

Relatório de Prática em Laboratório: Desenvolvendo aplicação de fluxo contínuo de dados utilizando Flume e Kafka

Aluno: Carlos Alberto Rocha Cardoso

Professor: Pedro Kássio

Introdução

Este relatório descreve prática desenvolvida em laboratório onde foram utilizados de forma integrada os softwares Flume e Kafka para atividade de ingestão de dados em uma aplicação de fluxo contínuo de dados. Flume e Kafka são softwares open-sources, mantidos pela Apache Foundation, amplamente utilizados no desenvolvimento de pipelines e fluxos para integração e transporte de dados, funcionando em modo distribuído com confiabilidade e escalabilidade. O objetivo do Flume é coletar, agregar e movimentar dados entre fontes e destinos. Já o Kafka funciona como uma fila para produtores e consumidores de dados.

Objetivo

O objetivo da prática descrita neste relatório foi aplicar e consolidar os conhecimentos teóricos compartilhados na disciplina de fluxos contínuos de dados sobre a etapa de ingestão de dados. Para isso foi desenvolvido um fluxo onde o Flume é utilizado para coletar dados em arquivo texto e Twitter, para posteriormente armazená-los em tópicos no Kafka.

Experimentos

Nos tópicos abaixo são apresentados os passos para criação do fluxo, incluindo os comandos e resultados.

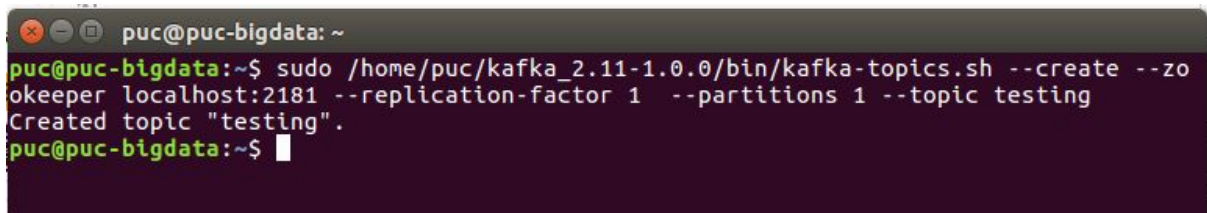
Criando os tópicos no Kafka: Os tópicos são estruturas do Kafka responsáveis por armazenar temporariamente os dados que estão sendo transportados e processados no fluxo. Nessa etapa foram criados os tópicos que receberão os dados que serão coletados pelo Flume.

1. Inicializando o Kafka

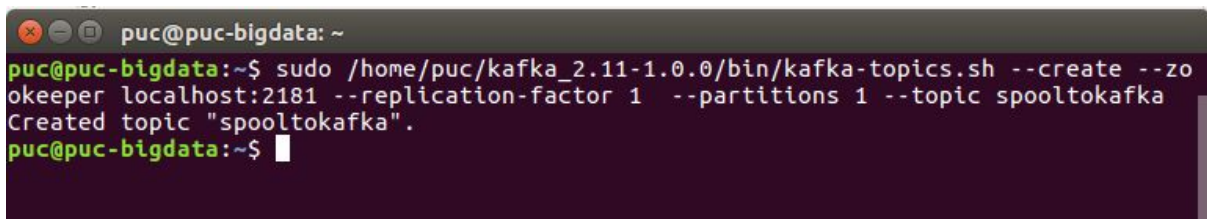


```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-server-start.sh /home/puc/kafka_2.11-1.0.0/config/server.properties
```

2. Criando tópicos no Kafka



```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic testing  
Created topic "testing".  
puc@puc-bigdata:~$
```



```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic spooltokafka  
Created topic "spooltokafka".  
puc@puc-bigdata:~$
```

```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic twittertopic  
Created topic "twittertopic".  
puc@puc-bigdata:~$
```

3. Listando os tópicos criados

```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-topics.sh --list --zookeeper localhost:2181  
spooltokafka  
testing  
twittertopic  
puc@puc-bigdata:~$
```

4. Inserindo strings no tópico testing - producer

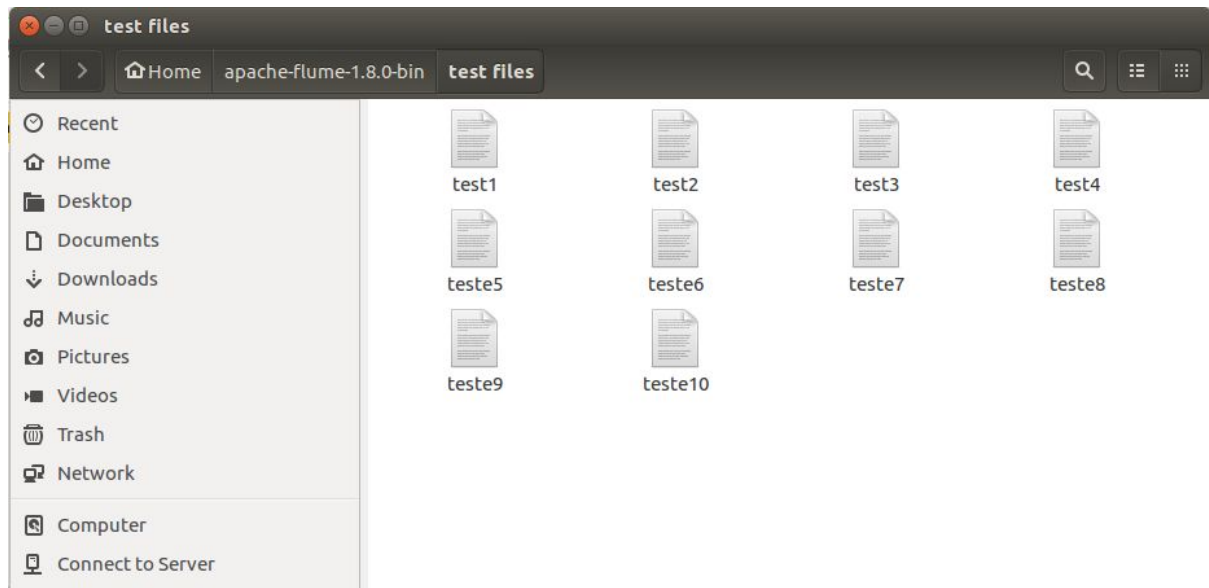
```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-console-producer.sh --broker-list localhost:9092 --topic testing  
>Iniciando teste do Kafka  
>Dado1  
>Dado2  
>Dado3  
>Fim do teste  
>
```

5. Listando strings do tópico testing - consumer

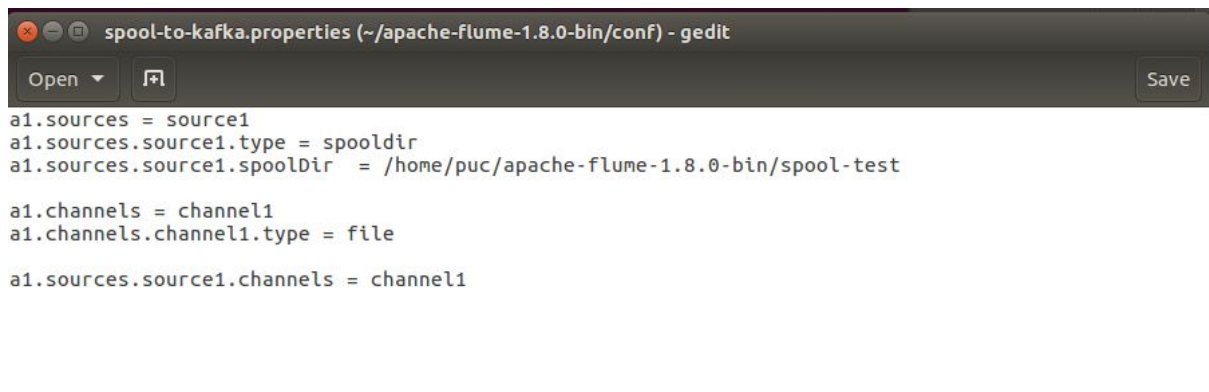
```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-console-consumer.sh --zookeeper localhost:2181 --topic testing --from-beginning  
Using the ConsoleConsumer with old consumer is deprecated and will be removed in a future major release. Consider using the new consumer by passing [bootstrap-server] instead of [zookeeper].  
Iniciando teste do Kafka  
Dado1  
Dado2  
Dado3  
Fim do teste
```

Coletando dados de arquivos texto com o Flume e inserindo no Kafka: Nessa etapa é configurado o agente do Flume para coleta dos dados dos arquivos texto e inserção desses dados em um tópico Kafka. O agente é formado por um source, onde são definidos os parâmetros conexão à fonte de dados, um sink, onde são definidos os parâmetros para conexão a destino dos dados, e um channel, que liga uma fonte aos destinos desejados.

1. Criando arquivos de teste para serem lidos pelo Flume



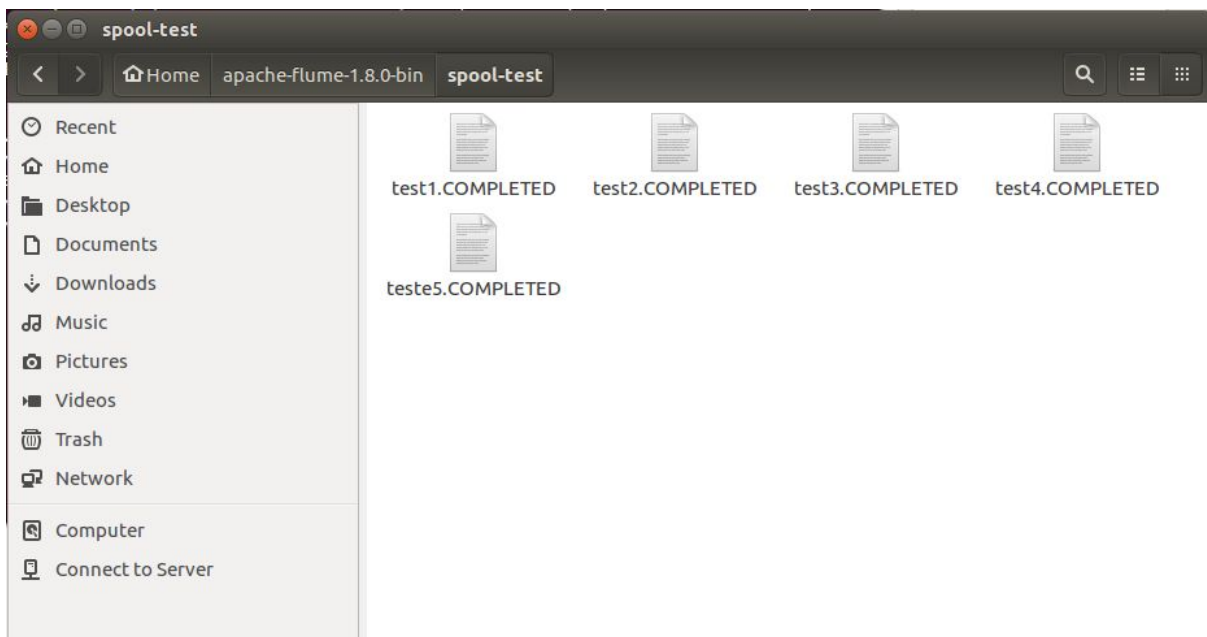
2. Configurando o agente do Flume - source

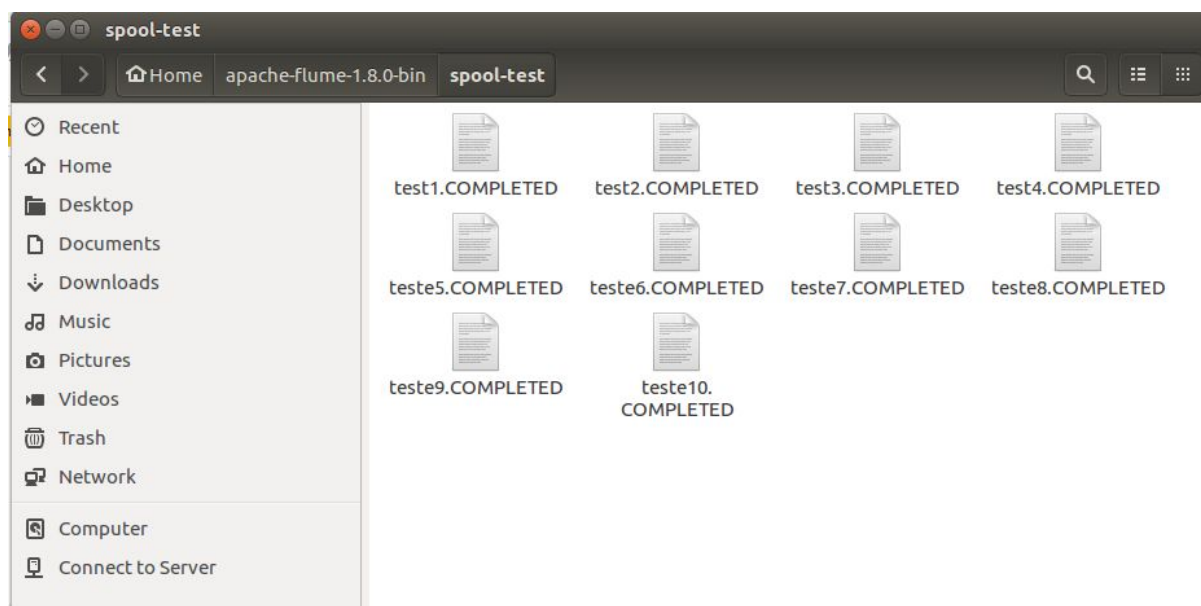


3. Executando o agente do Flume

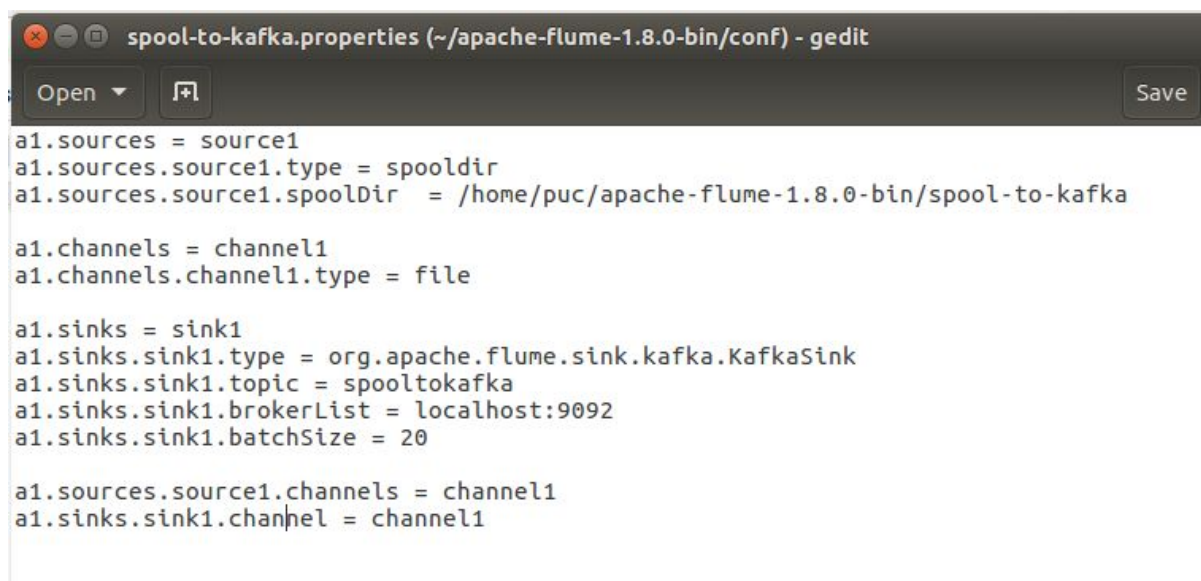
```
puc@puc-bigdata: ~/apache-flume-1.8.0-bin/conf
puc@puc-bigdata:~/apache-flume-1.8.0-bin/conf$ flume-ng agent --conf-file spool-
to-kafka.properties --name a1 -Dflume.root.logger=WARN,console
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hive libraries found via () for Hive access
+ exec /usr/lib/jvm/default-java/bin/java -Xmx20m -Dflume.root.logger=WARN,conso
le -cp '/home/puc/apache-flume-1.8.0-bin/lib/*:/lib/*' -Djava.library.path= org.
apache.flume.node.Application --conf-file spool-to-kafka.properties --name a1
log4j:WARN No appenders could be found for logger (org.apache.flume.lifecycle.Li
fecycleSupervisor).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
```

4. Validando ingestão dos arquivos pelo Flume

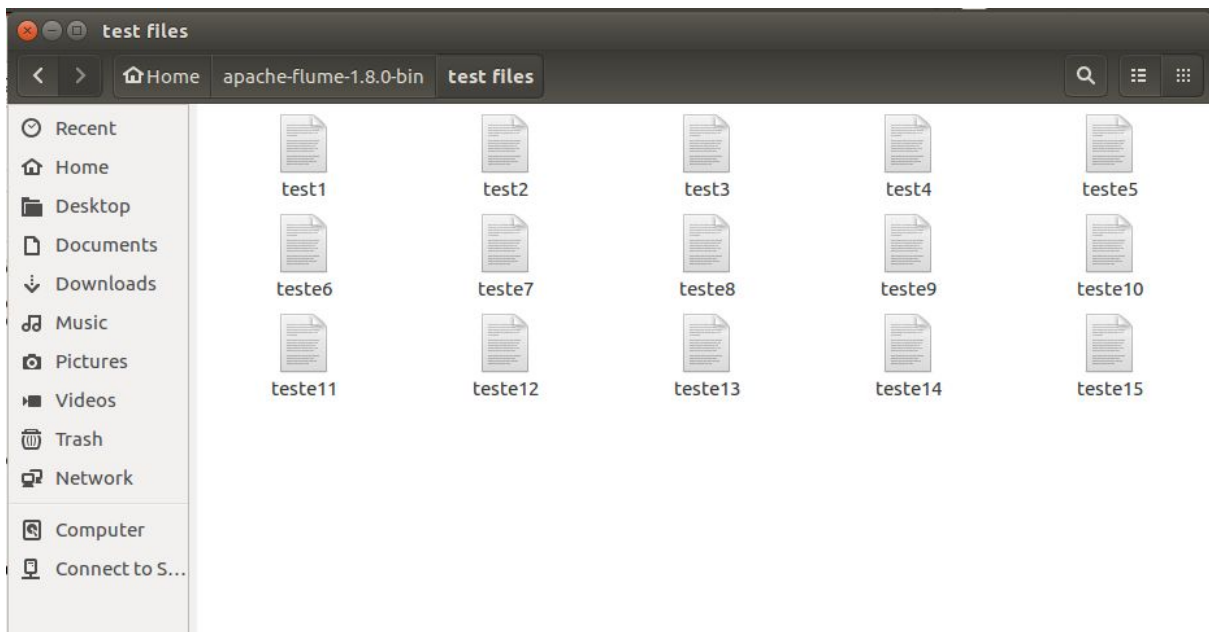




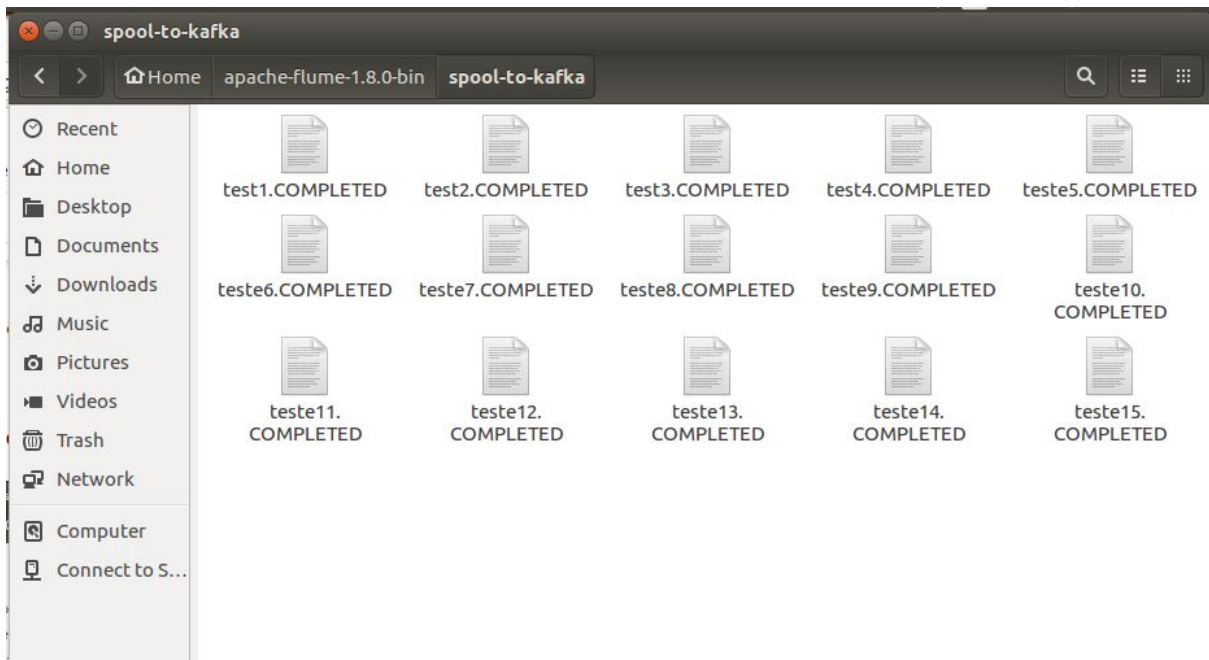
5. Configurando o agente do Flume para gravar no tópic Kafka spooltokafka



6. Arquivos de teste



7. Validando ingestão dos arquivos pelo Flume



8. Validando ingestão dos dados dos arquivos no tópico Kafka

```
puc@puc-bigdata: ~  
puc@puc-bigdata:~$ sudo /home/puc/kafka_2.11-1.0.0/bin/kafka-console-consumer.sh --zookeeper localhost:2181 --topic spooltokafka --from-beginning  
Using the ConsoleConsumer with old consumer is deprecated and will be removed in a future major release. Consider using the new consumer by passing [bootstrap-server] instead of [zookeeper].  
conteúdo teste 1  
conteúdo teste 2  
conteúdo teste 3  
conteúdo teste 4  
conteúdo teste 5  
conteúdo teste 6  
conteúdo teste 7  
conteúdo teste 8  
conteúdo teste 9  
conteúdo teste 10  
conteúdo teste 11  
conteúdo teste 12  
conteúdo teste 13  
conteúdo teste 14  
conteúdo teste 15
```

Coletando dados do Twitter com Flume e inserindo no Kafka: Nessa etapa foi configurado um agente do Flume de forma similar ao passo anterior, sendo que dessa vez o source foi parametrizado para coletar dados da rede social Twitter.

1. Configurando agente do Flume para coletar dados do Twitter e gravar no tópico Kafka twittertopic

```
twitter.properties (~/.apache-flume-1.8.0-bin/conf) - gedit  
Open Save  
# Naming the components on the current agent.  
a3.sources = Twitter  
a3.channels = MemChannel  
a3.sinks = kafkasink  
  
# Describing/Configuring the source  
a3.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource  
a3.sources.Twitter.consumerKey = #####  
a3.sources.Twitter.consumerSecret = #####  
a3.sources.Twitter.accessToken = #####  
a3.sources.Twitter.accessTokenSecret = #####  
a3.sources.Twitter.keywords = brasil, puc, big data, flume, kafka  
  
# Describing/Configuring the sink  
a3.sinks.kafkasink.type = org.apache.flume.sink.kafka.KafkaSink  
a3.sinks.kafkasink.topic = twittertopic  
a3.sinks.kafkasink.brokerList = localhost:9092  
a3.sinks.kafkasink.batchSize = 20  
  
# Describing/Configuring the channel agent3.channels.MemChannel.type = memory  
a3.channels.MemChannel.type = file  
  
# Binding the source and sink to the channel  
a3.sources.Twitter.channels = MemChannel  
a3.sinks.kafkasink.channel = MemChannel
```



```

puc@puc-bigdata: ~
อะไร มึงไม่รู้จักคิลข้อ 3 กับหอย #รักคุณไฉนนายจนกเงินEP14 https://t.c...a href="http://
twitter.com/download/android" rel="nofollow">Twitter for Android</a> in a birb house
in a birb house Cozy poaster/i heart frens/always joking unless I'm serious
in which case I'm just kidding?

@Cuzzin Frenny (spookt birb form)CuzzinFrenny
(2019-10-19T19:52:21ZbWhat was Jesus' Myers Briggs personality profile?a href="
http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
1185690175935938560NKRIPhanya orang desa yg bersyukur apa adanyaUai
yang177(2019-10-19T19:52:21Z|@Anggienatalien @ustadtengkuzul Haha,,, jahad kmu
a href="http://twitter.com/download/android" rel="nofollow">Twit
ter for Android</a>1185690175969484800Medan , Sumatera Utara"thait
ea for lyfe.
if1stiftahteaa(2019-10-19T19:52:21Z@T @jjaerink: @a
skmenfess Ya, semester 5 memang ketemu itu masa jenuh. Tapi, tolong jangan lupa
kan kami yang semester 7 yang dikejar lapora...a href="http://twitter.com/do
wnload/android" rel="nofollow">Twitter for Android</a>1185690175961206784

Buenos Aires, Argentina+Juanjo solisJuanjo13960796(2019-10-19T19:52:21ZnR
T @micaverbic: El único digno https://t.co/zQibk07Lrua href="http://twitte
r.com/download/android" rel="nofollow">Twitter for Android</a>https://pbs.tw
img.com/ext_tw_video_thumb/1185580279655780354/pu/img/u6d3I8pDzsDVw3vr.jpg~http
s://twitter.com/mvideosv/status/1185580364951150594/video/1&1185690175952867331
UCK ICE^insta: uo.slick please follow me in desperate
nqactusi
m_furu(2019-10-19T19:52:21Z@T @richbrian: "gold coast"

one of my fav songs ive ever made right now https://t.co/j1Bdv4Dl8Paa href=
"http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
https://pbs.twimg.com/ext_tw_video_thumb/1184271294398906368/pu/img/v1tN2MkGZiA3X
5Z5.jpg@https://twitter.com/richbrian/status/1184271598058123264/video/1&11856
90175952818179Corpus Christi, TXmile, laugh, and share happiness while ex
Trent1969Trent1969(2019-10-19T19:52:21Z@T @GOP: "Focusing on impeachmen
t doesn't deliver a single job, it doesn't fix health care, it doesn't fix our
broken immigration system. W...a href="http://twitter.com/download/iphone" r
el="nofollow">Twitter for iPhone</a>1185690175965413376a mereLes gens
parlent d'amour, moi j'te parle de c'que j'connais, la squalle /...@byBaby
_prt(2019-10-19T19:52:21Z@T @IeMaeva: quand il est trop beau et que tu l'aime
s trop mais qu'il t'énerv https://t.co/U0ZkG8np2Na href="http://twitter.c
om/download/iphone" rel="nofollow">Twitter for iPhone</a>
kWz
CPProcessed a total of 64 messages
puc@puc-bigdata: ~$

```

Conclusão

Nessa prática foi possível exercitar na prática conhecimentos teóricos compartilhados na disciplina de fluxo contínuos de dados sobre a etapa de ingestão de dados. Além disso foi possível experimentar os softwares Flume e Kafka e comprovar sua utilidade aos processos de fluxo contínuo de dados com garantias de confiabilidade e escalabilidade. O fluxo construído nesta prática compreendeu a etapa de ingestão dos dados. A partir desse ponto, os dados estariam em uma estrutura confiável para aguardar o processamento nas etapas posteriores do fluxo.