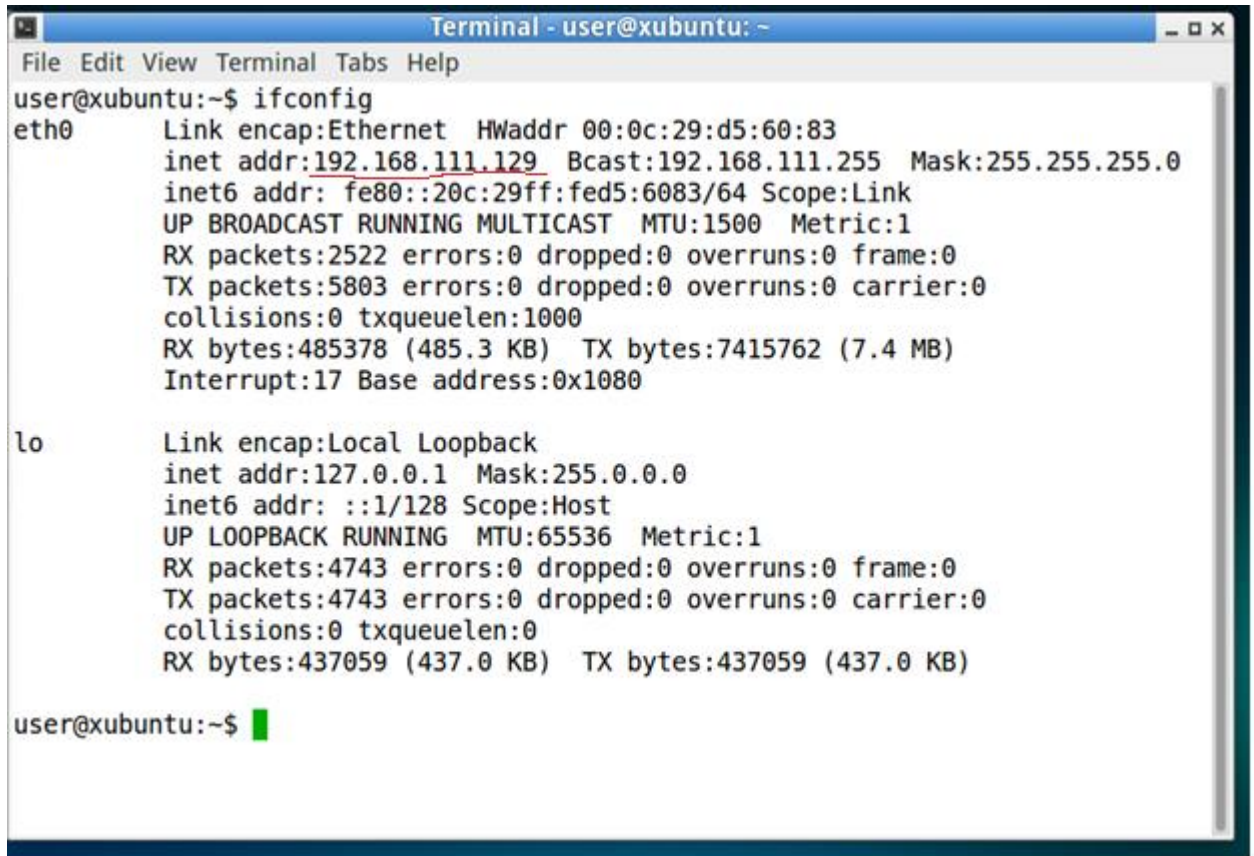


How to set up Streamsets with Kafka server to get live data from twitter.

1. Download the VM ware image file. Kafka and streamset is already installed on this image.
2. Open it with VM ware.
3. Open terminal window and check the IP address with "ifconfig" command.

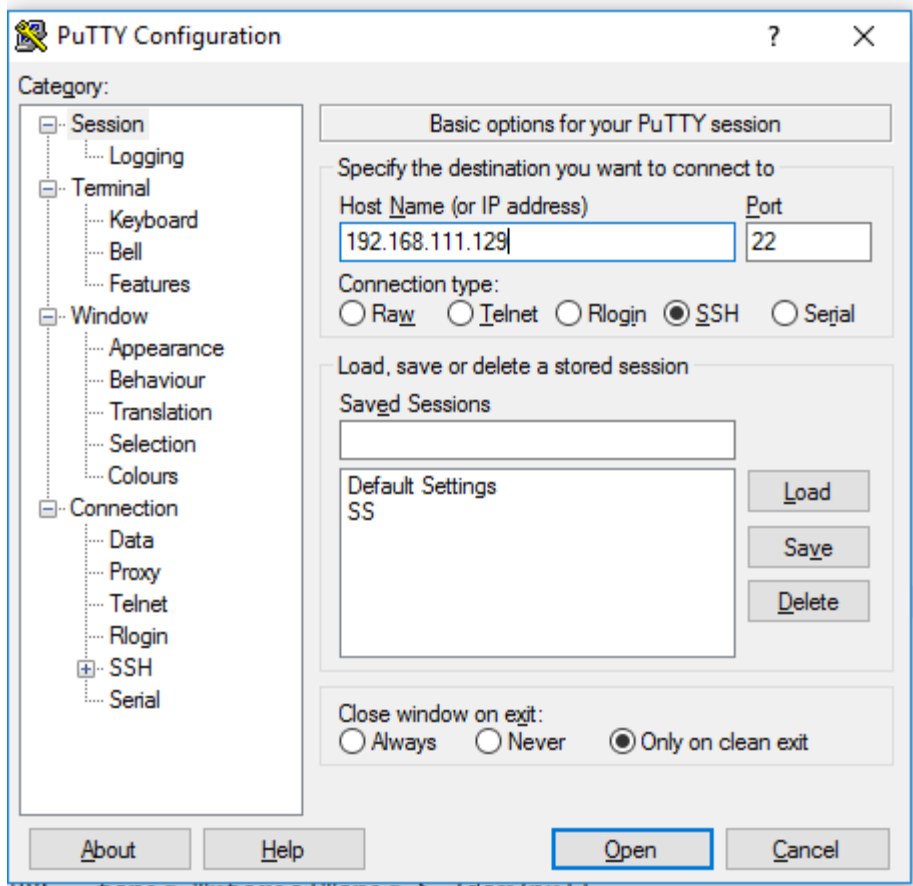
A screenshot of a terminal window titled "Terminal - user@xubuntu: ~". The window has a menu bar with "File", "Edit", "View", "Terminal", "Tabs", and "Help". The terminal shows the command "ifconfig" being executed. The output for the "eth0" interface is displayed, showing the IP address "192.168.111.129" underlined. The output for the "lo" interface is also shown. The prompt "user@xubuntu:~\$" is visible at the bottom.

```
Terminal - user@xubuntu: ~
File Edit View Terminal Tabs Help
user@xubuntu:~$ ifconfig
eth0      Link encap:Ethernet  HWaddr 00:0c:29:d5:60:83
          inet addr:192.168.111.129  Bcast:192.168.111.255  Mask:255.255.255.0
          inet6 addr: fe80::20c:29ff:fed5:6083/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:2522 errors:0 dropped:0 overruns:0 frame:0
          TX packets:5803 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:485378 (485.3 KB)  TX bytes:7415762 (7.4 MB)
          Interrupt:17 Base address:0x1080

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:4743 errors:0 dropped:0 overruns:0 frame:0
          TX packets:4743 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:437059 (437.0 KB)  TX bytes:437059 (437.0 KB)

user@xubuntu:~$
```

4. Open Putty and use above IP address to connect to terminal.



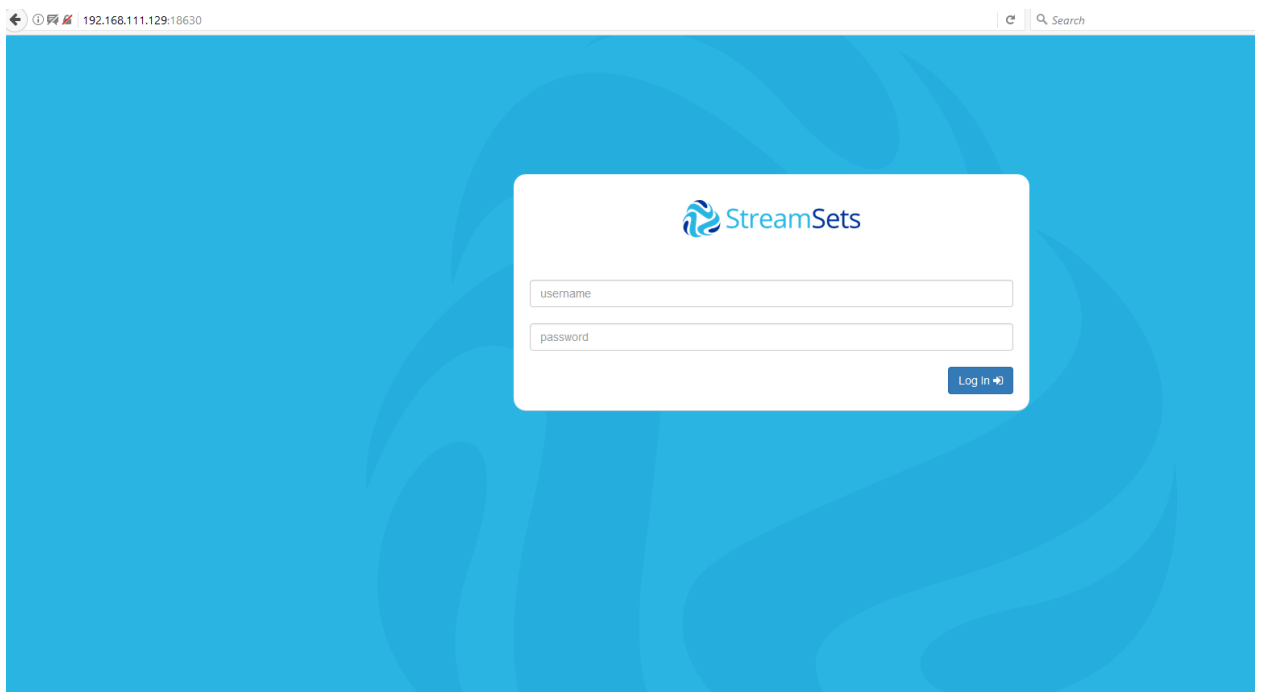
5. Use below details to connect with terminal.
 - a. Username 1 : User
 - b. Password : password
6. Login with SU and super user password is “password”, we need to login as super user as we have to run below command after login as super user.
 - a. `$ ulimit -n 9000`
7. Start Streamset application by below code.
 - a. `$ streamsets-datacollector-2.1.0.2/bin/streamsets dc`
8. You will get below screen; use the defined URL to open streamset in the browser (My case: <http://192.168.111.129:18630/>).

```
root@xubuntu: /home/user
login as: user
user@192.168.111.129's password:
Access denied
user@192.168.111.129's password:
Welcome to Ubuntu 14.04 LTS (GNU/Linux 3.13.0-24-generic i686)

 * Documentation:  https://help.ubuntu.com/

Last login: Mon Jun 26 16:17:59 2017 from 192.168.111.1
user@xubuntu:~$ su
Password:
root@xubuntu:/home/user# ulimit -n 9000
root@xubuntu:/home/user# streamsets-datacollector-2.1.0.2/bin/streamsets dc
Java HotSpot(TM) Server VM warning: ignoring option PermSize=256m; support was removed in 8.0
Java HotSpot(TM) Server VM warning: ignoring option MaxPermSize=512m; support was removed in 8.0
Running on URI : 'http://xubuntu:18630'
```

9. Login into stream set with user name (“admin”) and password (“admin”).



10. Login with KAFKA user in putty terminal and run KAFKA server, below are the commands to start it.

- a. \$ su - kafka
- b. \$ cd ~/kafka
- c. **Start KAFKA server:** \$ nohup ~/kafka/bin/kafka-server-start.sh
~/kafka/config/server.properties > ~/kafka/kafka.log 2>&1 &
- d. **Stop KAFKA server :** \$ nohup ~/kafka/bin/kafka-server-stop.sh &

11. All set to use Streamset and Kafka.

12. Below are the steps to create Twitter producer pipe line.

- a. Create a twitter app in twitter to get consumer key and token. Use this link <https://apps.twitter.com/>.
- b. Add HTTP client as source and configure it for twitter app.
- c. Add twitter URL as resource URL, check below image.
- d. Add twitter consumer key and token in credential tab.
- e. User other setting according to you requirement.
- f. Add field remover to clean the tweets.
- g. Add Kafka producer to create a stream in Kafka server
- h. Add Broker URI and select stage library to Kafka.

The screenshot displays the StreamSets Pipelines interface. At the top, the pipeline is named 'TweeterData_1'. The pipeline consists of three stages: 'TweeterData_1' (a blue box with a Twitter icon), 'FilterTweets' (a grey box with a funnel icon), and 'NFLKUP' (a grey box with a Kafka icon). The 'TweeterData_1' stage is selected, and its configuration is shown in the bottom panel. The configuration includes a Resource URL, Headers, Mode (Streaming), HTTP Method (GET), and Data Format (JSON). The Resource URL is set to 'https://stream.twitter.com/1.1/statuses/filter.json?track=%23SuperBowl'. The Headers section is empty. The Mode is set to 'Streaming'. The HTTP Method is set to 'GET'. The Data Format is set to 'JSON'. The bottom panel also shows tabs for 'Info', 'General', 'HTTP', 'Pagination', 'Credentials', 'Proxy', 'SSL/TLS', 'Text', 'JSON', 'Delimited', 'XML', 'Log', 'Avro', 'Binary', 'Protobuf', 'Datagram', and 'Whole File'. The 'HTTP' tab is selected.

The first screenshot shows the 'TweeterData_1' pipeline configuration in the StreamSets Data Collector. The pipeline consists of three components: 'TweeterData_1' (a blue box with a Twitter icon), 'FilterTweets' (a grey box with a funnel icon), and 'NFLKP' (a grey box with a Kafka icon). The 'Credentials' tab is selected, showing fields for 'Consumer Key', 'Consumer Secret', 'Token', and 'Token Secret'. The second screenshot shows the 'NFLKP' component configuration. The 'Kafka' tab is selected, showing fields for 'Broker URI' (192.168.111.129:9092), 'Runtime Topic Resolution' (unchecked), 'Topic' (superbow1), 'Partition Strategy' (Round Robin), 'One Message per Batch' (unchecked), and 'Data Format' (SDC Record).

13. Below are the steps to create Twitter consumer pipe line
- Add Kafka consumer and configure it according to your IP address
 - Add filed remover, field flattener and expression evaluator to clean the data again.
 - Add local file system as destination and configure it according to your needs.

StreamSets

Pipelines / Tweeter_Consumer

KFTConsumer → FilterData → DataFlattener → MAPToListMap → WriteDataInDirectory

KFTConsumer

Info Configuration Raw Preview

General Kafka Text JSON Delimited XML Log Avro Binary Protobuf Datagram Whole File

Data Format: SDC Record

Broker URI: 192.168.111.129:9092

ZooKeeper URI: 192.168.111.129:2181

Consumer Group: streamsetsDataCollector

Topic: superbowL

Produce Single Record: ☐

Max Batch Size (records): 1000

Batch Wait Time (ms): 2000

Kafka Configuration: Enter Name

Switch to bulk edit mode

StreamSets

Pipelines / Tweeter_Consumer

KFTConsumer → FilterData → DataFlattener → MAPToListMap → WriteDataInDirectory

WriteDataInDirectory

Info Configuration

General Output Files Late Records Text JSON Delimited Avro Binary Protobuf Datagram Whole File

Data Format: JSON

Files Prefix: ado-\${docId() }

Directory in Header: ☐

Directory Template: /home/user/streamsets-datacollector-2.1.0.2/SuperBowlTweets/\${YYYY()}-\${MM()}-\${DD()}-\${hh() }

Data Time Zone: UTC (UTC)

Time Basis: \${time:now() }

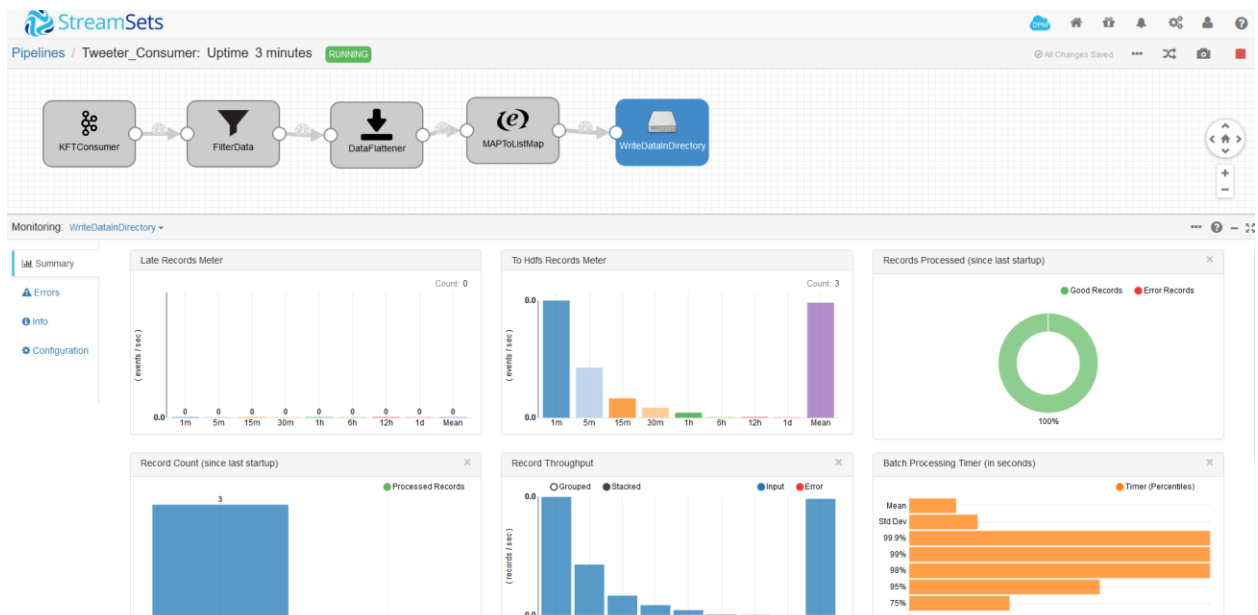
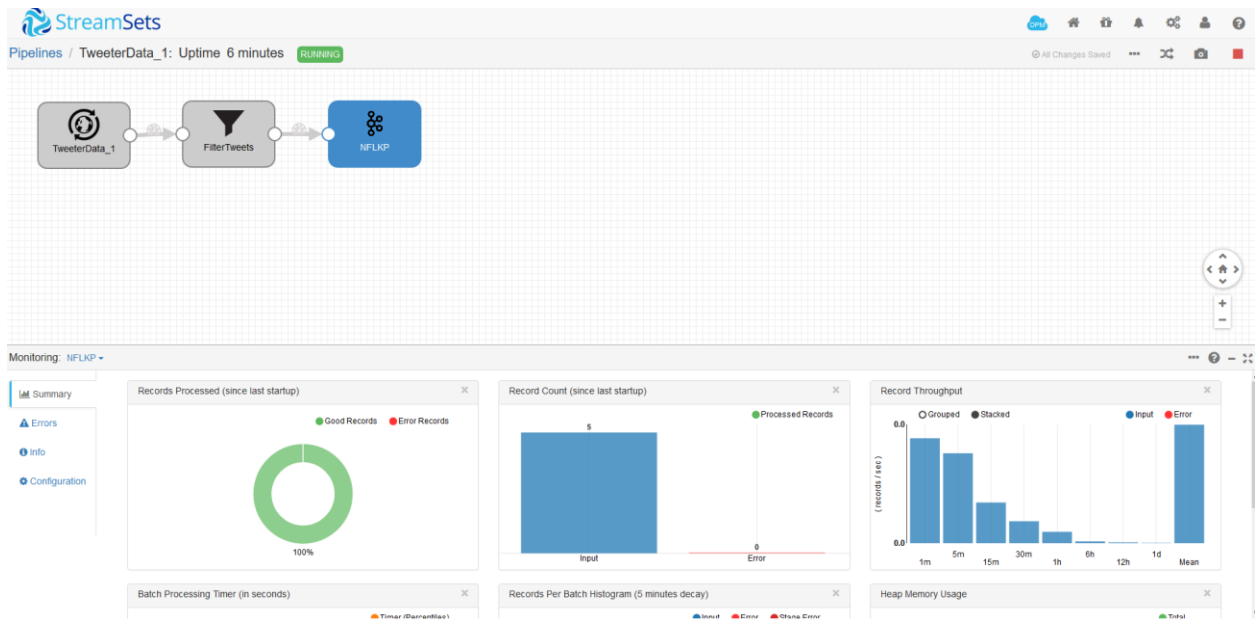
Max Records in File: 10000

Max File Size (MB): 0

Idle Timeout: \${1 * 1000000}

Compression Codec: None

14. Run both the pipeline to get the data.



Note: Everything for both pipeline already set up, you just need to use your tweeter token, consumer key, URL and IP address to start the pipelines