Real Time Weather Data Sink into HDFS using Flume

Prerequisite

- Please ensure you have python installed on your ubuntu.

Follow the below steps to ingest weather data to hdfs:

1. Create a file weather.conf as follows:
    - Open terminal and type:

    gedit weather.conf

    - In weather.conf type

```
WeatherAgent.sources=pstream
WeatherAgent.channels=memoryChannel
WeatherAgent.sinks=HDFS
WeatherAgent.sources.pstream.type=exec
WeatherAgent.sources.pstream.command = python
/home/saurabh/Documents/print_weatherInfo.py
WeatherAgent.sinks.HDFS.type=hdfs
WeatherAgent.sinks.HDFS.hdfs.path= /Weather_feed
WeatherAgent.sinks.HDFS.hdfs.fileType=DataStream
WeatherAgent.sinks.HDFS.hdfs.writeFormat=Text
WeatherAgent.channels.memoryChannel.type=memory
WeatherAgent.channels.memoryChannel.capacity=1000
WeatherAgent.channels.memoryChannel.transactionCapacity=100
WeatherAgent.sinks.HDFS.channel=memoryChannel
WeatherAgent.sources.pstream.channels=memoryChannel
```

    - Move this weather.conf to /usr/local/flume/conf as:

```
sudo mv weather.conf /usr/local/flume/conf
```

2. Create a python script as follows:
    - Open termial and type:

    cd /Desktop/

gedit print_weatherInfo.py

- In print_weatherinfo.py type:

  import urllib2

  response =

  urllib2.urlopen('http://api.apixu.com/v1/current.json?

  key=50e18156d7704eac99d93832191704&q&q=jaipur')

  html = response.read()

  print(html)

- Save and exit

3. Now run the flume job using command:

   flume-ng agent -n WeatherAgent -c conf -f /usr/local/flume/conf/weather.conf

Note:

Here we are using a sample www.apixu.com api i.e. http://api.apixu.com/v1/current.json?key=50e18156d7704eac99d93832191704&q&q=jaipur'

In order to get desired data create you own www.apixu.com api and replace it with the above api in print_weatherInfo.py file.

And in this documentation /Weather_feed is the destination directory in HDFS where data is stored, you can change it accordingly in weather.conf file.