## Stack

Python3, Kafka, Elastic stack

## Used Python third party module

Kafka, json, elasticsearch, tweepy, configparser, textblob

## Workflow logic

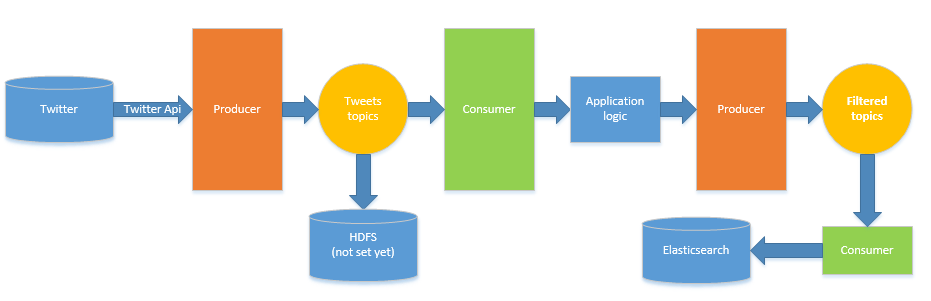


Figure 1 Data pipe workflow logic

## Code startup

1. Modify ***config.cfg*** to set up the twitter API authorization.
2. Kafka configuration can be found in Kafka\_config.rar. This is the configuration I used for Kafka.
3. Run **conusumer.py**
4. Run **producer.py**
5. Run  **Kafka\_to\_es\_bulk.py**

## Comments

* The data extraction from topics to HDFS is not set in my workflow
* There is a bug(TO\_DO) in the consumer code:

It can only extract countries from the twitter txt, but not cities. Therefore for many records, the column ‘key\_words’ is empty.

For example, if the twitter text is ‘I love Milan’. The column ‘key\_words’ will be empty. But it’s supposed to be ‘Italy’.

This bug can be fixed by:

* 1. Using multiple producers and partitions. Assign only one country name to each producer and set partition name. Therefore each partition will only contain records for one country.
  2. Improve the logic in TO\_DO part. Let it map between cities and countries.