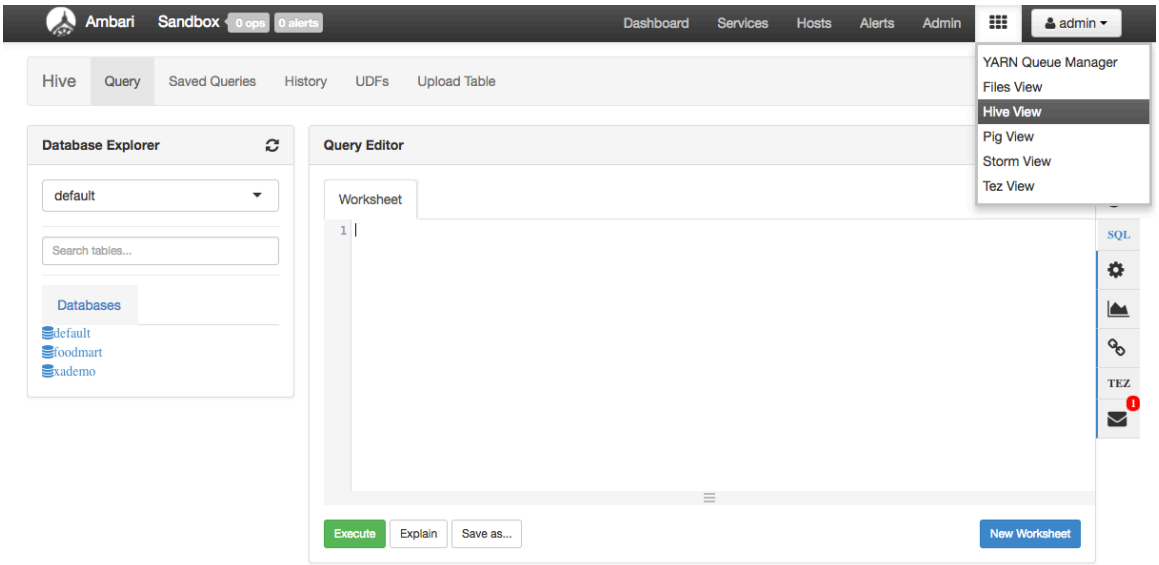
```
498938372096|1519694611664|Tue Feb 27 01:23:31 +0000 2018|Pacific Time (US & Canada)|2018SocialMedia|Social Media
502172233728|1519694612435|Tue Feb 27 01:23:32 +0000 2018|Pacific Time (US & Canada)|BlogWithKara|Kara Smith|924
503078158337|1519694612651|Tue Feb 27 01:23:32 +0000 2018|Pacific Time (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

Analyse de données avec Hive

Après la persistance des tweets dans HDFS, la possibilité d'effectuer des analyses sur les données est presque illimitée. En utilisant Hive, SQL peut être utilisé pour analyser les Tweets. Les sections suivantes créeront la table Hive et exécuteront des requêtes de base sur les tweets.

Vous pouvez accéder à Hive via la vue Hive disponible sur Ambari:

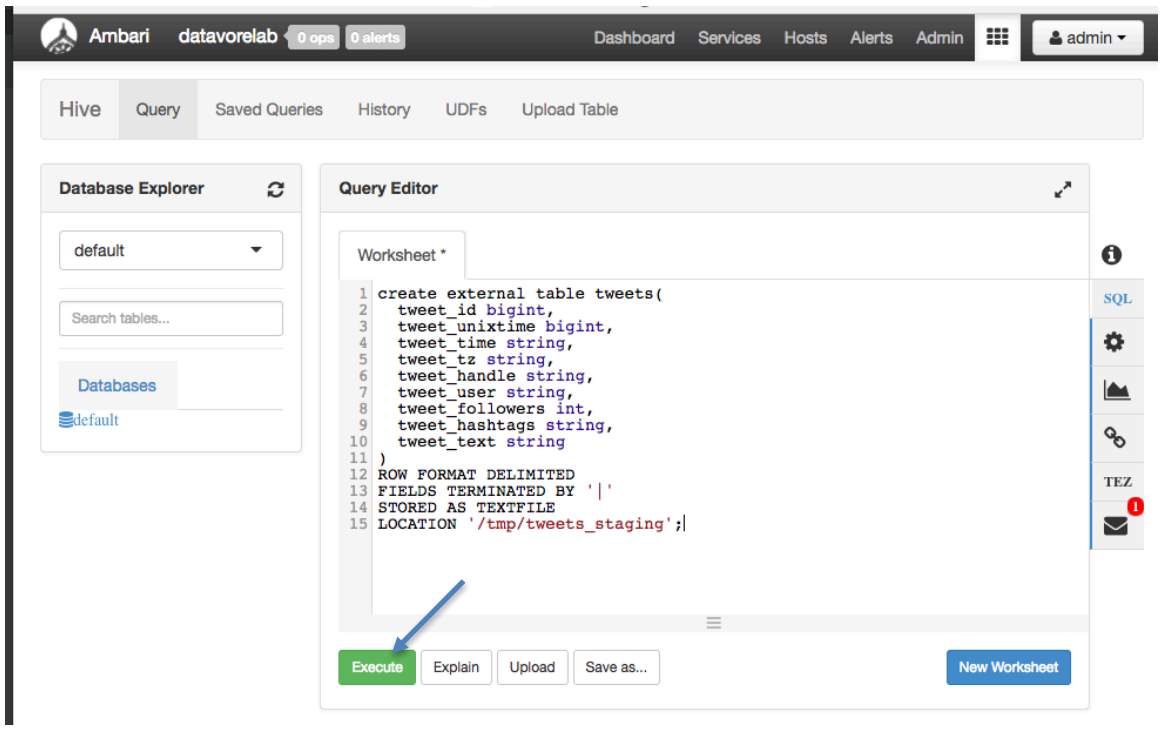


Créer la table Hive

Créez la table ruche pour permettre d'interroger les tweets en collant la requête disponible au lien suivant sur la feuille de calcul:

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

Cliquez sur le bouton Exécuter:

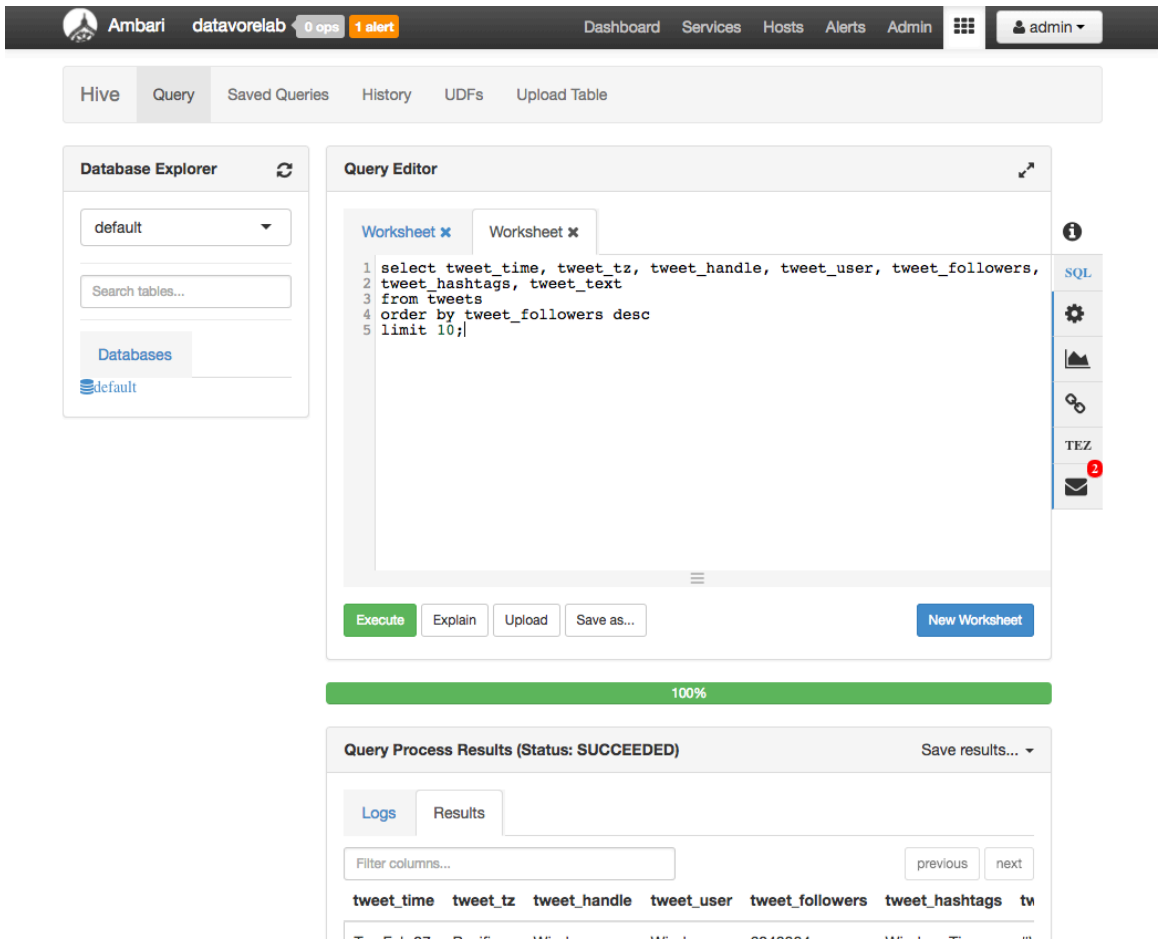


Interroger des données à l'aide de SQL sur Hive

Cliquez sur Nouvelle feuille de calcul (bouton bleu à droite de l'écran). Copiez et collez la requête trouvée sur le lien suivant dans la nouvelle feuille de calcul:

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

Cette requête trouvera les 10 tweets les plus influents (en fonction du nombre d'adeptes). Cliquez sur Exécuter.

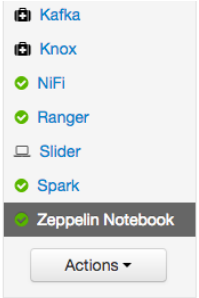


Visualisation de données avec Zeppelin

Dans cette section, nous allons visualiser les données Twitter à l'aide de Zeppelin. SparkSQL sera utilisé pour analyser les données.

Toutes les requêtes de cette section sont disponibles sur le lien suivant:
https://github.com/oascofare/Datavore2018/blob/master/zeppelin-queries.sql

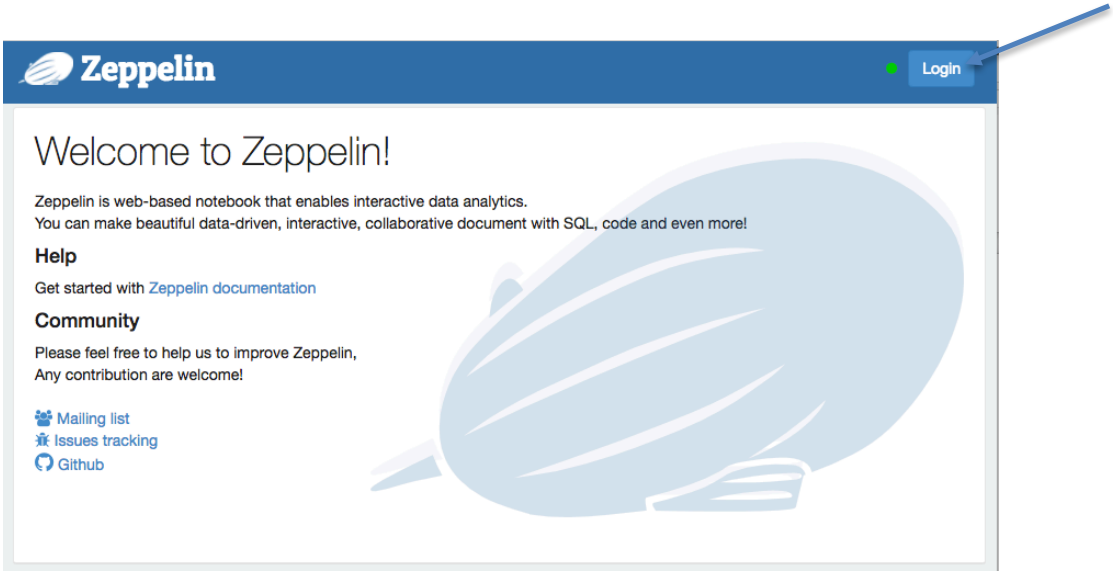
Assurez-vous que le service Zeppelin est démarré à Ambari:



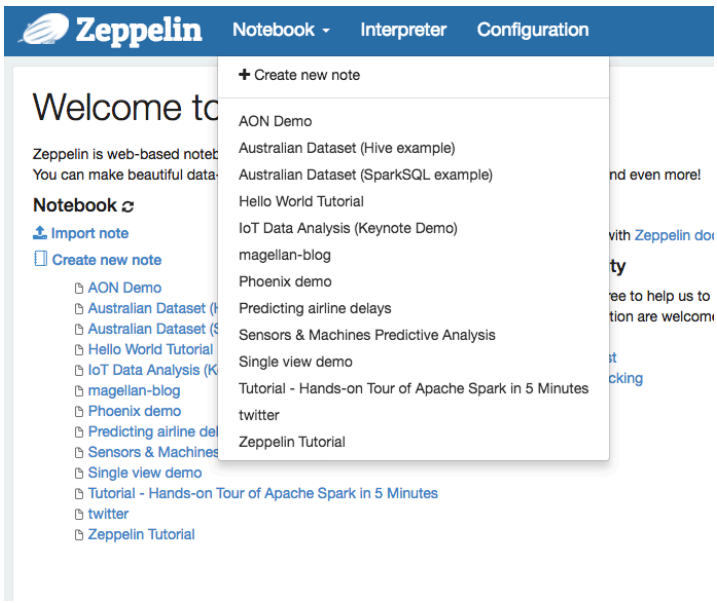
Accédez à Zeppelin en utilisant l'URL suivante: http://<public-ip>:9995/#/

Cliquez sur Connexion en haut à droite de l'écran et utilisez les informations d'identification suivantes:

Username: admin
Password: admin



Utilisez le menu déroulant Bloc-notes pour créer une nouvelle note appelée : "Twitter Analysis"



Après avoir créé la note, ouvrez-la à l'écran Bloc-notes vide et tapez la commande suivante. **%sql** est l'interpréteur SparkSQL.

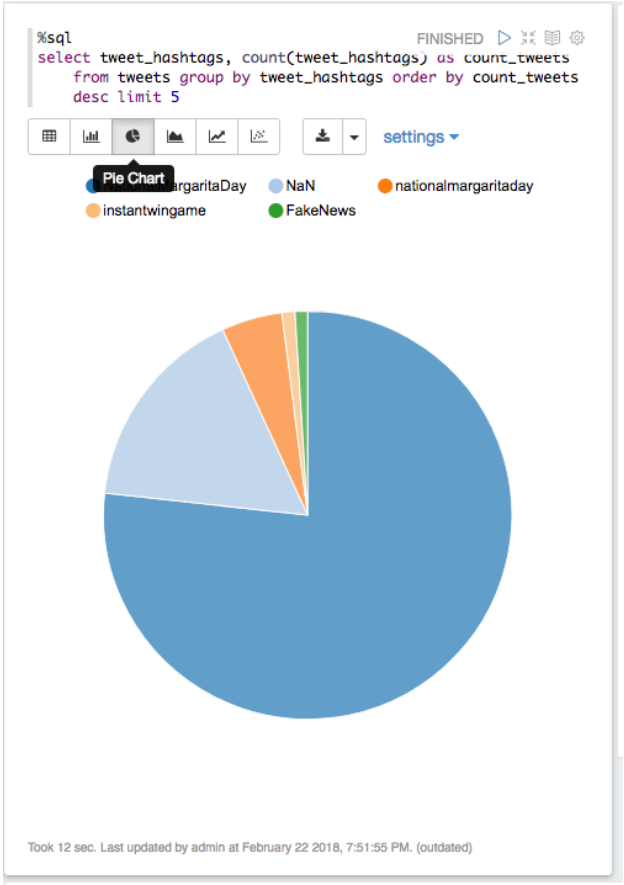
```
%sql
--Top 10 tweets
select tweet_handle, tweet_user, tweet_followers, tweet_hashtags,
tweet_text from tweets order by tweet_followers desc limit 10
```

La requête ci-dessus est similaire à la requête soumise à partir de la vue Hive et donnera les 10 tweets les plus influents (en fonction du nombre de followers). Dans ce cas, Hortonworks utilise le moteur de traitement Spark, via l'API de SparkSQL, pour effectuer des analyses sur les données.

Sur un nouveau paragraphe, la requête ci-dessous comptera les 5 hashtags les plus populaires (basés sur le nombre de 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
```

Cliquez sur l'icône du graphique à secteurs pour visualiser les données:



Dans un paragraphe suivant, nous pouvons essayer de visualiser le débit de tweet par minute pour la période pendant laquelle les données ont été collectées en utilisant la requête suivante:

```
%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
```

Cliquez sur l'icône du graphique à barres pour visualiser les données sous forme de graphique à barres:

```
%sql
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
```

FINISHED

settings

Bar Chart Grouped Stacked num_tweets

129 120 100 80 60 40 20 0

2018-02-22 22:41:00 2018-02-22 22:57:00 2018-02-22 23:13:00 2018-02-22 23:29:00 2018-02-22 23:45:00 2018-02-23 00:01:00 2018-02-23 00:17

Took 11 sec. Last updated by admin at February 22 2018, 7:59:40 PM. (outdated)

Si le temps vous le permet, vous pouvez éventuellement ajouter des paragraphes supplémentaires à ce bloc-notes Zeppelin pour explorer davantage l'ensemble de données.

Real-Time Data Visualization with Apache Druid and Superset

Superset est un outil de Business Intelligence doté de nombreuses fonctionnalités pour la conception, la maintenance et la narration de données via des visualisations de données significatives pour des données en temps réel.

Dans cette section, nous allons explorer la visualisation de données avec Superset pour le jeu de données Twitter, qui a été ingéré dans Druid, via l'application SAM.

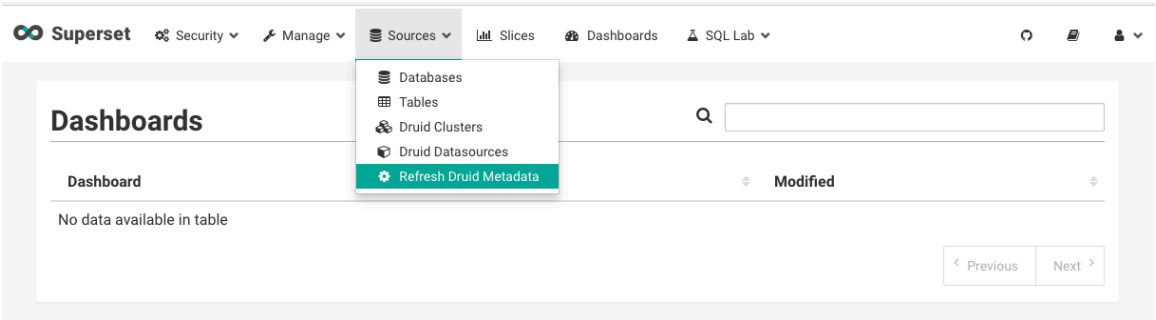
Pour accéder à Superset, utilisez l'URL suivante:
http://<vm-public-ip>:9087/login

Pour vous connecter, utilisez les informations d'identification suivantes:

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

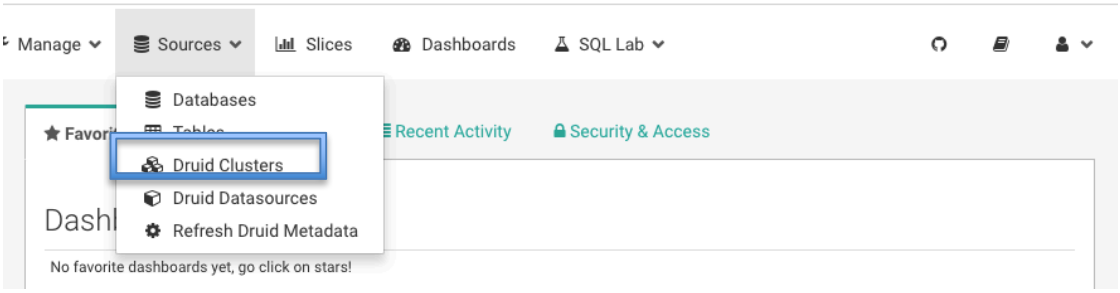
Locate the Druid Twitter datasource

Une fois dans le menu principal de Superset, cliquez sur Sources dans le menu, puis sélectionnez 'Refresh' Druid Metadata :

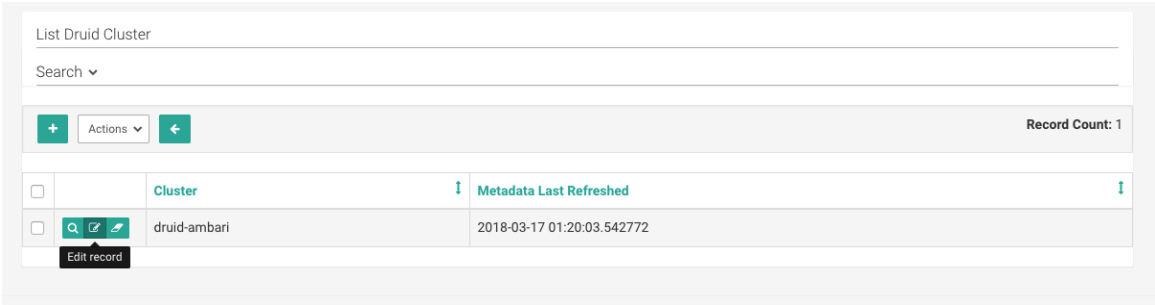


La source de données twitter-dsn créée par l'application SAM devrait maintenant apparaître. L'initialisation de la nouvelle source de données sous Druid prend environ 30 minutes.

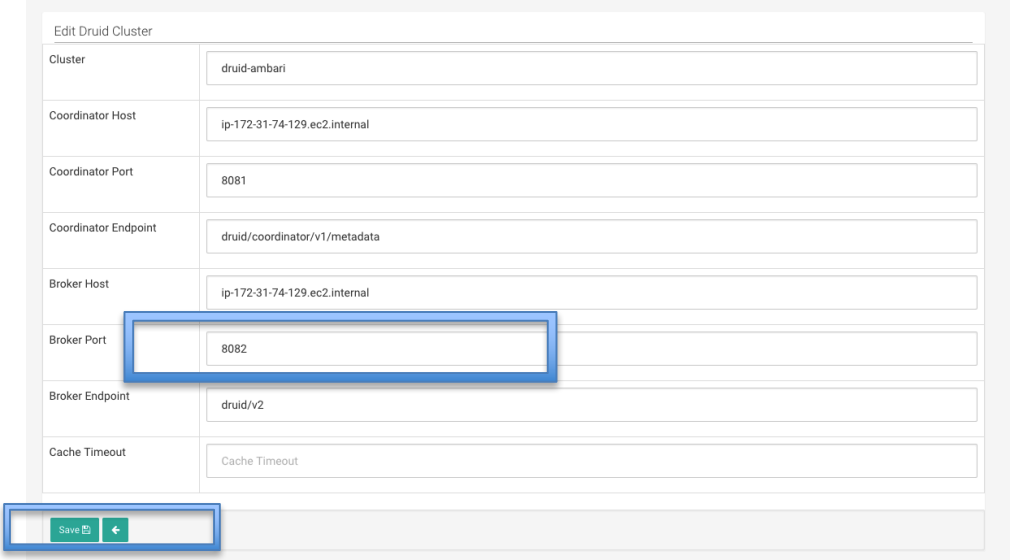
Si vous ne pouvez pas utiliser votre source de données ou obtenez une erreur lorsque vous cliquez sur le lien hypertexte sur Twitter-DSN; allez dans le menu source en haut, puis cliquez sur Druid Clusters.



Then click on “edit record”



Mettez à jour le port du courtier sur 8082, puis cliquez sur Enregistrer.



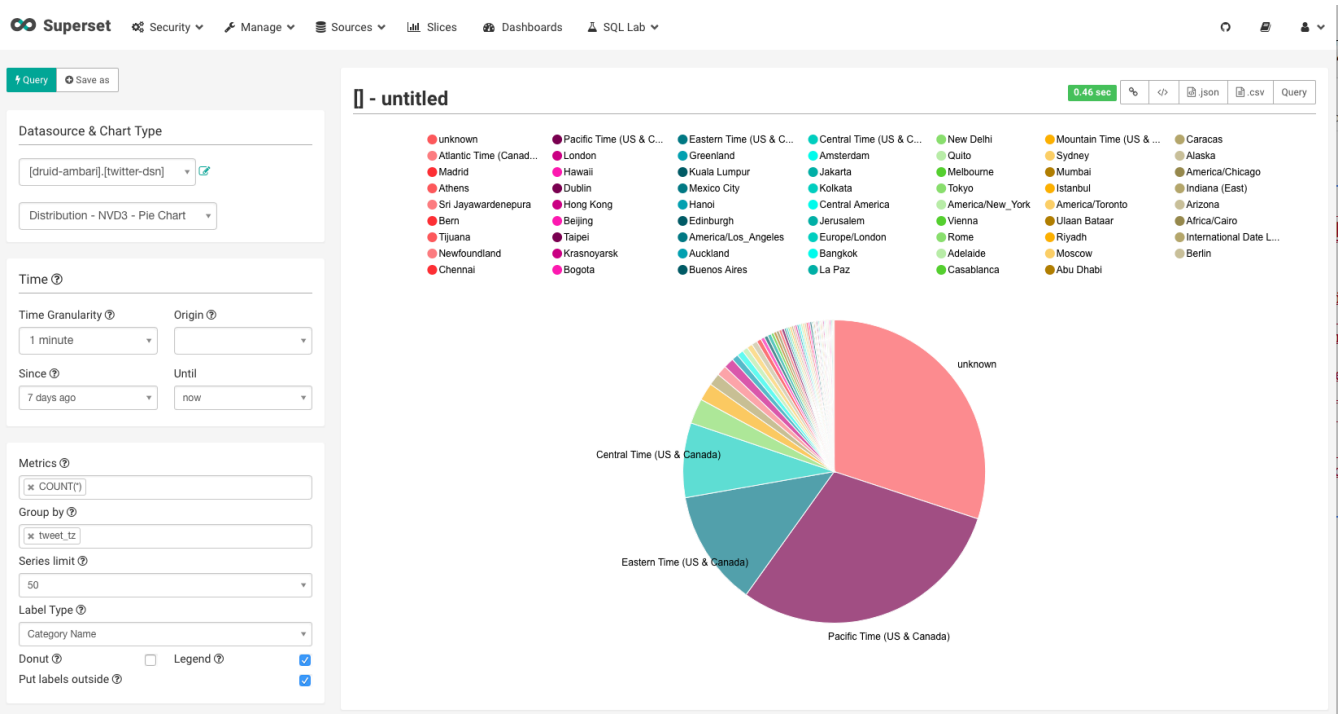
Si vous ne pouvez toujours pas voir votre source de données, vous voudrez peut-être revenir à l'étape précédente pour surveiller l'application Storm et vous assurer que les n-uplets sont écrits sur Druid. Une fois confirmé, vous devez attendre plus de temps avant d'actualiser à nouveau les métadonnées du Druid.

Create a visualization and a dashboard

Nous pouvons maintenant générer au moins une visualisation pour vérifier la région du fuseau horaire la plus active en termes de nombre de tweet. Cliquez sur le lien hypertexte Twitter-DSN. Une nouvelle fenêtre s'ouvrira (voir ci-dessous). Entrez les champs suivants:

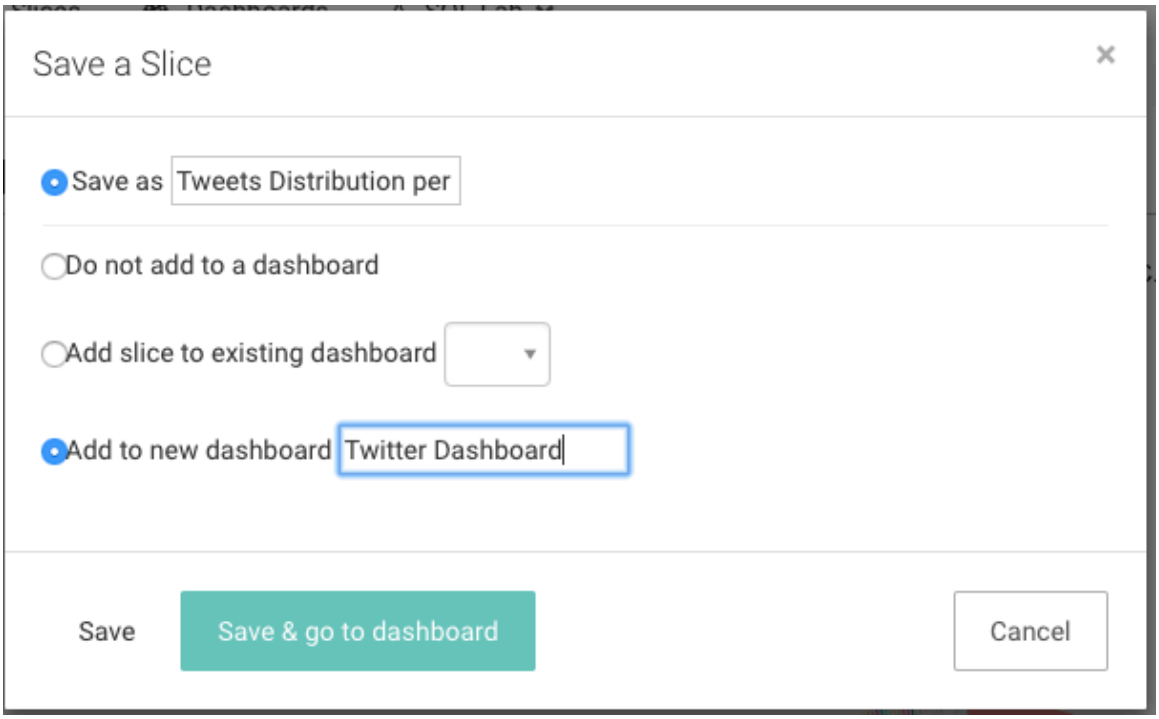
- Chart Type: **Distribution – NVD3 – Pie Chart**
- Time Granularity: **1 minute**
- Since: **7 days ago**
- Metrics: **COUNT(*)**
- Group by: **tweet_tz**

Une fois terminé, cliquez sur Requête (bouton vert) en haut à gauche de l'écran.
Vous devriez voir la visualisation apparaître.

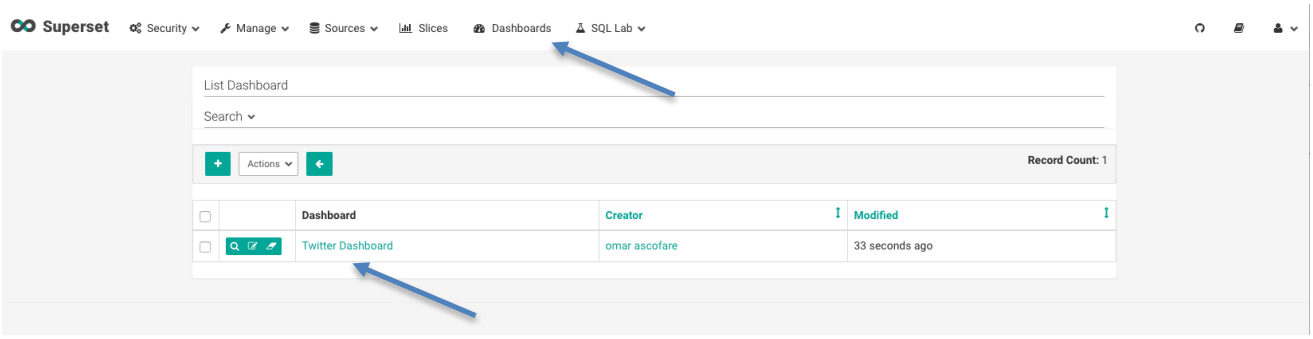


Cliquez sur Enregistrer sous:

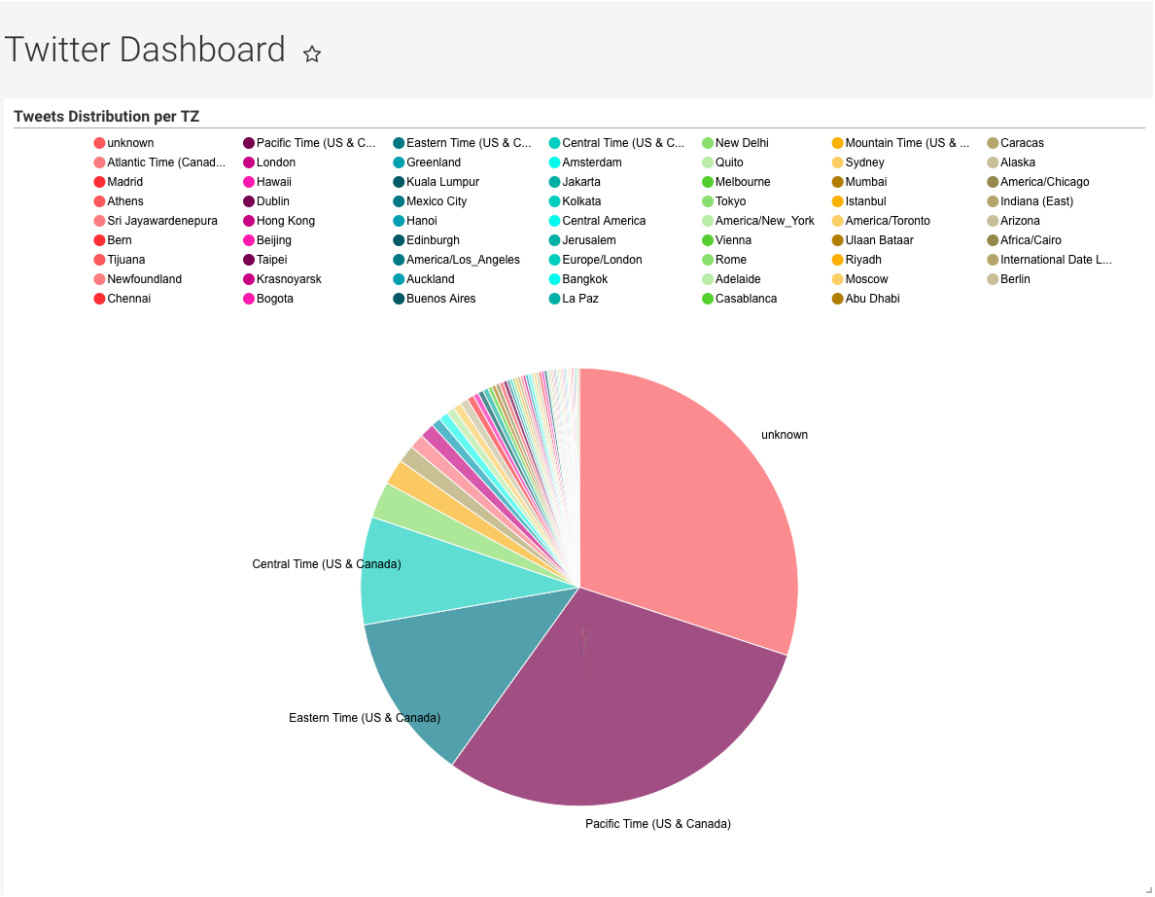
- Save as: **Tweets Distribution per TZ**
- Add to new dashboard: **Twitter Dashboard**



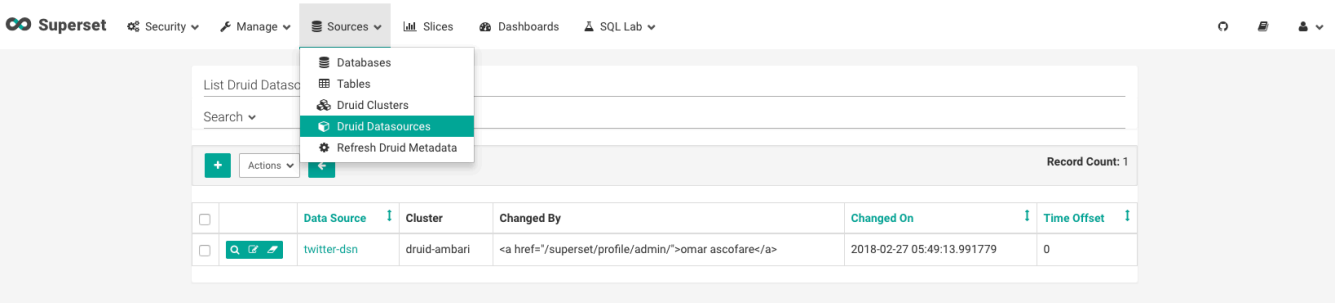
Cliquez sur Enregistrer. Ensuite, cliquez sur Tableaux de bord dans le menu du haut, puis sur l'hyperlien Twitter des tableaux de bord:



Vous devriez maintenant voir votre tableau de bord avec une seule visualisation:



Nous pouvons ajouter une autre visualisation au tableau de bord. Sélectionnez Sources dans le menu du haut et cliquez sur le lien hypertexte Twitter-DSN:

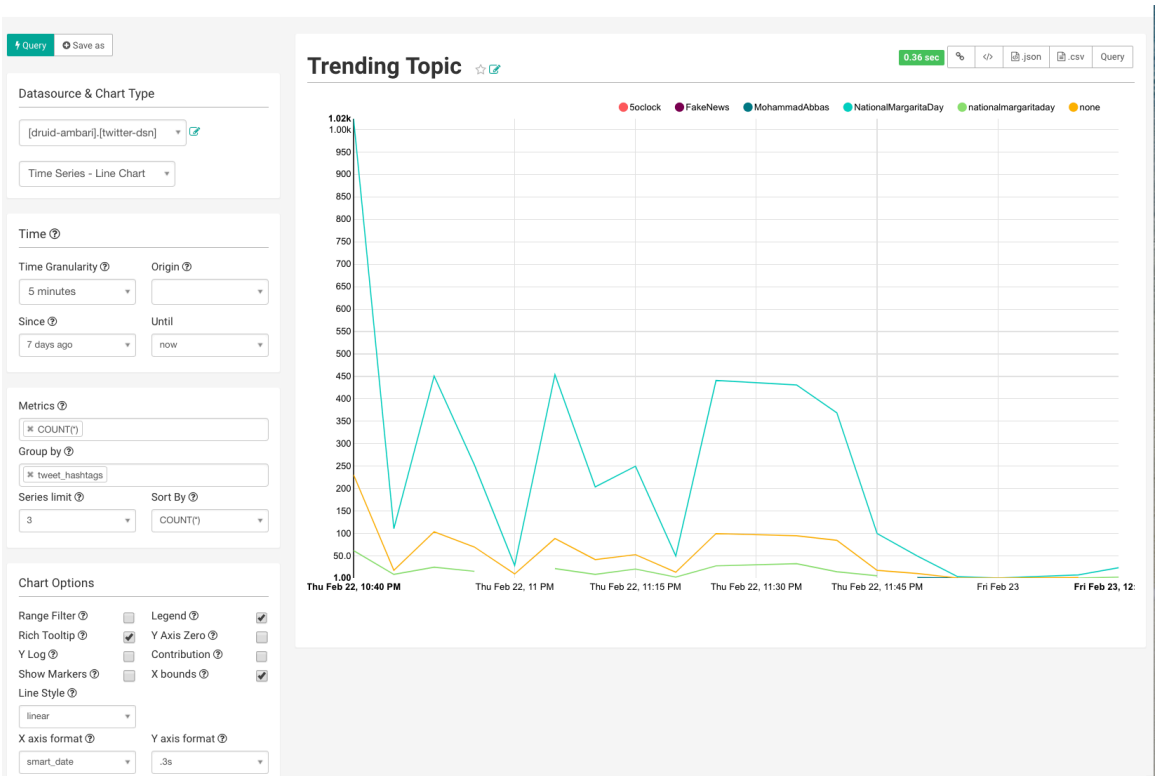


Entrez les champs suivants:

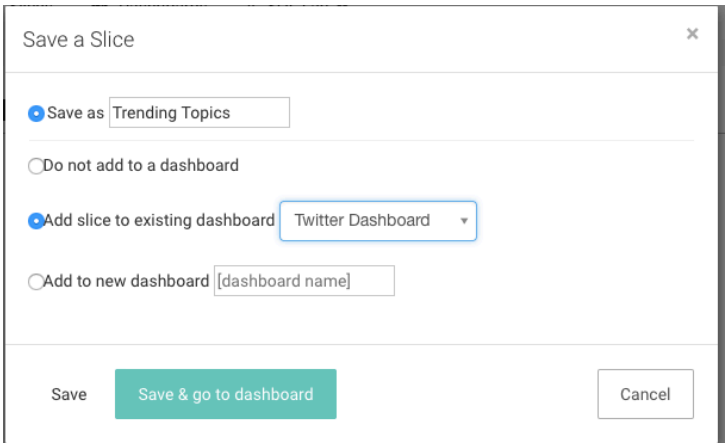
- Chart Type: **Time-Series – Line Chart**
- Time Granularity: **5 minutes**
- Since: **7 days ago**
- Metrics: **COUNT(*)**
- Group by: **tweet_hashtags**
- Series limit: **3**

- Sort By: **COUNT(*)**

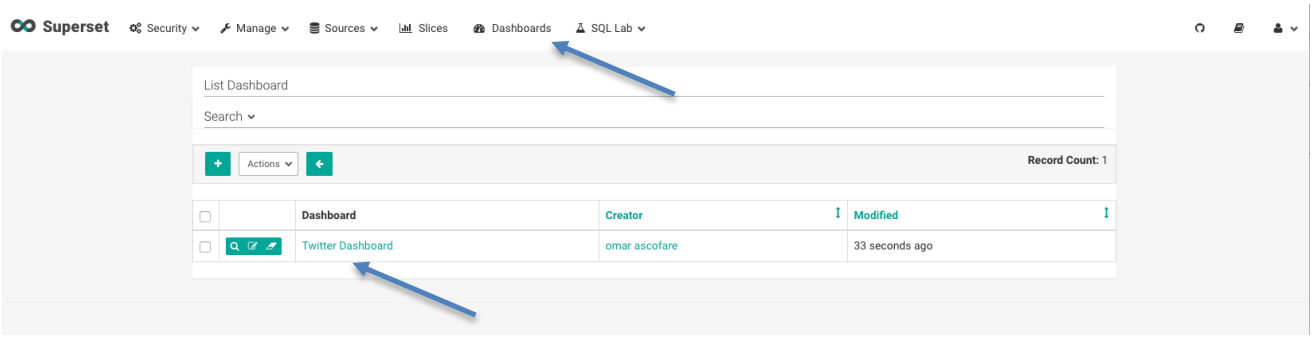
Une fois terminé, cliquez sur Requête (bouton vert) en haut à gauche de l'écran. Vous devriez voir la visualisation apparaître.



Vous pouvez maintenant enregistrer cette diapositive et l'ajouter à Twitter Dashboard. Cliquez sur le bouton "Enregistrer sous":



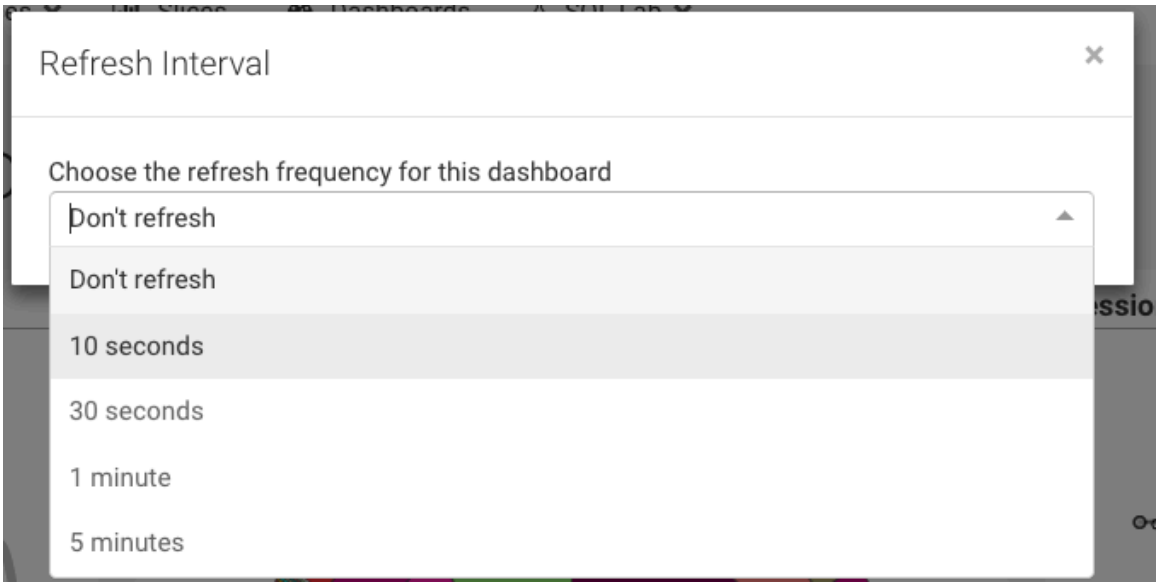
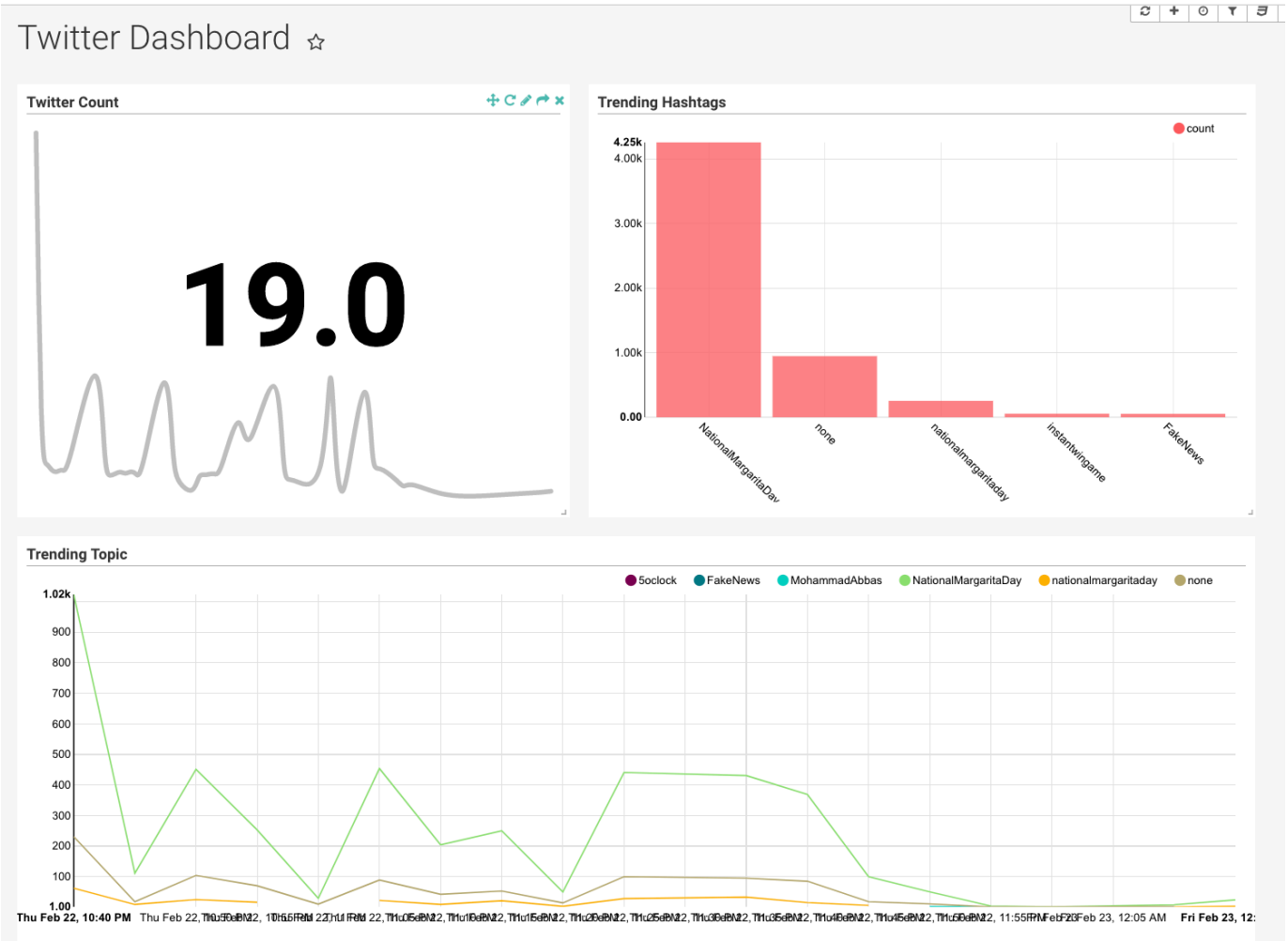
Cliquez sur Enregistrer. Ensuite, cliquez sur Tableaux de bord dans le menu du haut, puis sur l'hyperlien Twitter des tableaux de bord:



Quelques minutes peuvent être nécessaires pour exécuter les requêtes initiales requises pour votre visualisation et voir la nouvelle visualisation ajoutée sur votre tableau de bord.

Notez que ces requêtes peuvent être actualisées automatiquement à un intervalle défini pour obtenir un tableau de bord dynamique. Cet intervalle peut être défini si vous cliquez sur l'icône d'horloge en haut à droite du tableau de bord:

Vous pouvez ensuite choisir éventuellement l'intervalle d'actualisation:



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: