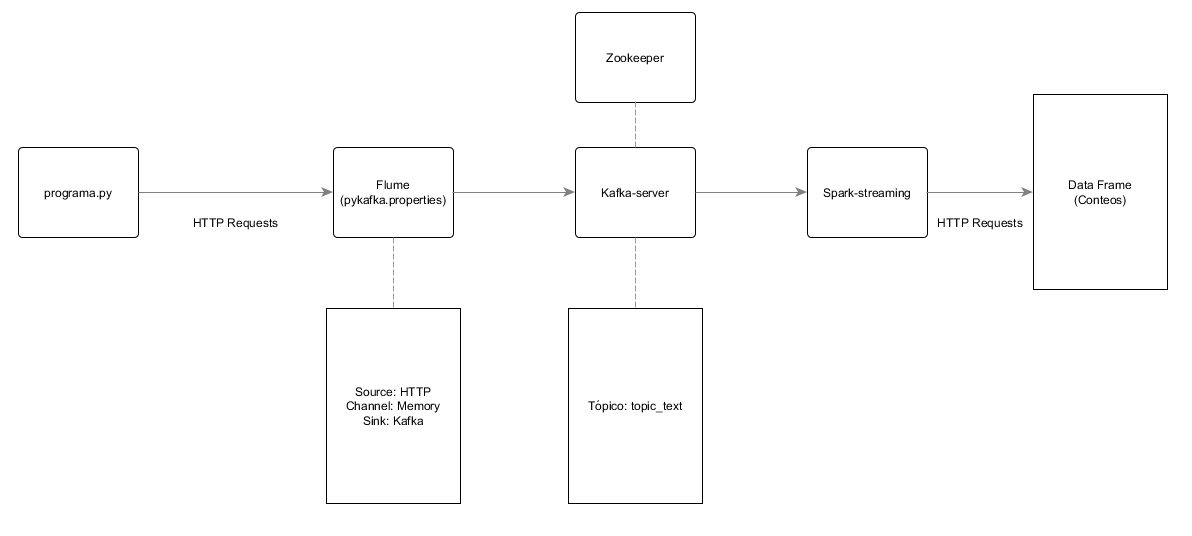
# Creación de flujo desde Python hasta Spark-Streaming



## Python

### Crear programa de Python

El programa lee un archivo y envía requests HTTP:

**flume\_kafka.py**

import requests

import json

import sys

from time import sleep

f= open('sabines','r')

data= f.readlines()

url\_flume = 'http://localhost:9260'

headers = {'content-type': 'application/json'}

for d in data:

print('>>>{}'.format(d))

payload = [{'headers': {}, 'body': d }]

response = requests.post(url\_flume, data=json.dumps(payload),headers=headers)

sleep(1)

print('Terminado.')

### El archivo de muestra:

**sabines**

Espero curarme de ti en unos días.

Debo dejar de fumarte, de beberte, de pensarte.

Es posible.

Siguiendo las prescripciones de la moral en turno.

Me receto tiempo, abstinencia, soledad.

¿Te parece bien que te quiera nada más una semana?

No es mucho, ni es poco, es bastante.

En una semana se puede reunir todas las palabras de amor

que se han pronunciado sobre la tierra y se les puede prender fuego.

Te voy a calentar con esa hoguera del amor quemado.

Y también el silencio.

Porque las mejores palabras del amor están entre dos gentes que no se dicen nada.

Hay que quemar también ese otro lenguaje

lateral y subversivo del que ama.

(Tú sabes cómo te digo que te quiero cuando digo:

"qué calor hace", "dame agua", "¿sabes manejar?",

"se hizo de noche"...

Entre las gentes, a un lado de tus gentes y las mías,

te he dicho "ya es tarde", y tú sabías que decía "te quiero".)

Una semana más para reunir todo el amor del tiempo.

Para dártelo. Para que hagas con él lo que tú quieras:

guardarlo, acariciarlo, tirarlo a la basura.

No sirve, es cierto.

Sólo quiero una semana para entender las cosas.

Porque esto es muy parecido a estar saliendo de un manicomio para entrar a un panteón.

## Flume

### Crear archivo de configuración

**pykafka.properties**

# Name the components on this agent

agent1.sources = http1

agent1.sinks = kafka1

agent1.channels = memory1

# Describe/configure the source

agent1.sources.http1.type = http

agent1.sources.http1.bind = 0.0.0.0

agent1.sources.http1.port = 9260

# Describe the sink

agent1.sinks.kafka1.type = org.apache.flume.sink.kafka.KafkaSink

agent1.sinks.kafka1.kafka.bootstrap.servers= localhost:9092

agent1.sinks.kafka1.kafka.topic= topic\_text

agent1.sinks.kafka1.kafka.flumeBatchSize = 20

agent1.sinks.kafka1.kafka.producer.acks = 1

agent1.sinks.kafka1.kafka.producer.linger.ms = 1

agent1.sinks.kafka1.kafka.producer.compression.type = snappy

# Use a channel which buffers events in memory

agent1.channels.memory1.type = memory

agent1.channels.memory1.capacity = 1000

agent1.channels.memory1.transactionCapacity = 100

# Bind the source and sink to the channel

agent1.sources.http1.channels = memory1

agent1.sinks.kafka1.channel = memory1

## Kafka

**NOTA: Se asume que Zookeeper y Kafka-server están arriba.**

### Crear tópico:

bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic topic\_text

## Spark-streaming

### Descargar archivo \*.jar

Descargar spark-streaming-1.0-SNAPSHOT.jar del Repo.

## Iniciar

**NOTA: Se asume que Zookeeper y Kafka-server están arriba.**

1. Iniciar el agente de Flume

$FLUME\_HOME/bin/flume-ng agent --conf conf --conf-file conf/pykafka.properties --name agent1 -Dflume.root.logger=DEBUG,console

1. Ejecutar spark-submit

$SPARK\_HOME/bin/spark-submit --class com.sparkbyexamples.spark.streaming.kafka.SparkStreamingConsumeKafka --packages org.apache.spark:spark-sql-kafka-0-10\_2.11:2.4.0 spark-streaming-1.0-SNAPSHOT.jar

1. Ejecutar flume\_kafka.py

python3 flume\_kafka.py

