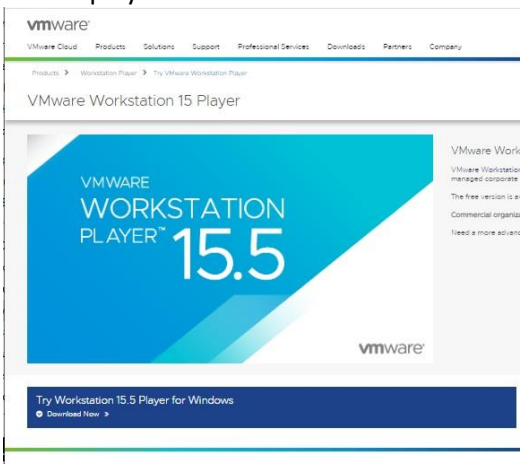
Acquiring Twitter Data
Using NiFi & Kafka

Hortonworks HDF Sandbox

Introduction: This article is an updated version of the tutorial here. It includes fixes to few errors in the original implementation. This tutorial works on Windows 10 with vMWare workstation player 15. It was not tested on Mac or with Oracle VirtualBox VM.

Pre-requisitea:

1. vMplayer to install the Hortonworks sandnox, Download and install vMware workstation player



2. Twitter API developer account.

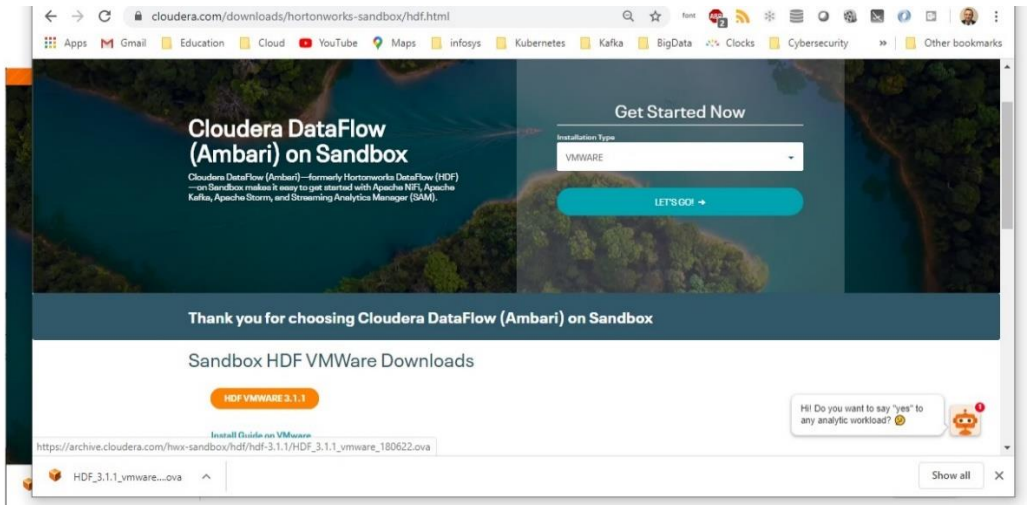
Follow instructions here in order to create a Twitter developer account.

Save the 4 keys/token that are provided by Twitter: **consumer API Key, consumer API Secret key, access Token and access Token Secret**

==========

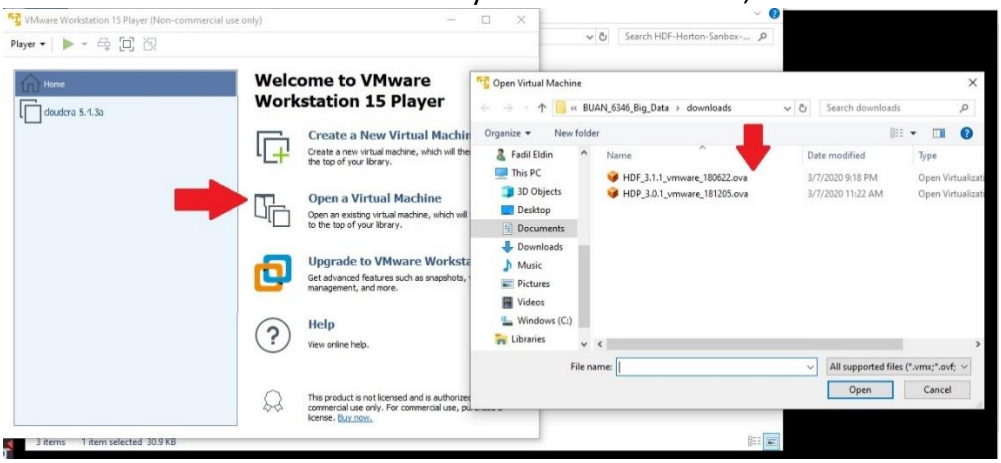
Install and run the HDF Hortonworks sandbox:

Download the image ova file of the Horton Works HDF from this page, select vMware, accept the terms and download the file.

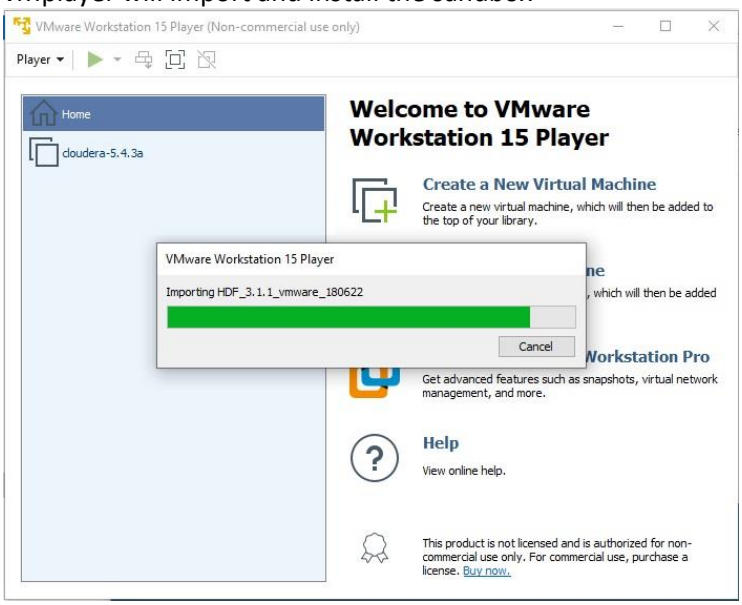


At the time of the creation of this tutorial, the name of the downloaded files was **HDF_3.1.1_vmware_180622.ova**

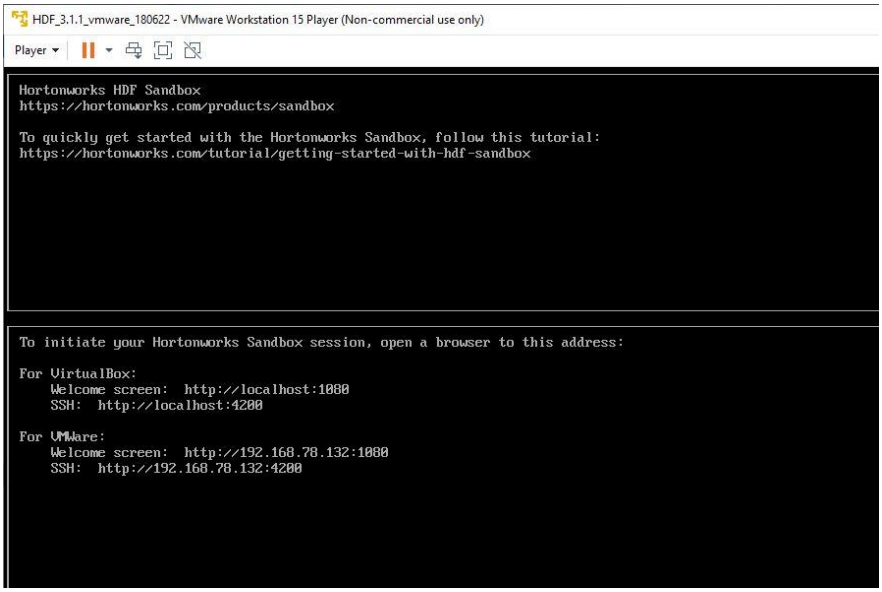
Open (vMware workstation 15 Player) on the right pane, click "**Open A Virtual Machine**" take the Windows file browser to where you save the .ova file, select the file and click **open**



vMplayer will import and install the sandbox



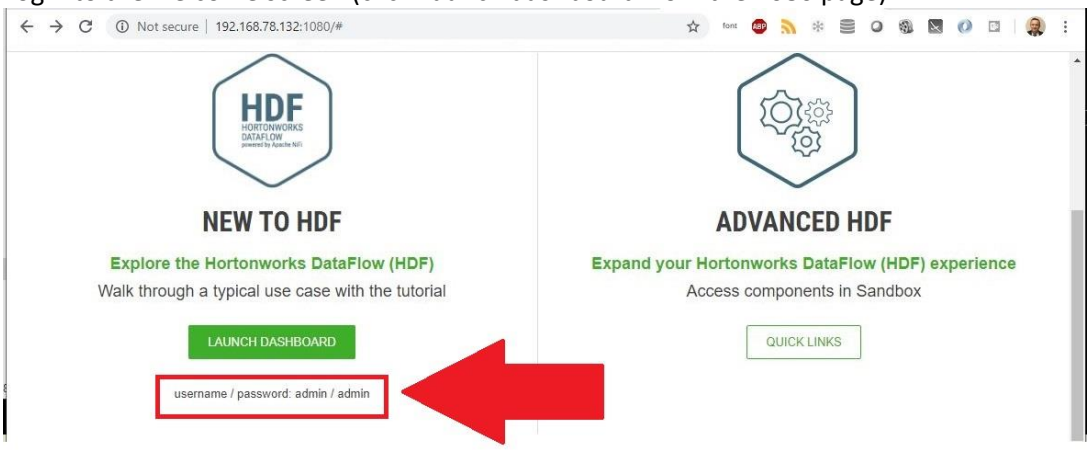
At the end of a successful installation, you will see a page like this:



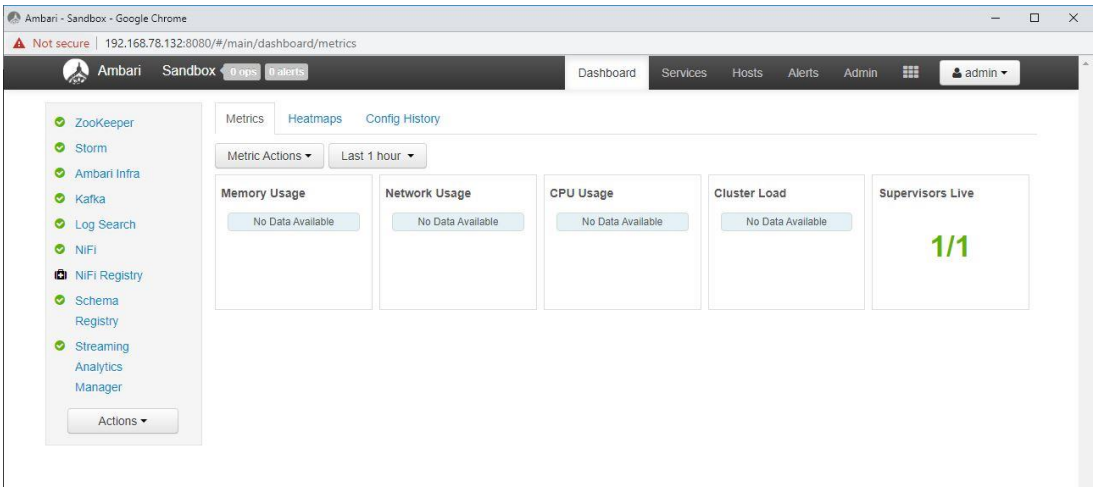
The default username/Password for the welcome screen (See vMware, port 1080) is **admin/admin**
And the default for the SSH (port 4200) is **root/Hadoop**

Preparing the SandBox:

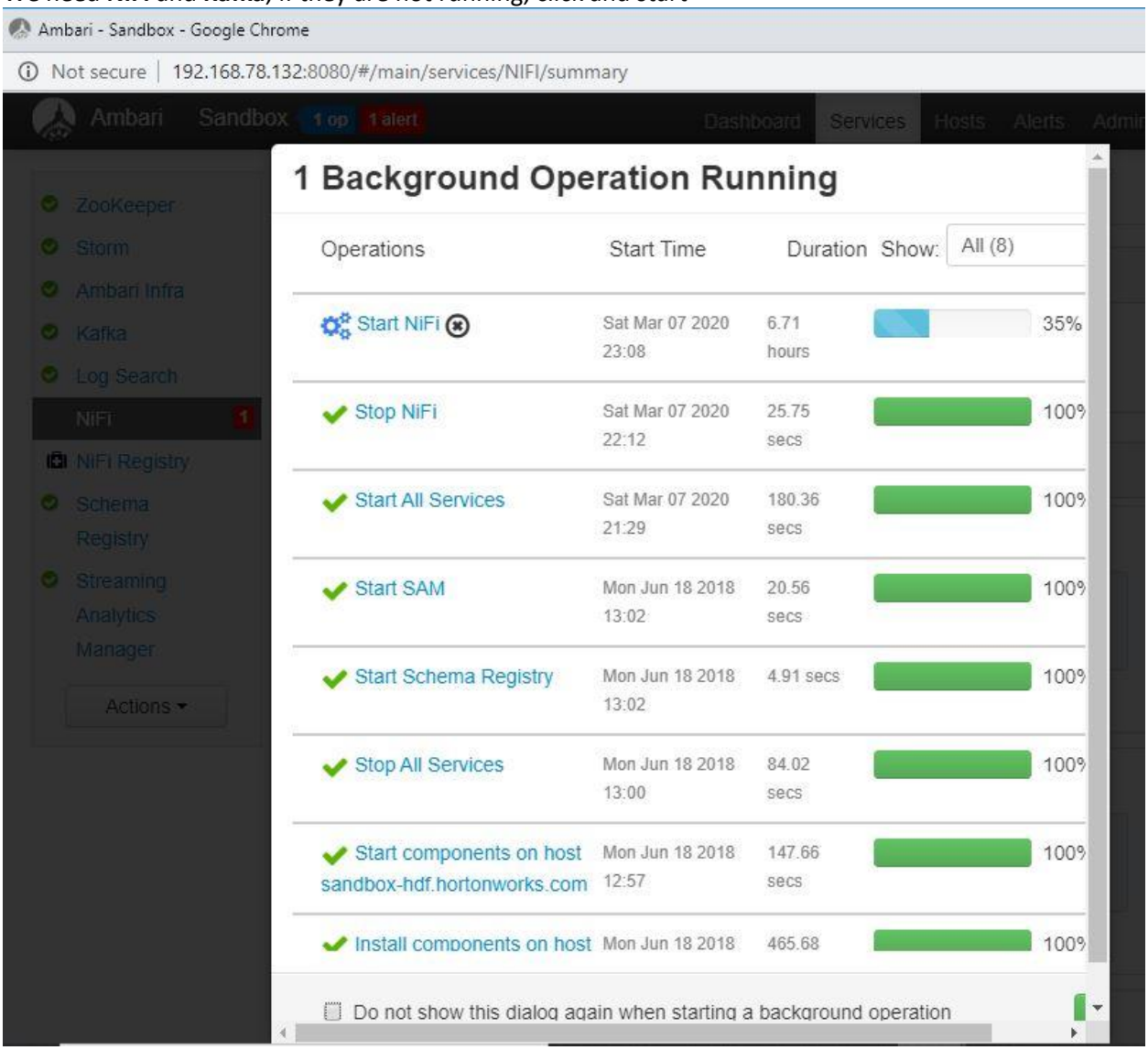
Login to the welcome screen (click Launch dashboard from the 1080 page)



and make sure all services are running



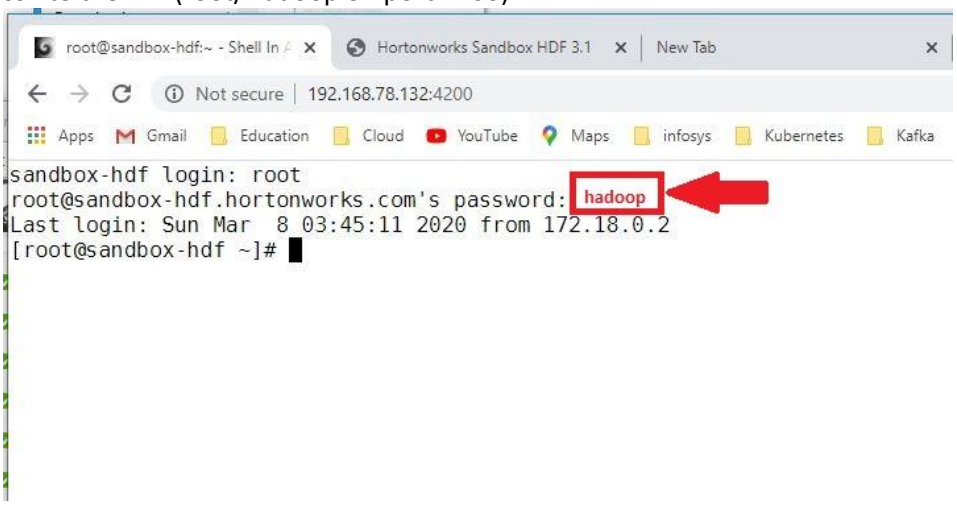
We need **NiFi** and **Kafka**, if they are not running, click and start



.

Sync the VM clock, this is needed to run Twitter, if the data on your VM is in the past then Twitter will reject the API call

ssh to the VM (root/Hadoop on port 4200)



Run the following commands

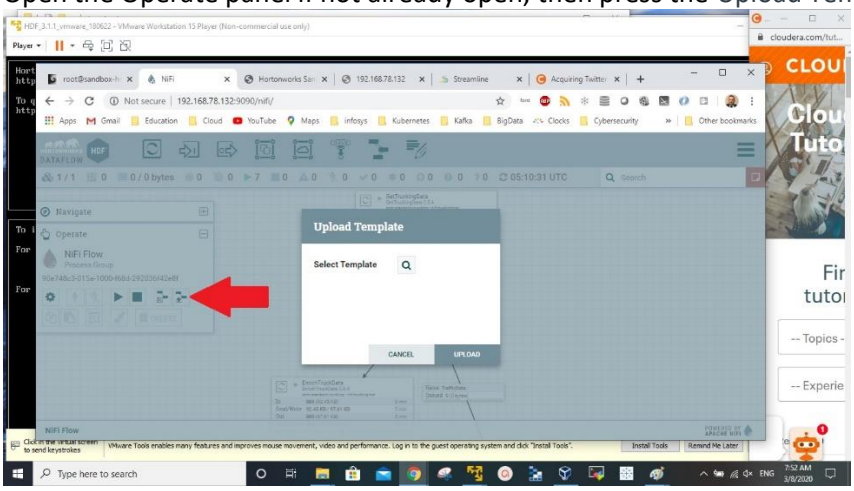
```
sudo yum install -y ntp
sudo service ntpd stop
sudo ntpdate pool.ntp.org
sudo service ntpd start
```

Import NiFi Flow For Twitter Acquisition and storage

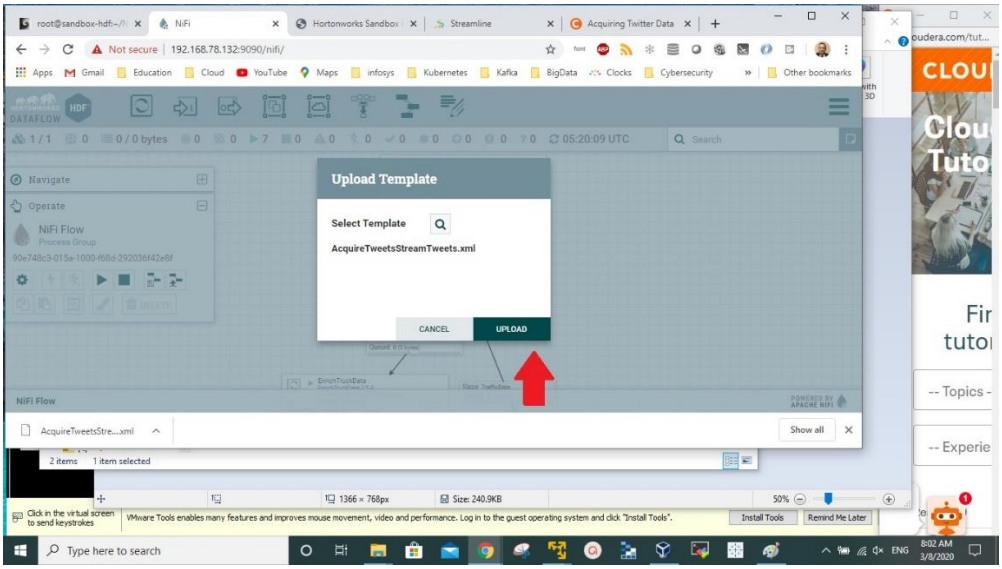
Download the NiFi xml template file from AcquireTweetsStreamTweets.xml to your local computer.

open HDF NiFi UI at http://YOUR-IP:9090/nifi.

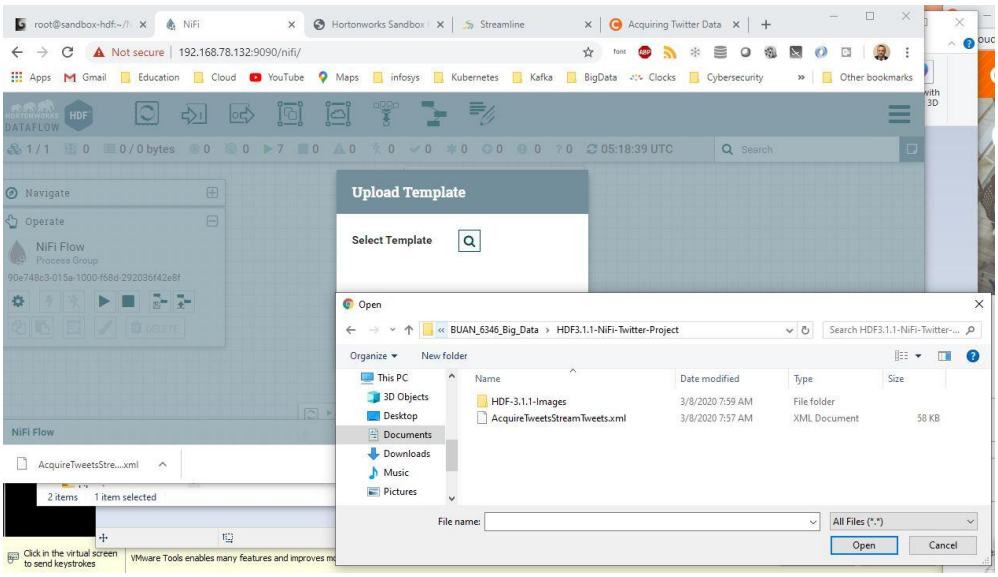
Open the Operate panel if not already open, then press the Upload Template icon  .

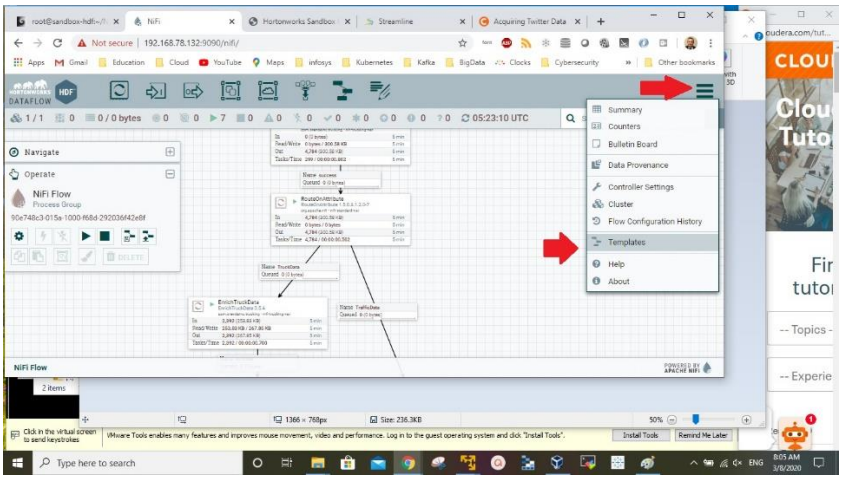


Press on Select Template icon  .



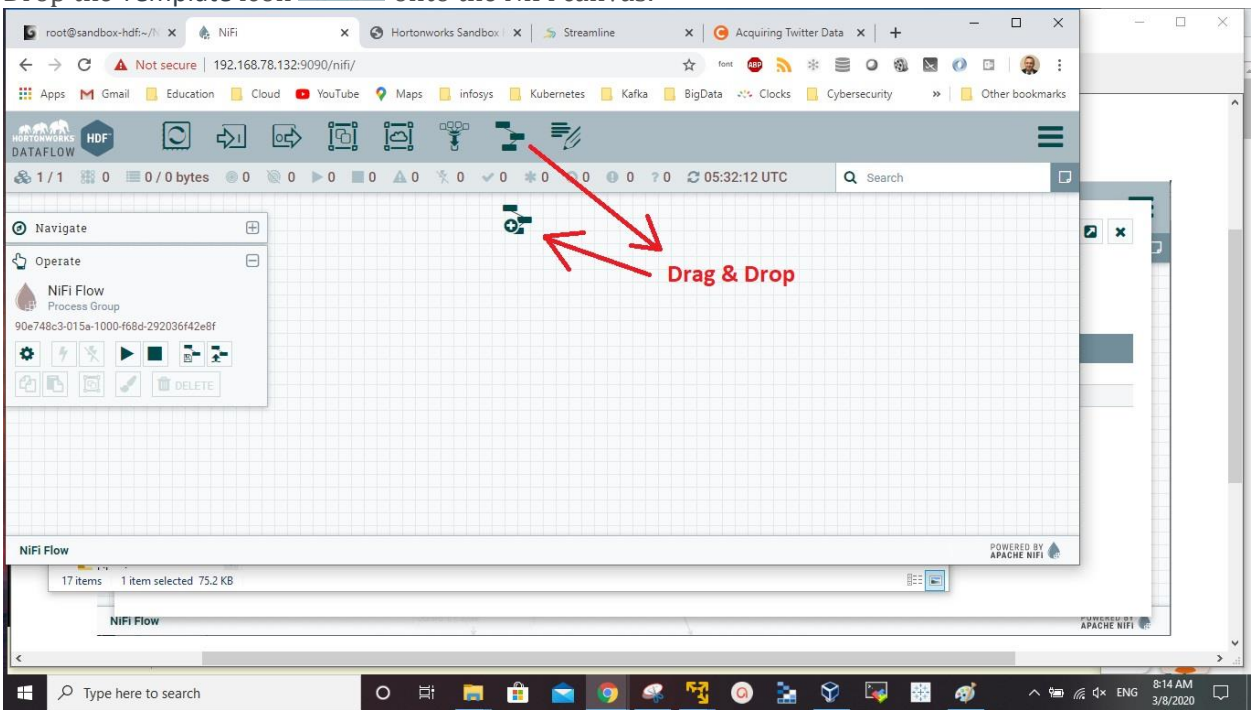
The file browser on your local computer will appear, find AcquireTweetsStreamTweets.xml template you just downloaded, then press Open, then press UPLOAD.



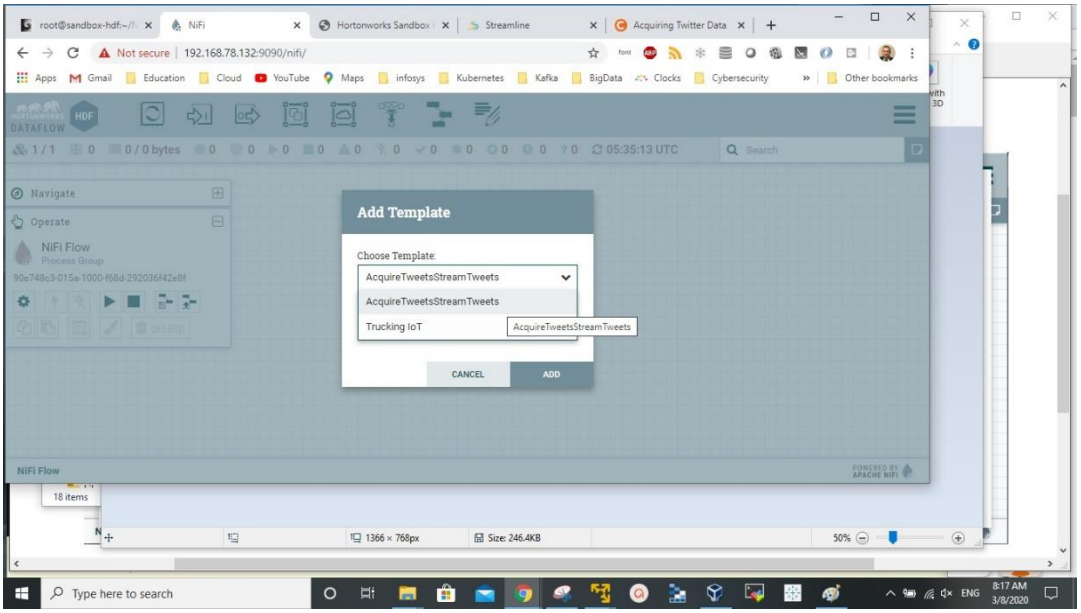


You should receive a notification that the Template successfully imported. Press OK to acknowledge.

Drop the Template icon  onto the NiFi canvas.

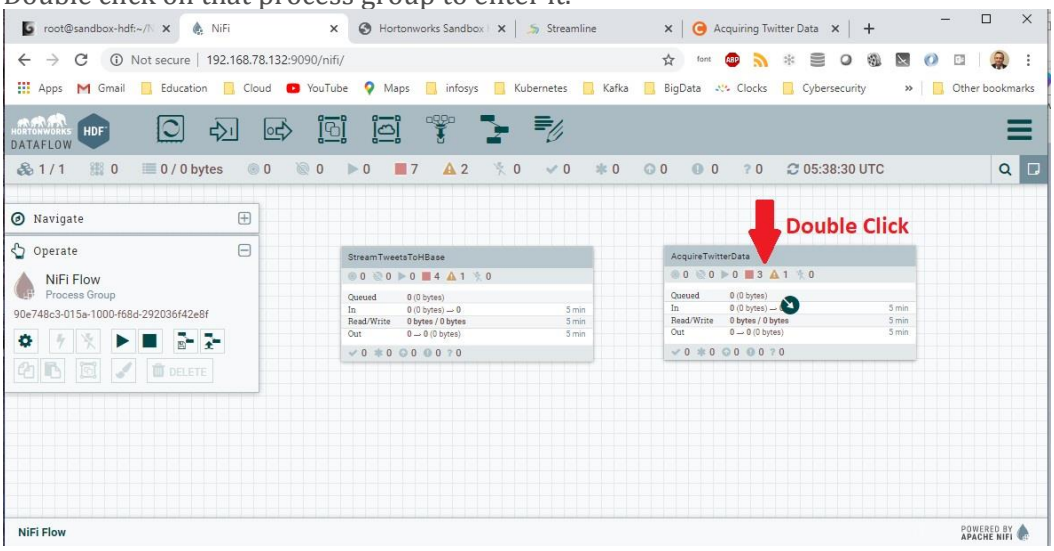


Add Template called AcquireTweetsStreamTweets.

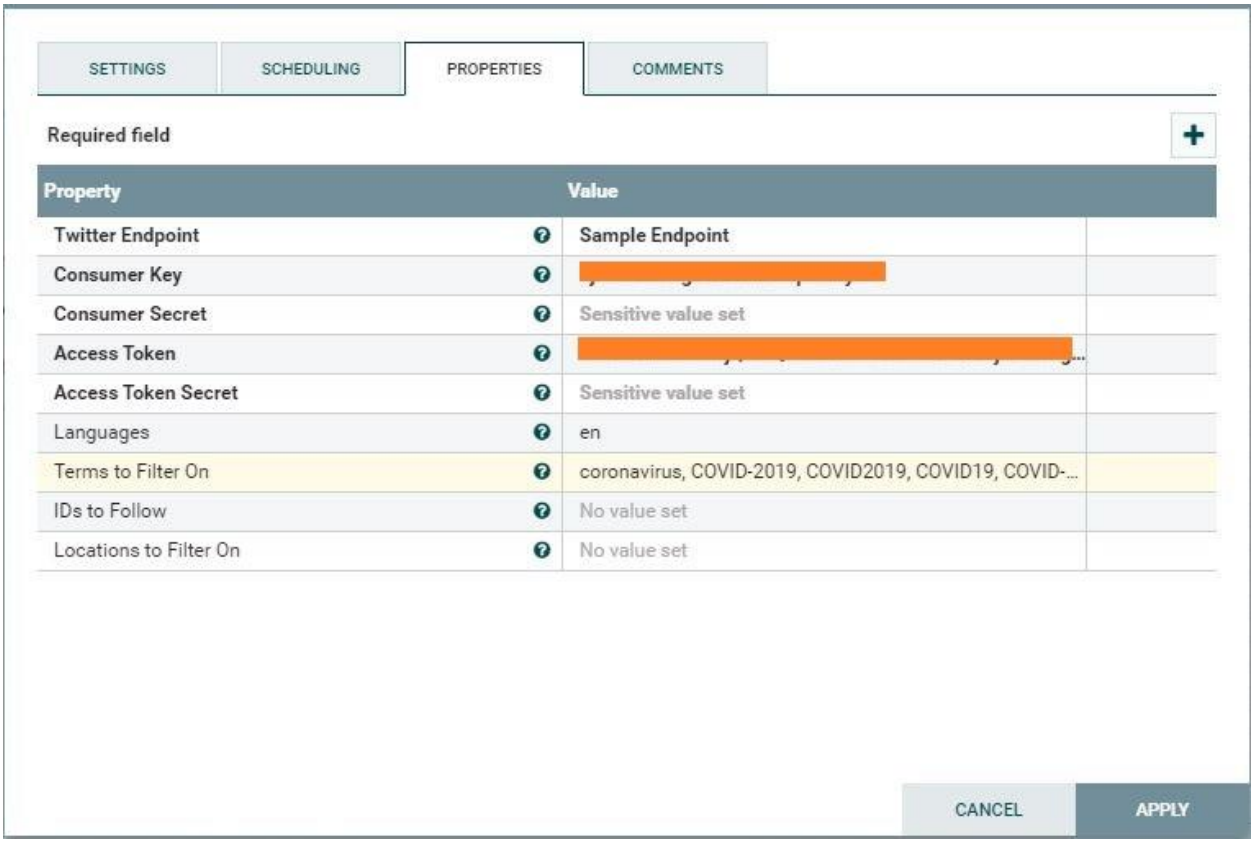


You will notice on the process group called AcquireTwitterData, there is one yellow warning.

Double click on that process group to enter it.

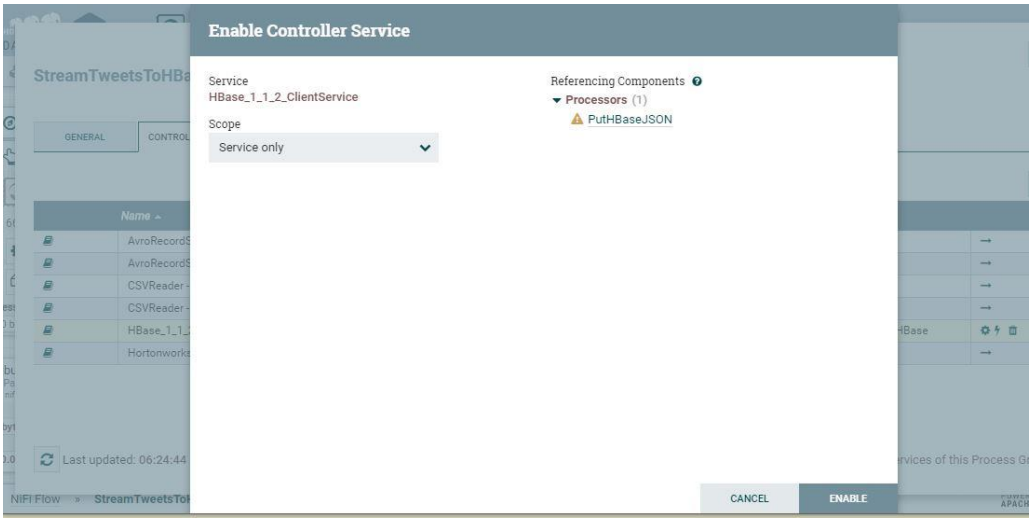


Zoom in if needed. GrabGardenHose processor has the warning. The reason is that we need to update the Consumer API Key and Consumer API Secret Key and the Access Token and Access Token Secret in the processor's properties table for the warning to go away.
Modify what terms you want to filter On.

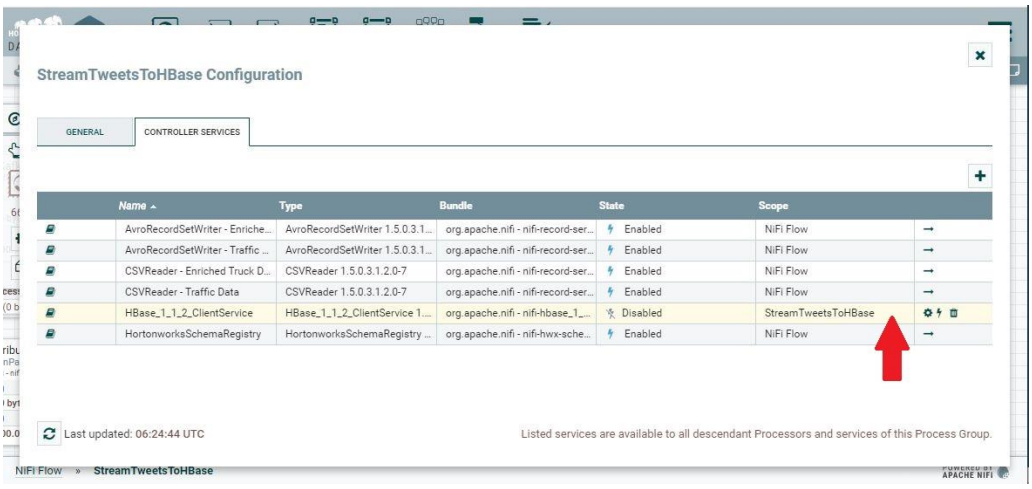


Start the NiFi flow. Hold control + mouse click on each process group, then click the start option.

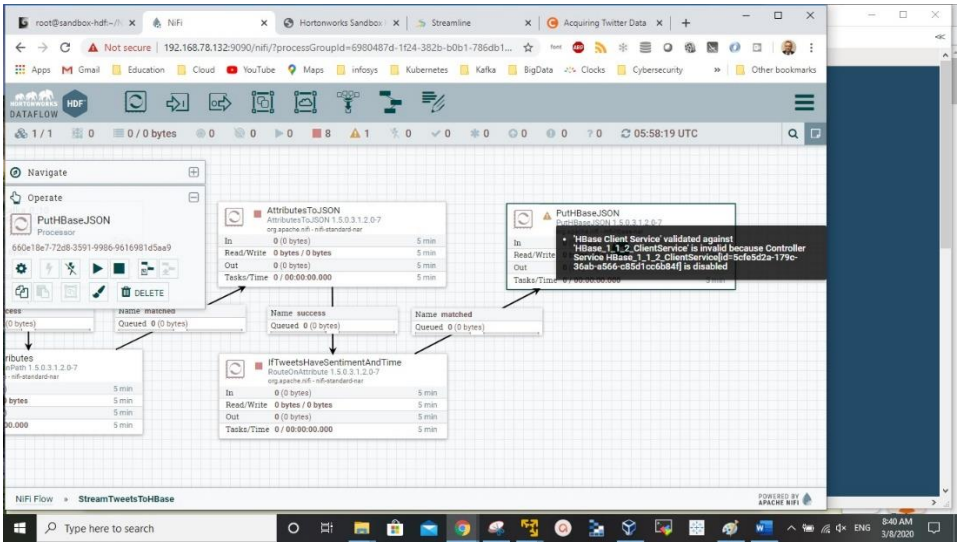
Select StreamStweets To HBAse, click configuration  and enable StreamTweetsToHBase


Enable Controller Service
Service
HBase_1_1_2_ClientService
Scope
Service only
Referencing Components
Processors (1)
PutHBaseJSON
CANCEL
ENABLE

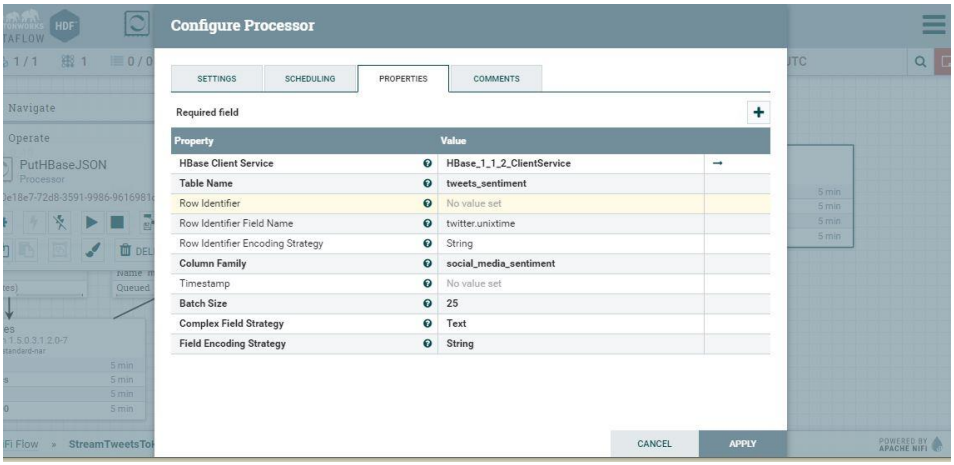
.


StreamTweetsToHBase Configuration
GENERAL
CONTROLLER SERVICES
Name
Type
Bundle
State
Scope
AvroRecordSetWriter - Enriche...
AvroRecordSetWriter 1.5.0.3.1...
org.apache.nifi - nifi-record-ser...
Enabled
NiFi Flow
AvroRecordSetWriter - Traffic ...
AvroRecordSetWriter 1.5.0.3.1...
org.apache.nifi - nifi-record-ser...
Enabled
NiFi Flow
CSVReader - Enriched Truck D...
CSVReader 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-record-ser...
Enabled
NiFi Flow
CSVReader - Traffic Data
CSVReader 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-record-ser...
Enabled
NiFi Flow
HBase_1_1_2_ClientService
HBase_1_1_2_ClientService 1....
org.apache.nifi - nifi-hbase_1_...
Disabled
StreamTweetsToHBase
HortonworksSchemaRegistry
HortonworksSchemaRegistry ...
org.apache.nifi - nifi-hwx-sche...
Enabled
NiFi Flow
Last updated: 06:24:44 UTC
Listed services are available to all descendant Processors and services of this Process Group.
NiFi Flow » StreamTweetsToHBase

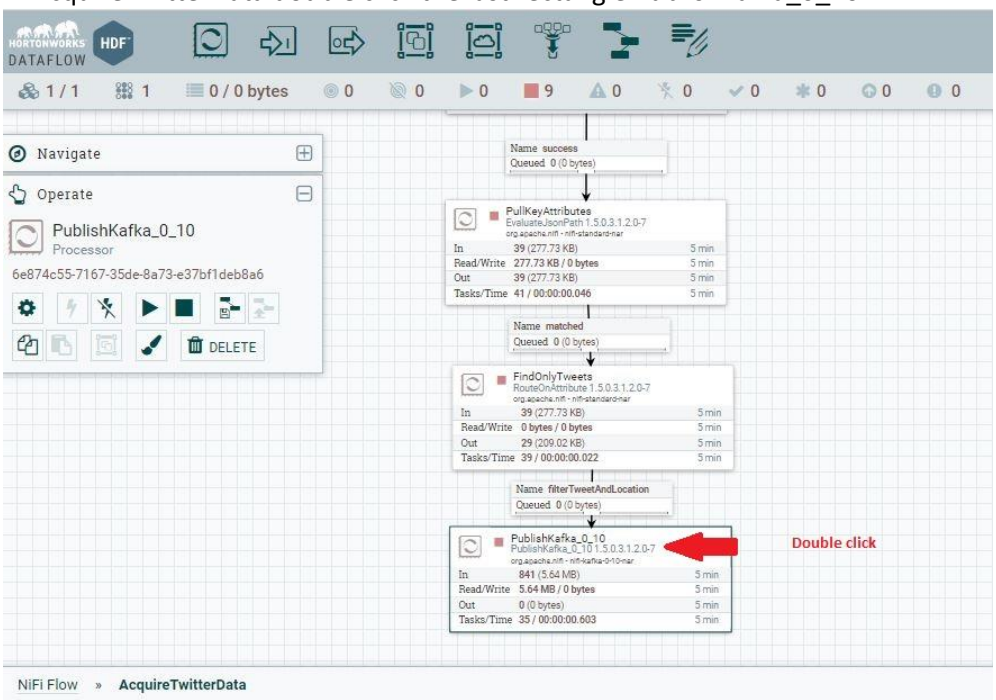
.


Navigate
Operate
PutHBaseJSON
Processor
660e18e7-72d8-3591-9986-9616981d5aa9
DELETE
AttributesToJSON
AttributesToJSON 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar
PutHBaseJSON
'HBase Client Service' validated against 'HBase_1_1_2_ClientService' is invalid because Controller Service HBase_1_1_2_ClientService[id=5cfe5d2a-179c-36ab-a566-c85d1cc6b84f] is disabled
Name matched
Queued 0 (0 bytes)
Name success
Queued 0 (0 bytes)
IfTweetsHaveSentimentAndTime
RouteOnAttribute 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar
NiFi Flow » StreamTweetsToHBase
Type here to search

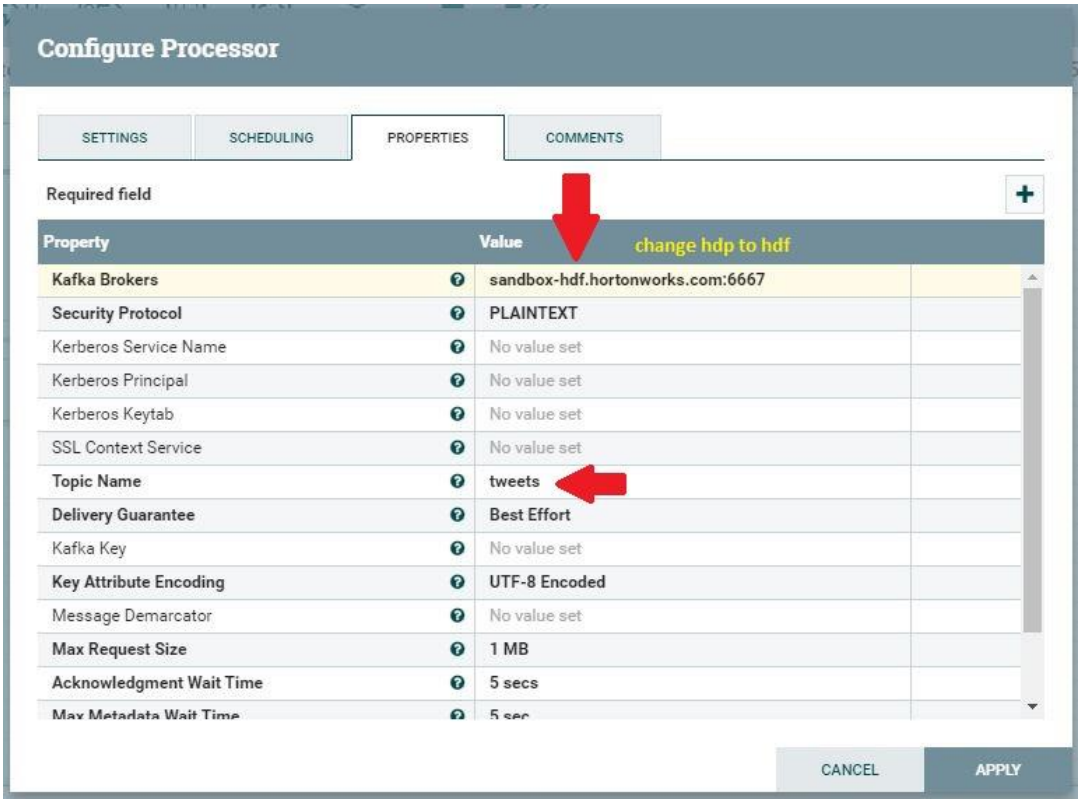
.



In AcquireTwitterData double click the last rectangle PublishKafka_0_10



.



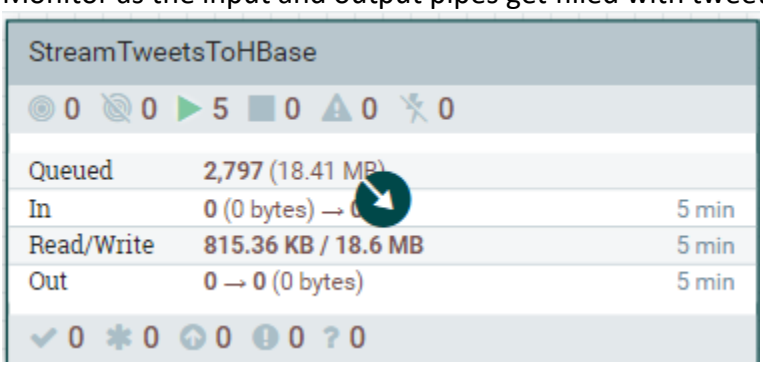
In the properties, change hdp to hdf change the topic name to tweets.

Once NiFi writes tweet data to Kafka on HDP, you can check the provenance events quickly by looking at the PublishKafka_0_10 processor inside the AcquireTwitterData process group.

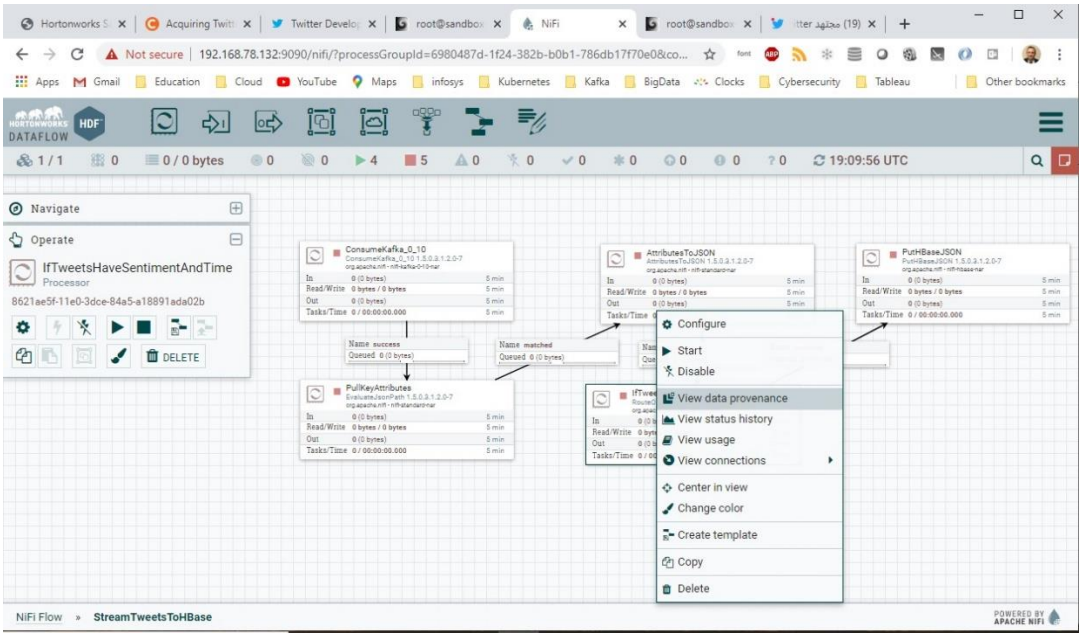
To turn off a process group, you can do so by holding control + mouse click on for instance the AcquireTwitterData process group, then choose stop option.

Run the flow and inspect the results.

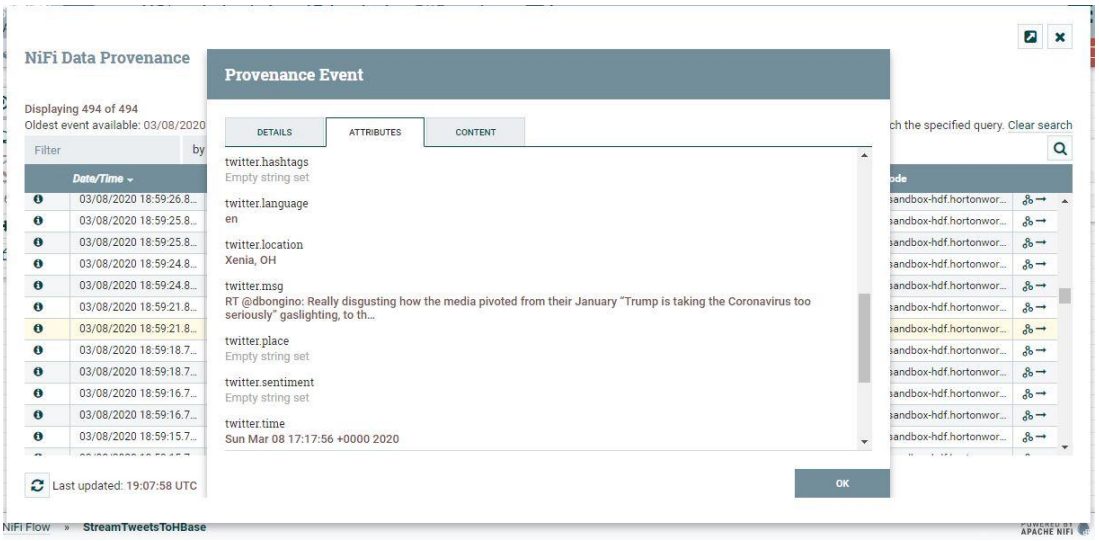
On NiFi flow main window, click CTRIL-A and Run the flow.
Monitor as the input and output pipes get filled with tweets,



They are filtered according to the keywords we put and analyzed.
Here are few sample tweet



.



This Tutorial on [LinkedIn]
Thanks