

APACHE FLUME



FERRAMENTA DE INGESTÃO DE DADOS STREAMING

TÓPICOS

- HISTÓRIA
- DEFINIÇÃO
- OUTRAS SOLUÇÕES
- APLICAÇÃO
- VISÃO GERAL
- ARQUITETURA
- TIPOS DE FLUXO
- TRANSAÇÕES
- CONFIGURAÇÃO VM
- CONFIGURAÇÃO DO AGENTE
- CONFIGURAÇÃO APP TWITTER
- CONFIGURAÇÃO CLOUDERA
- PROBLEMA ENCONTRADO
- PROBLEMA SOLUCIONADO
- AJUSTES
- TABELAS HIVE
- LIGANDO O AGENTE
- DESLIGANDO O AGENTE
- ANÁLISE
- CONCLUSÕES FINAIS

HISTÓRIA

- EM 2011 O FLUME ENTROU NA INCUBADORA DA FUNDAÇÃO APACHE
- FOI CONSTRUÍDO PELOS ENGENHEIROS DA CLOUDERA PARA ATENDER A AGREGAÇÃO DE DADOS DE LOG EM LARGA ESCALA NO HADOOP
- NO MESMO ANO FOI INTRODUZIDO PELA PRIMEIRA VEZ NA DISTRIBUIÇÃO CDH3 DA CLOUDERA
- DEPOIS DISSO, PASSOU A SER UM PROJETO APACHE DE NÍVEL SUPERIOR, FEZ VÁRIOS LANÇAMENTOS ESTÁVEIS E CRESCERU SIGNIFICATIVAMENTE EM FUNCIONALIDADES
- O FLUME ESTÁ ATIVAMENTE IMPLANTADO E EM USO EM TODO O MUNDO EM GRANDE QUANTIDADE DE CENTRAL DE DADOS, ÀS VEZES ABRANGENDO LIMITES CONTINENTAIS.

DEFINIÇÃO

O FLUME É UM MECANISMO, UMA FERRAMENTA, ISTO É, UM SERVIÇO DE INGESTÃO DE DADOS PARA COLETAR, AGREGAR E TRANSPORTAR GRANDES QUANTIDADES DE FLUXO DE DADOS (STREAMING), COMO POR EXEMPLO: ARQUIVOS DE LOG, EVENTOS, ETC. DE VÁRIAS FONTES PARA UM ARMAZENAMENTO DE DADOS CENTRALIZADO (HBASE, HDFS...)

OUTRAS SOLUÇÕES

PARA ENVIAR DADOS STREAMING (ARQUIVOS DE LOG, EVENTOS, ETC.) DE VÁRIAS FONTES PARA O HDFS:

- **FACEBOOK'S SCRIBE** – O SCRIBE É UMA FERRAMENTA IMENSAMENTE POPULAR QUE É USADA PARA AGREGAR E TRANSMITIR (STREAMING) DADOS DE LOG. ELE É PROJETADO PARA DIMENSIONAR UM NÚMERO MUITO GRANDE DE NÓS E SER ROBUSTO EM RELAÇÃO A FALHAS DE NÓS E DE REDE
- **APACHE KAFKA** – O KAFKA FOI DESENVOLVIDO PELA APACHE SOFTWARE FOUNDATION. É UM AGENTE DE MENSAGENS DE CÓDIGO ABERTO. USANDO A KAFKA, PODEMOS LIDAR COM FEEDS COM ALTA TAXA DE TRANSFERÊNCIA (*HIGH-THROUGHPUT*) E BAIXA LATÊNCIA.

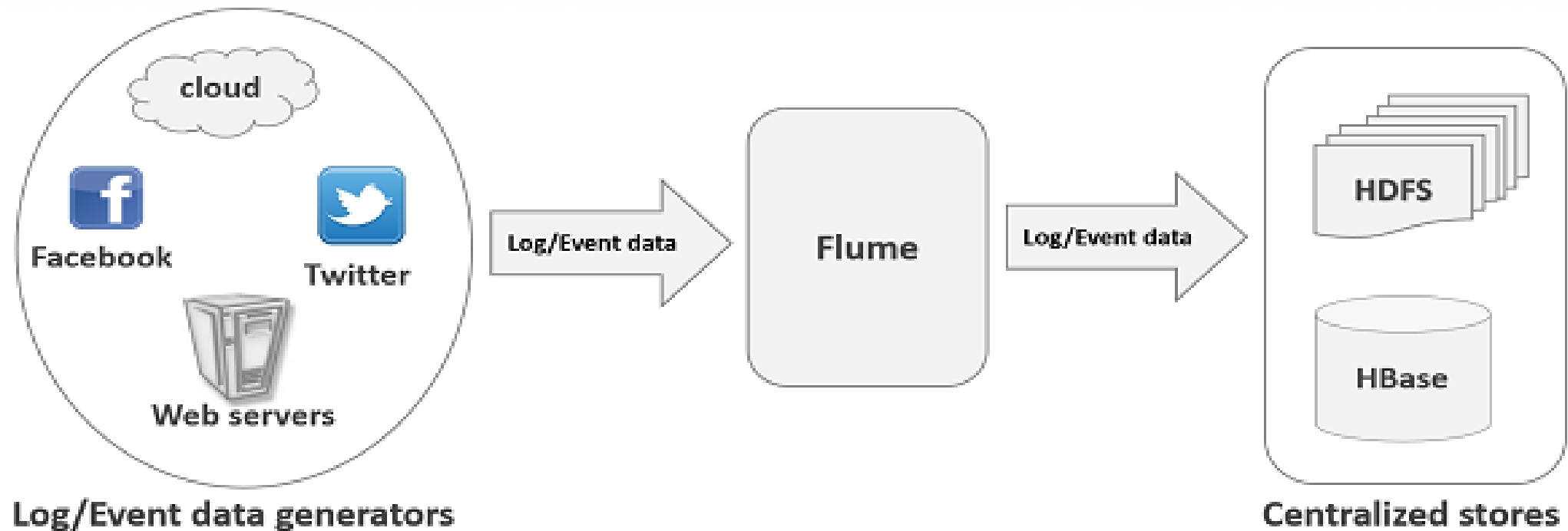
APLICAÇÃO

QUANDO USAR O FLUME?

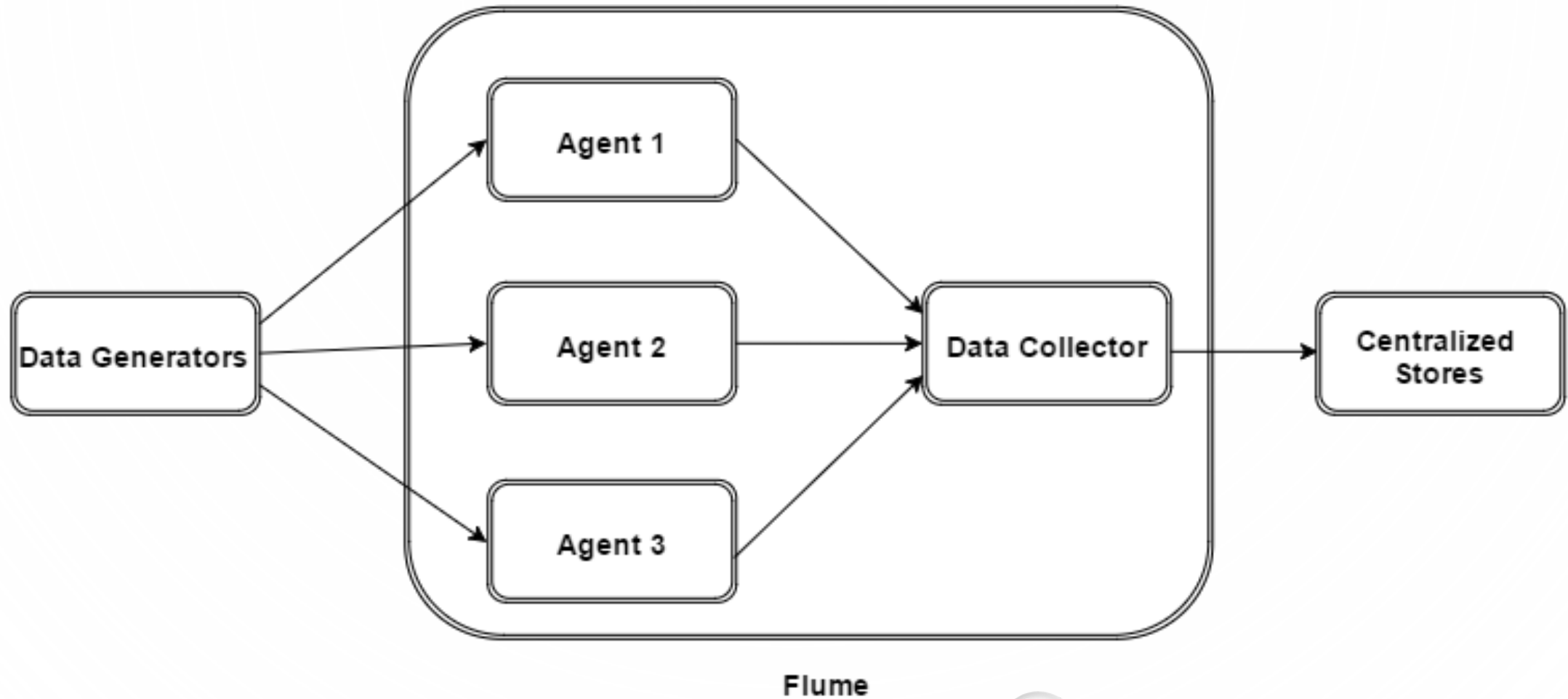
SUPONHA QUE UMA APLICAÇÃO WEB DE COMÉRCIO ELETRÔNICO QUEIRA ANALISAR O COMPORTAMENTO DOS CLIENTES DE UMA DETERMINADA REGIÃO. PARA FAZER ISSO, ELES PRECISARIAM MOVER OS DADOS DE LOG DISPONÍVEIS PARA O HADOOP E DEPOIS ANALISÁ-LOS. ESSE É UM CASO ONDE O FLUME PODE SER A FERRAMENTA ADEQUADA.

VISÃO GERAL

- O FLUME ATUA COMO UM BUFFER ENTRE OS GERADORES DE DADOS E O DESTINO FINAL.



ARQUITETURA



ARQUITETURA

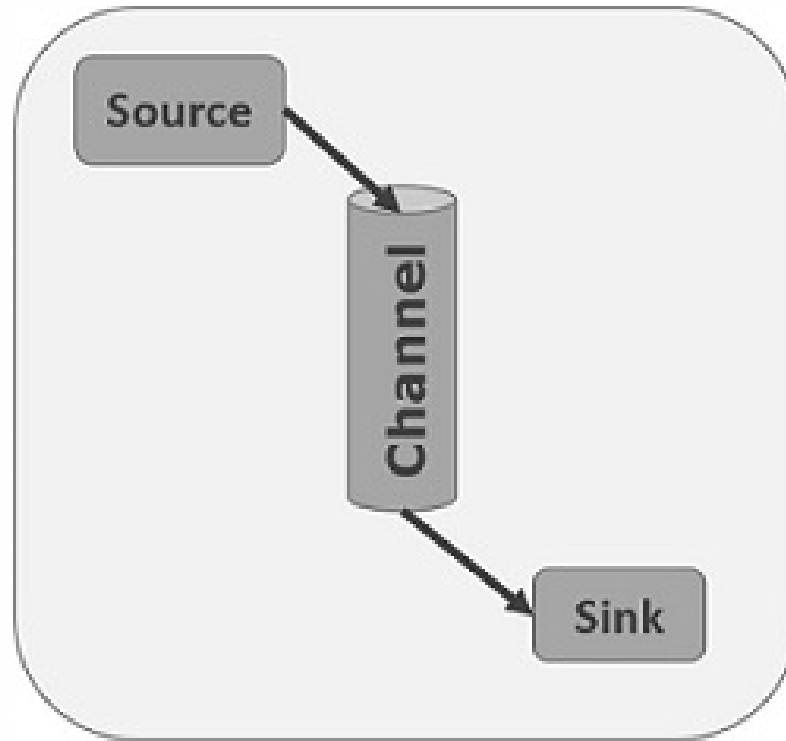
- UM EVENTO (EVENT) É A MENOR UNIDADE DE DADOS QUE TRANSITA NO FLUME
- UM EVENT TEM UM CABEÇALHO (HEADER) OPCIONAL E O DADO EM SI (PAYLOAD)



Flume event

ARQUITETURA

- UM AGENTE É UM PROCESSO INDEPENDENTE RODANDO POR TRÁS (DAEMON) EM UMA JVM



Flume Agent

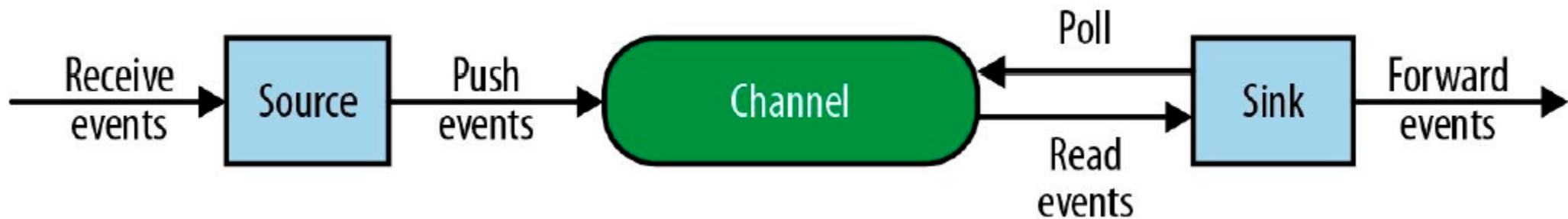
Recebem os dados (eventos) de clientes ou outros agentes e encaminha-o para seu próximo destino (coletor ou agente).

ARQUITETURA

- UM AGENTE FLUME CONTÉM TRÊS COMPONENTES PRINCIPAIS:
 - SOURCE - RECEBE DADOS DOS GERADORES DE DADOS E TRANSFERE-OS PARA UM OU MAIS CHANNEL NA FORMA DE EVENTOS FLUME.
 - CHANNEL - É UM ARMAZENAMENTO TRANSITÓRIO QUE RECEBE OS EVENTOS DO SOURCE E OS PROTEGE ATÉ SEREM CONSUMIDOS PELO SINK. ATUA COMO UMA PONTE ENTRE OS SOURCE E OS SINKS.
 - SINK - ARMAZENAM OS DADOS DE FORMA CENTRALIZADA NO HBASE E HDFS. CONSOME OS DADOS (EVENTOS) DOS CANAIS E OS ENTREGA AO DESTINO. O DESTINO DO SINK PODE SER OUTRO AGENTE OU UM ARMAZENAMENTO CENTRALIZADO (HBASE, HDFS)

ARQUITETURA

- DENTRO DE UM AGENTE FLUME COM 1 SOURCE, 1 CHANNEL E 1 SINK.



ARQUITETURA

- MAIS ALGUNS COMPONENTES QUE DESEMPENHAM UM PAPEL VITAL:
 - INTERCEPTORS - SÃO USADOS PARA ALTERAR/INSPECIONAR EVENTOS FLUME QUE SÃO TRANSFERIDOS ENTRE O SOURCE E O CHANNEL.
 - CHANNEL SELECTORS - EXISTEM DOIS TIPOS DE SELETORES DE CANAIS:
 - DEFAULT CHANNEL SELECTORS - SELETORES DE CANAIS DE REPLICAÇÃO QUE REPLICAM TODOS OS EVENTOS EM CADA CANAL.
 - MULTIPLEXING CHANNEL SELECTORS - DECIDE PARA QUE CANAL ENVIAR UM EVENTO COM BASE NO ENDEREÇO NO CABEÇALHO (HEADER) DESSE EVENTO
 - COLLECTORS – COLETA OS DADOS DOS AGENTES, OS DADOS DE TODOS OS COLLECTORS SÃO AGREGADOS E ENVIADOS PARA O ARMAZENAMENTO (HBASE, HDFS).

TIPOS DE FLUXO

- **MULTI-HOP FLOW** – QUANDO EXISTEM VÁRIOS AGENTES E ANTES DE ATINGIR O DESTINO FINAL, UM EVENTO PERCORRE MAIS DE UM AGENTE.
- **FAN-OUT FLOW** – FLUXO DE DADOS DE UMA FONTE PARA VÁRIOS CANAIS.
 - **REPLICATING** – OS DADOS SÃO REPLICADOS EM TODOS OS CANAIS CONFIGURADOS.
 - **MULTIPLEXING** – OS DADOS SÃO ENVIADOS PARA UM CANAL SELECIONADO QUE É DEFINIDO NO CABEÇALHO DO EVENTO.
- **FAN-IN FLOW** – FLUXO DE DADOS EM QUE OS DADOS SÃO TRANSFERIDOS DE MUITAS FONTES PARA UM CANAL.

TRANSAÇÕES

- PARA CADA EVENTO, OCORREM DUAS TRANSAÇÕES: UMA NO REMETENTE E OUTRA NO DESTINATÁRIO. O REMETENTE ENVIA EVENTOS PARA O DESTINATÁRIO. LOGO APÓS RECEBER OS DADOS, O DESTINATÁRIO FINALIZA SUA PRÓPRIA TRANSAÇÃO E ENVIA UM SINAL DE "RECEBIDO" PARA O REMETENTE.
- DEPOIS DE RECEBER O SINAL, O REMETENTE FINALIZA SUA TRANSAÇÃO. (O REMETENTE NÃO FINALIZA SUA TRANSAÇÃO ATÉ RECEBER UM SINAL DO DESTINATÁRIO).

CONFIGURAÇÃO VM

- PARA ESSE PROJETO FOI USADA A MÁQUINA CLOUDERA-QUICKSTART-VM-5.13.0-0 PARA VIRTUALBOX, DISPONÍVEL PARA DOWNLOAD EM:

https://www.cloudera.com/downloads/quickstart_vms/5-13.html

IMPORTE A MÁQUINA E CONFIGURE, PARA ISSO VÁ EM:

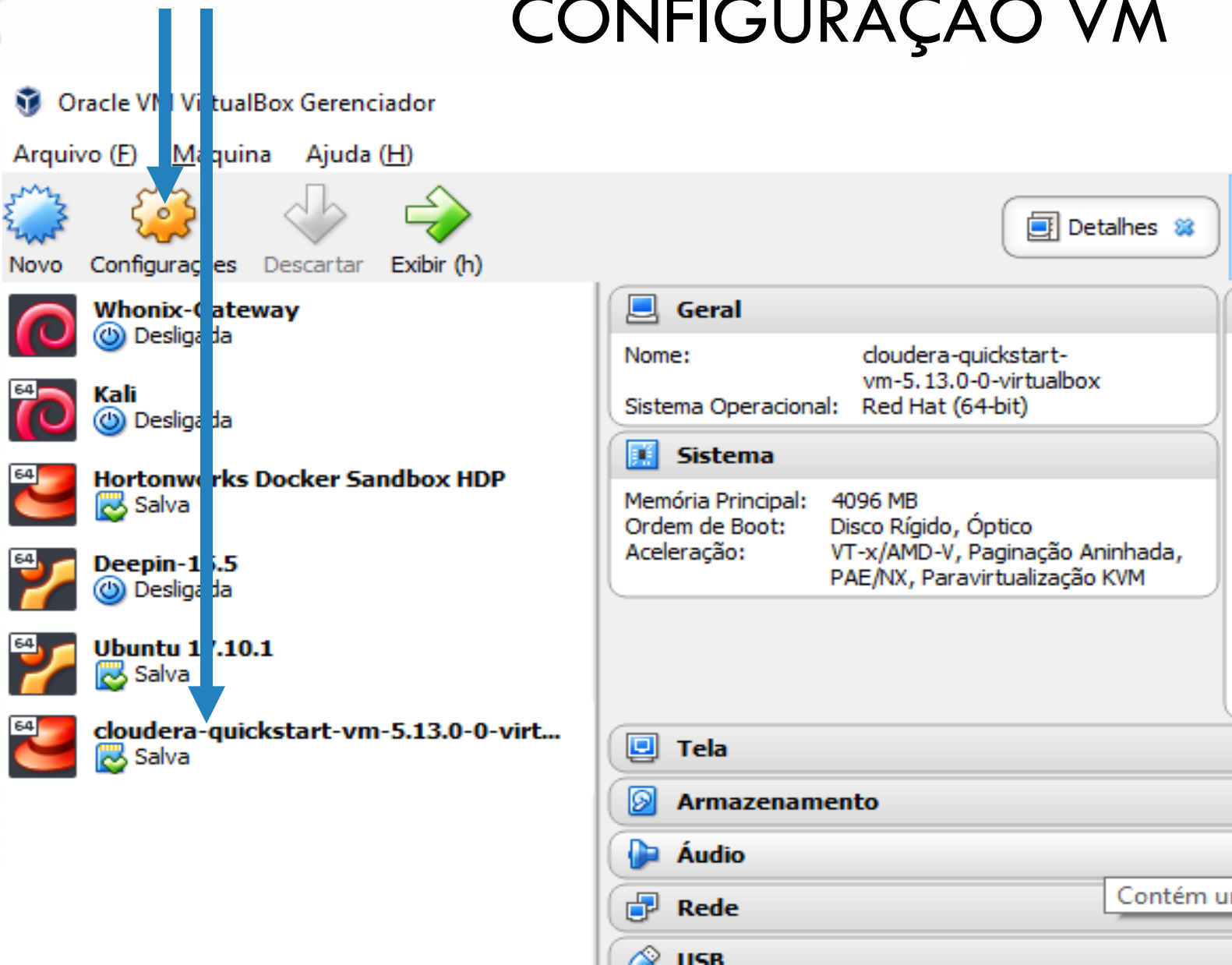
CONFIGURAÇÃO / REDE / CLICAR NA ABA ADAPTADOR 2

EM CONECTADO A: PLACA DE REDE EXCLUSIVA DE HOSPEDEIRO (HOST-ONLY)

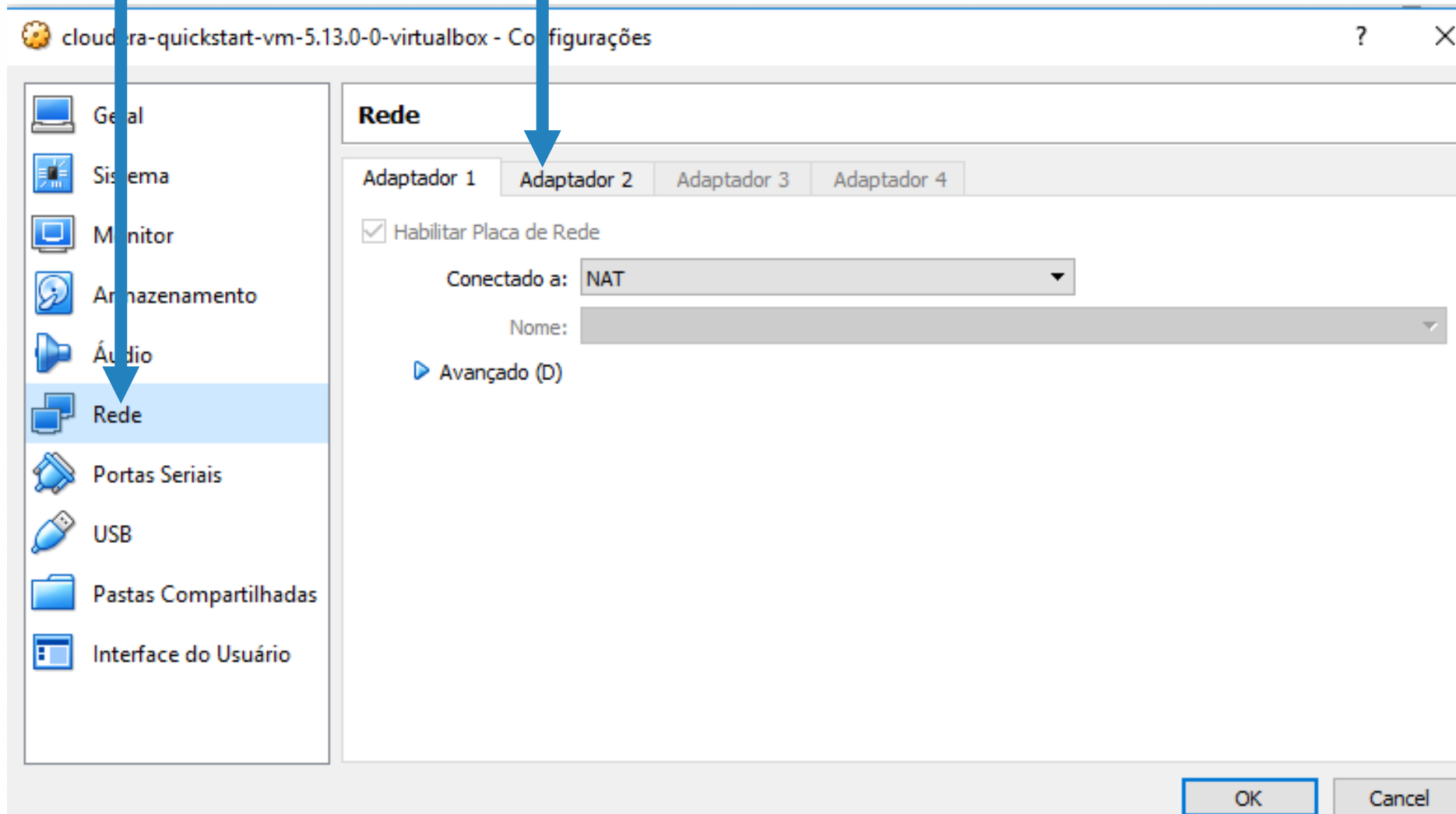
EM NOME: VIRTUALBOX HOST-ONLY ETHERNET ADAPTER

CLICAR EM OK

CONFIGURAÇÃO VM



CONFIGURAÇÃO VM



CONFIGURAÇÃO VM

cloudera-quickstart-vm-5.10.0-0-virtualbox - Configurações

? X

- Geral
- Sistema
- Monitor
- Armazenamento
- Áudio
- Rede**
- Portas Seriais
- USB
- Pastas Compartilhadas
- Interface do Usuário

Rede

Adaptador 1

Adaptador 2

Adaptador 3

Adaptador 4

☒ Habilitar Placa de Rede

Conectar a: Placa de rede exclusiva de hospedeiro (host-only)

Nome: VirtualBox Host-Only Ethernet Adapter

▶ Avançado (D)

OK

Cancel

CONFIGURAÇÃO VM

- O USER PODE SER ROOT OU CLOUDERA
- PASSWORD É CLOUDERA PARA OS DOIS USERS
- DEPOIS DE INICIADA ABRIR UM TERMINAL E DÁ UM IFCONFIG PARA PEGAR O IP DO ETH1
- COM ESSE IP É POSSÍVEL ACESSAR A MÁQUINA VIA SSH (PUTTY) E WINSCP NA PORTA 22
- A MÁQUINA TEM UMA INTERFACE WEB TAMBÉM

NO BROWSER ACESSE: [http:// 192.168.56.101:8888](http://192.168.56.101:8888)

CASO FAÇA UM UPDATE COMPLETO NA MÁQUINA USE:

<http://192.168.56.101:8888/hue>

CONFIGURAÇÃO DO AGENTE

- CADA SOURCE TERÁ UMA LISTA SEPARADA DE PROPRIEDADES. A PROPRIEDADE DENOMINADA "TYPE" É COMUM A TODOS OS SOURCES E É USADA PARA ESPECIFICAR O TIPO DE SOURCE QUE SERÁ USADO.

Nome do agente



Tipo



Nome



Parâmetro de configuração



AGENT_NAME.SOURCES.SOURCE_NAME.TYPE = VALUE

AGENT_NAME.SOURCES.SOURCE_NAME.PROPERTY1 = VALUE

AGENT1.SOURCES.SOURCE1.PORT = 4144

Nome
do
agente

Tipo

Valor

CONFIGURAÇÃO DO AGENTE

Nomea os componentes do agente

a1.sources = r1

a1.sinks = k1

a1.channels = c1

Configura o source

a1.sources.r1.type = netcat

a1.sources.r1.bind = localhost

a1.sources.r1.port = 44444

Descreva o sink

a1.sinks.k1.type = logger

Use um canal que armazene os eventos na memória

a1.channels.c1.type = memory

a1.channels.c1.capacity = 1000

a1.channels.c1.transactionCapacity = 100

Vincula o source ao sink e ao canal

a1.sources.r1.channels = c1

a1.sinks.k1.channel = c1

CONFIGURAÇÃO APP TWITTER

- ACESSAR O ENDEREÇO E CRIAR UMA APP: [HTTPS://APPS.TWITTER.COM/](https://apps.twitter.com/)
- CRIAR LOGIN E SENHA E LOGAR
- CRIAR UMA NOVA APP CLICANDO EM [CREATE NEW APP](#)
- DEFINIR OS DETALHES DA APLICAÇÃO: NOME, DESCRIÇÃO, WEBSITE, ETC
- CLICAR EM [CREATE MY ACCESS TOKEN](#) PARA GERAR AS CHAVES DA APP PARA USAR NA CONFIGURAÇÃO DO FLUME

CONFIGURAÇÃO CLOUDERA

- ADICIONAR A LINHA ABAIXO EM `/etc/hosts` DA MÁQUINA CLOUDERA:

`199.59.148.138 stream.twitter.com`

- ALGUMAS VEZES É NECESSÁRIO ATUALIZAR O DATETIME DA MÁQUINA ANTES DE RODAR O AGENTE:

`sudo ntpdate ntp.ubuntu.com`

- CRIAR O DIRETÓRIO NO HDFS PARA SALVAR OS ARQUIVOS FLUMEDATA

`hadoop fs -mkdir -p /twitteranalytics/`

NO DIRETÓRIO DA MÁQUINA CLOUDERA `/etc/flume-ng/conf/` CRIAR O ARQUIVO `flume_process_twitter` E INSERIR O CONTEÚDO A SEGUIR:

TWITTER.CONF

```
TwitterAgent.sources=Twitter  
TwitterAgent.channels=MemChannel  
TwitterAgent.sinks=HDFS
```

Describing/Configuring the source

```
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource  
#TwitterAgent.sources.Twitter.type=org.apache.flume.source.twitter.TwitterSource  
TwitterAgent.sources.Twitter.consumerKey=xxxxxxxxxxxxxxxx  
TwitterAgent.sources.Twitter.consumerSecret=xxxxxxxxxxxxxxxx  
TwitterAgent.sources.Twitter.accessToken=xxxxxxxxxxxxxxxx  
TwitterAgent.sources.Twitter.accessTokenSecret=xxxxxxxxxxxx  
TwitterAgent.sources.Twitter.keywords=Suas, Keywords,...
```

TWITTER.CONF

Describing/Configuring the sink

TwitterAgent.sinks.HDFS.channel=MemChannel

TwitterAgent.sinks.HDFS.type=hdfs

TwitterAgent.sinks.HDFS.hdfs.path=/twitteranalytics/incremental

TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream

TwitterAgent.sinks.HDFS.hdfs.writeformat=Text

TwitterAgent.sinks.HDFS.hdfs.batchSize=1 000

TwitterAgent.sinks.HDFS.hdfs.rollSize=0

TwitterAgent.sinks.HDFS.hdfs.rollCount=1 0000

TwitterAgent.sinks.HDFS.hdfs.rollInterval=600

TwitterAgent.channels.MemChannel.type=memory

TwitterAgent.channels.MemChannel.capacity=1 0000

TwitterAgent.channels.MemChannel.transactionCapacity=1 000

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sinks.HDFS.channel = MemChannel

AJUSTES

- SE NÃO TIVER SIDO CRIADO O ARQUIVO `twitter.conf` em `/etc/flume-ng/conf/`, COPIE ELE PARA LÁ USANDO:

```
cp twitter.conf /etc/flume-ng/conf/
```

PROBLEMA ENCONTRADO

USANDO `org.apache.flume.source.twitter.TwitterSource` NA CONFIGURAÇÃO DO AGENTE, O ARQUIVO GERADO É TOTALMENTE ILEGÍVEL.

Objavro.schemaä

```
{ "type": "record", "name": "Doc", "doc": "adoc", "fields": [ { "name": "id", "type": "string" }, { "name": "user_friends_count", "type": [ "int", "null" ] }, { "name": "user_location", "type": [ "string", "null" ] }, { "name": "user_description", "type": [ "string", "null" ] }, { "name": "user_statuses_count", "type": [ "int", "null" ] }, { "name": "user_followers_count", "type": [ "int", "null" ] }, { "name": "user_name", "type": [ "string", "null" ] }, { "name": "user_screen_name", "type": [ "string", "null" ] }, { "name": "created_at", "type": [ "string", "null" ] }, { "name": "text", "type": [ "string", "null" ] }, { "name": "retweet_count", "type": [ "long", "null" ] }, { "name": "retweeted", "type": [ "boolean", "null" ] }, { "name": "in_reply_to_user_id", "type": [ "long", "null" ] }, { "name": "source", "type": [ "string", "null" ] }, { "name": "in_reply_to_status_id", "type": [ "long", "null" ] }, { "name": "media_url_https", "type": [ "string", "null" ] }, { "name": "expanded_url", "type": [ "string", "null" ] } ] } ø □ > ‡ Û ó °, ç i ç À ß, æ Æ $ 967438715848200192 ì † ž | DaniðŸ“
Danimarquez21 (2018-02-24T16:38:58Z @ðŸ™ ðŸ™ https://t.co/11X88sku4T ^<a href="https://mobile.twitter.com" rel="nofollow">Twitter
Lite</a> ^https://pbs.twimg.com/media/DW0H2moWsAEKuXv.jpg
†https://twitter.com/Danmarquez21/status/967438715848200192/photo/1 $967438720071819267 ì Mic4Mimundo òŸ

' lili francoðŸŒŠ Alberto61898426 (2018-02-24T16:38:59Z æRT @byforviconte: @AngeldebritoOk Las dos son lindas! La diferencia q Micaela no
es para nada falsa.Igual no entiendo porque quieren enfâ€| š<a href="http://www.twitter.com" rel="nofollow">Twitter for Windows Phone</a>
$967438720067661825 ¶ □ ABCD ABCDABCD98 (2018-02-24T16:38:59Z òRT @AwatefMM: ÛŠøŸøø·ÛŠ Û„Û„ÛfÛ·ÛŠøª Û·øø·ÛŠ
Û„øŸÛ‡Û„Û‡øŸ ÛŠøŸøø¬ø¹Û„ÛŠ øŸÛ† ø·øŸøiøŸÛ„Û„Û‡ øŸø³ÛfÛ† ÛÛŠÛ‡øŸ ÛŠøŸø±ø· æ<a href="http://twitter.com/download/iphone"
rel="nofollow">Twitter for iPhone</a> $967438720050892800 ¼ Portugal ,FaÃŸo ficheiros MP4 para o YouTube. FaÃŸo muitos tweets a brincar Ú®
ØÑ T7agox T7agoxOficial (2018-02-24T16:38:59Z îRT @GenotDzn: Tavas a precisar de uma nova @T7agoxOficial Espero que gostes :3
https://t.co/UiVaQ1MDNK æ<a href="http://twitter.com/download/iphone"
```

PROBLEMA ENCONTRADO

IMPOSSÍVEL A DESSERIALIZAÇÃO DESSES ARQUIVOS!

ERRO:

OK

Failed with exception

java.io.IOException:org.apache.avro.AvroRuntimeException: java.io.IOException:
Block size invalid or too large for this implementation: -40

Time taken: 0.156 seconds

PARA CONTORANAR O PROBLEMA FOI USADO:

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

PROBLEMA SOLUCIONADO

`TwitterAgent.sources.Twitter.type=org.apache.flume.source.twitter.TwitterSource`

`TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource`

MAS, PARA USAR `com.cloudera.flume.source.TwitterSource` SÃO NECESSÁRIOS DOIS ARQUIVOS:

`flume-sources-1.0-SNAPSHOT.jar`

`hive-serdes-1.0-SNAPSHOT.jar`

ESSES ARQUIVOS PRECISAM SER COLOCADOS EM PASTAS ESPECÍFICAS COMO VEREMOS MAIS A FRENTE.

LINKS PARA DOWNLOAD:

<https://github.com/toticavalcanti/projeto-flume-twitter-spark/raw/master/flume-sources-1.0-SNAPSHOT.jar>

<https://github.com/toticavalcanti/projeto-flume-twitter-spark/raw/master/hive-serdes-1.0-SNAPSHOT.jar>

AJUSTES

- CRIAR A PASTA `/twitteranalytics/` NO HDFS

```
hadoop fs -mkdir -p /twitteranalytics/
```

- CRIAR O DIRETÓRIO `/usr/lib/flume-ng/plugins.d/twitter-streaming/lib/`:

```
mkdir -p /usr/lib/flume-ng/plugins.d/twitter-streaming/lib/
```

- CRIAR O DIRETÓRIO `/var/lib/flume-ng/plugins.d/twitter-streaming/lib/`:

```
mkdir -p /var/lib/flume-ng/plugins.d/twitter-streaming/lib/
```

- COPIAR O ARQUIVO `flume-sources-1.0-SNAPSHOT.jar` PARA DENTRO DAS DUAS PASTAS CRIADAS:

```
cp flume-sources-1.0-SNAPSHOT.jar /usr/lib/flume-ng/plugins.d/twitter-streaming/lib/
```

```
cp flume-sources-1.0-SNAPSHOT.jar /var/lib/flume-ng/plugins.d/twitter-streaming/lib/
```

AJUSTES

- COPIE O `hive-serdes-1.0-SNAPSHOT.jar` PARA A PASTA `/usr/lib/hive/lib`

`sudo cp hive-serdes-1.0-SNAPSHOT.jar /usr/lib/hive/lib`

AJUSTES

- PARE O HIVE:

```
sudo service hive-server2 stop
```

- DEPOIS REINICIE:

```
sudo service hive-server2 start
```

CRIANDO TABELAS HIVE

- BAIXE O ARQUIVO `create_twitter_schema.hql` NO LINK:

https://github.com/toticavalcanti/projeto-flume-twitter-spark/blob/master/create_twitter_schema.hql

- NO PROMPT DO SHELL, PARA GERAR AS TABELAS HIVE, DIGITE:

```
hive -f Create_Twitter_Schema.hql
```

SÃO GERADAS AS TABELAS E VIEW:

`base_tweets`

`candidate_score`

`incremental_tweets`

`reconcile_view`

LIGANDO O AGENTE

- AGORA É SÓ RODAR O AGENTE FLUME, PARA ISSO ENTRE NA PASTA `/etc/flume-ng/conf` COM O COMANDO:

```
cd /etc/flume-ng/conf
```

E RODE:

```
flume-ng agent -n twitteragent -c conf -f /etc/flume/conf/twitter.conf -dflume.root.logger=debug,console -  
n TwitterAgent
```

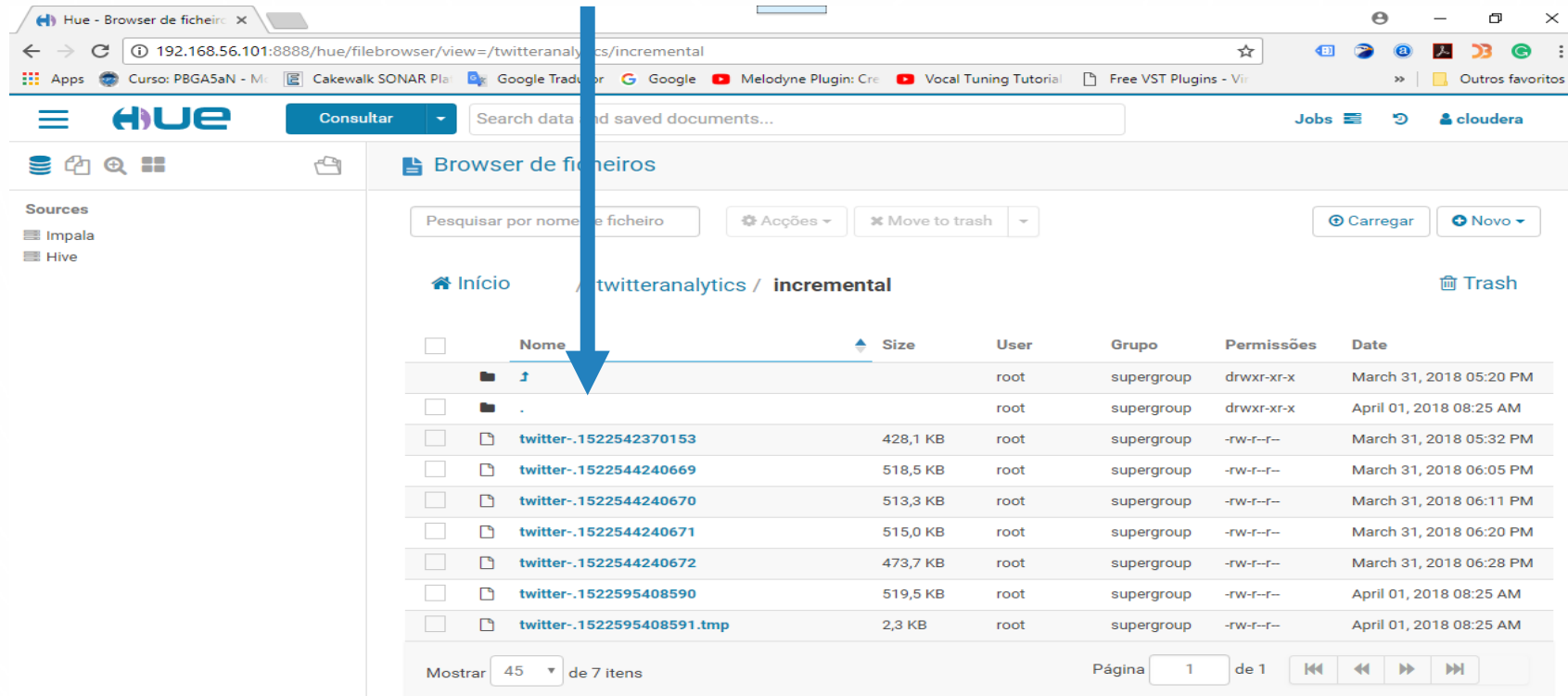
DESLIGANDO O AGENTE

- PARA PARAR O AGENTE USE:

ctrl c

TABELAS HIVE

- ARQUIVOS GERADOS NA PASTA **twitteranalytics/incremental**



Hue - Browser de ficheiros

192.168.56.101:8888/hue/filebrowser/view=twitteranalytics/incremental

Consultar

Search data and saved documents...

Jobs cloudera

Browser de ficheiros

Pesquisar por nome de ficheiro

Acções










Move to trash

Carregar

Novo

Início / twitteranalytics / incremental

Trash

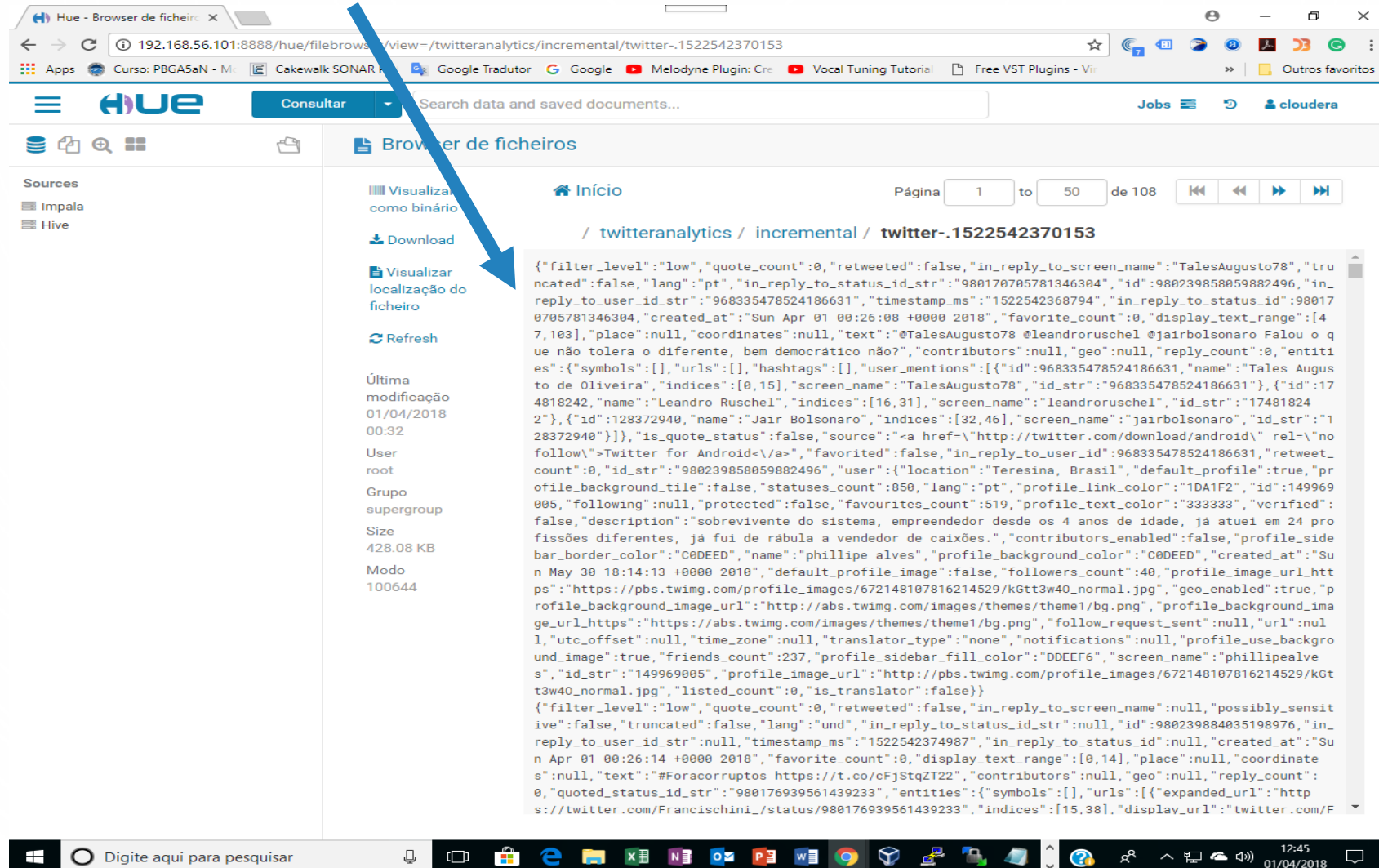
	Nome	Size	User	Grupo	Permissões	Date
<input type="checkbox"/>			root	supergroup	drwxr-xr-x	March 31, 2018 05:20 PM
<input type="checkbox"/>			root	supergroup	drwxr-xr-x	April 01, 2018 08:25 AM
<input type="checkbox"/>	 twitter-.1522542370153	428,1 KB	root	supergroup	-rw-r--r--	March 31, 2018 05:32 PM
<input type="checkbox"/>	 twitter-.1522544240669	518,5 KB	root	supergroup	-rw-r--r--	March 31, 2018 06:05 PM
<input type="checkbox"/>	 twitter-.1522544240670	513,3 KB	root	supergroup	-rw-r--r--	March 31, 2018 06:11 PM
<input type="checkbox"/>	 twitter-.1522544240671	515,0 KB	root	supergroup	-rw-r--r--	March 31, 2018 06:20 PM
<input type="checkbox"/>	 twitter-.1522544240672	473,7 KB	root	supergroup	-rw-r--r--	March 31, 2018 06:28 PM
<input type="checkbox"/>	 twitter-.1522595408590	519,5 KB	root	supergroup	-rw-r--r--	April 01, 2018 08:25 AM
<input type="checkbox"/>	 twitter-.1522595408591.tmp	2,3 KB	root	supergroup	-rw-r--r--	April 01, 2018 08:25 AM

Mostrar 45 de 7 itens

Página 1 de 1

CONTEÚDO DOS ARQUIVOS

• CONTEÚDO DOS ARQUIVOS GERADOS PELO AGENTE



The screenshot displays the Hue web interface, specifically the 'Browser de ficheiros' (File Browser) view. The browser window shows the URL `192.168.56.101:8888/hue/filebrowser/view=/twitteranalytics/incremental/twitter-.1522542370153`. The interface includes a search bar, navigation buttons, and a sidebar with 'Sources' (Impala, Hive) and 'Visualizar como binário' (View as binary) options. A blue arrow points to the 'Visualizar localização do ficheiro' (View file location) button. The main content area shows the file path `/ twitteranalytics / incremental / twitter-.1522542370153` and a JSON file viewer displaying the content of the file. The JSON data includes fields like `filter_level`, `quote_count`, `retweeted`, `in_reply_to_screen_name`, `truncated`, `lang`, `in_reply_to_status_id_str`, `id`, `in_reply_to_user_id_str`, `timestamp_ms`, `created_at`, `favorite_count`, `display_text_range`, `place`, `coordinates`, `text`, `contributors`, `geo`, `reply_count`, `entities`, `symbols`, `urls`, `hashtags`, `user_mentions`, `indices`, `screen_name`, `id_str`, `location`, `default_profile`, `profile_background_tile`, `statuses_count`, `lang`, `profile_link_color`, `id`, `following`, `protected`, `favorites_count`, `profile_text_color`, `verified`, `description`, `contributors_enabled`, `profile_sidebar_border_color`, `name`, `phillipe alves`, `profile_background_color`, `created_at`, `default_profile_image`, `followers_count`, `profile_image_url_https`, `profile_background_image_url`, `profile_background_image_url_https`, `follow_request_sent`, `url`, `utc_offset`, `time_zone`, `translator_type`, `notifications`, `profile_use_background_image`, `friends_count`, `profile_sidebar_fill_color`, `screen_name`, `phillipealves`, `id_str`, `profile_image_url`, `profile_image_url_https`, `listed_count`, `is_translator`, `filter_level`, `quote_count`, `retweeted`, `in_reply_to_screen_name`, `possibly_sensitive`, `truncated`, `lang`, `und`, `in_reply_to_status_id_str`, `id`, `in_reply_to_user_id_str`, `timestamp_ms`, `created_at`, `favorite_count`, `display_text_range`, `place`, `coordinates`, `text`, `quoted_status_id_str`, `entities`, `symbols`, `urls`, `expanded_url`, `twitter.com/Francischini_/status/980176939561439233`, `indices`, `display_url`, `twitter.com/F`.

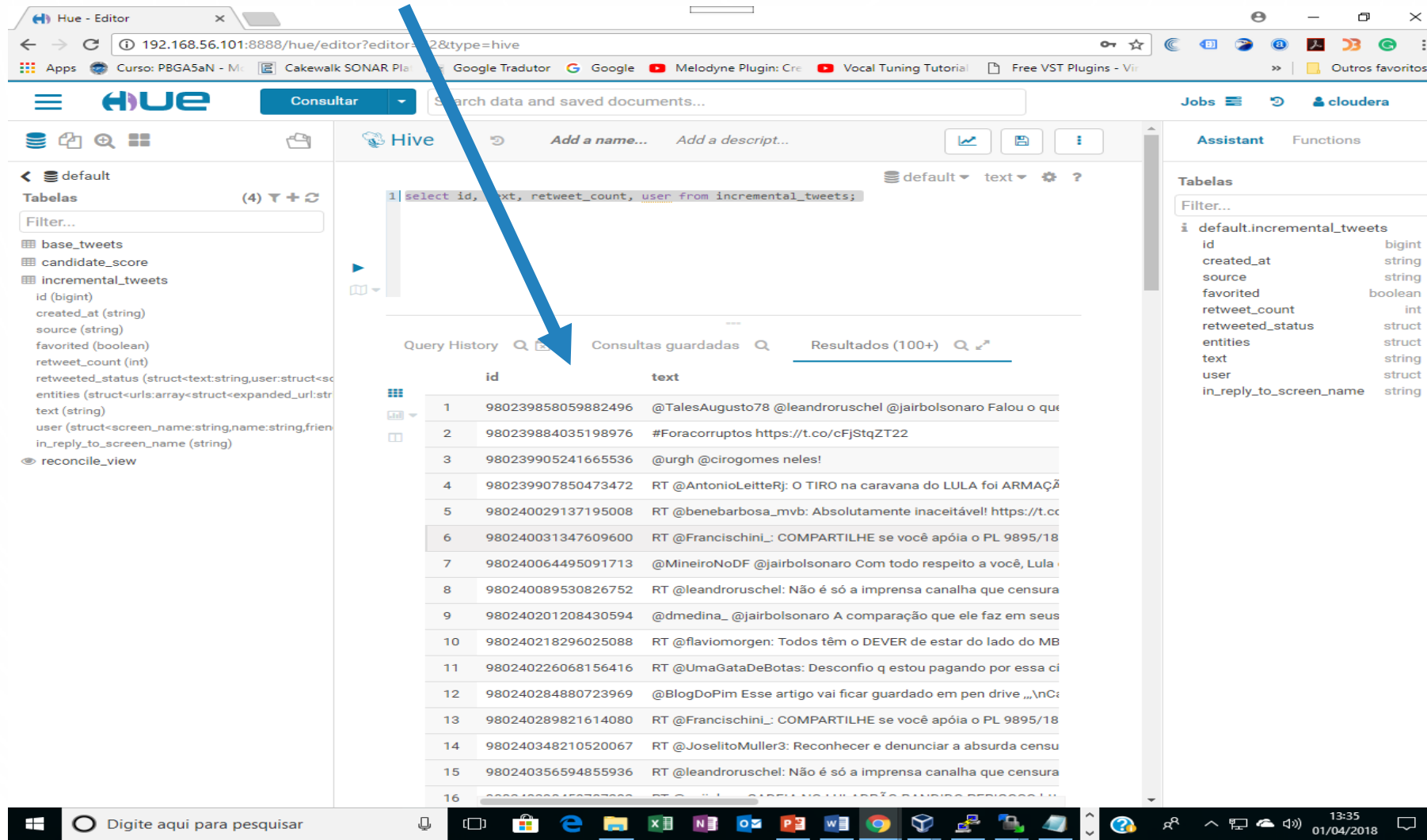
CONSULTA

- É POSSÍVEL TAMBÉM FAZER UMA CONSULTA À TABELA HIVE QUE FOI CRIADA ANTERIORMENTE COM O SCRIPT `create_twitter_schema.hql`.

```
select id, text, retweet_count, user from incremental_tweets;
```

RESULTADO CONSULTA

• RESULTADO DA CONSULTA HIVE



The screenshot shows the Hue web interface with a Hive query executed. The query is `select id, text, retweet_count, user from incremental_tweets;`. The result is displayed in a table with 16 rows and 3 columns: `id`, `text`, and `retweet_count`. The table is titled "Resultados (100+)".

	id	text	retweet_count
1	980239858059882496	@TalesAugusto78 @leandroruschel @jairbolsonaro Falou o que	
2	980239884035198976	#Foracorrupitos https://t.co/cFjStqZT22	
3	980239905241665536	@urgh @cirogomes neles!	
4	980239907850473472	RT @AntonioLeitteRj: O TIRO na caravana do LULA foi ARMAÇÃ	
5	980240029137195008	RT @benebarbosa_mvb: Absolutamente inaceitável! https://t.co	
6	980240031347609600	RT @FrancischiniL: COMPARTILHE se você apóia o PL 9895/18	
7	980240064495091713	@MineiroNoDF @jairbolsonaro Com todo respeito a você, Lula	
8	980240089530826752	RT @leandroruschel: Não é só a imprensa canalha que censura	
9	980240201208430594	@dmedina_ @jairbolsonaro A comparação que ele faz em seus	
10	980240218296025088	RT @flaviomorgen: Todos têm o DEVER de estar do lado do MB	
11	980240226068156416	RT @UmaGataDeBotas: Desconfio q estou pagando por essa ci	
12	980240284880723969	@BlogDoPim Esse artigo vai ficar guardado em pen drive „\nC	
13	980240289821614080	RT @FrancischiniL: COMPARTILHE se você apóia o PL 9895/18	
14	980240348210520067	RT @JoselitoMuller3: Reconhecer e denunciar a absurda censu	
15	980240356594855936	RT @leandroruschel: Não é só a imprensa canalha que censura	
16	9802403615037300	RT @Lula: NÃO É SÓ A IMPRENSA CANALHA QUE CENSURA	

ANÁLISE

- PARA A ANÁLISE FOI UTILIZADO O SPARK, UMA TECNOLOGIA DE COMPUTAÇÃO DISTRIBUÍDA EM CLUSTER EXTREMAMENTE RÁPIDA.
- PARA QUE O SPARK ENXERGUE AS TABELAS HIVE RODE OS COMANDOS:

```
rm -r /etc/spark/conf/hive.xml
```

```
sudo ln -s /etc/hive/conf/hive-site.xml /etc/spark/conf/hive-site.xml
```

ANÁLISE

- ALGUNS ARQUIVOS NECESSÁRIOS AO FUNCIONAMENTO:
 - AFINN (FINN ÅRUP NIELSEN) - UMA LISTA DE PALAVRAS EM INGLÊS CLASSIFICADAS POR VALOR NUMÉRICO INTEIRO ENTRE MENOS CINCO (NEGATIVO) E MAIS CINCO (POSITIVO). O ARQUIVO FOI ADAPTADO PARA O PORTUGUÊS PARA REALIZAÇÃO DESSE EXPERIMENTO
 - CANDIDATE MAPPING.TXT
- LINK PARA DOWNLOAD DOS ARQUIVOS:

<https://github.com/toticavalcanti/projeto-flume-twitter-spark/tree/master/>

ANÁLISE

- PARA SIMPLIFICAR, COMO DETERMINAMOS SE UM TWEET DEVE SER ATRIBUÍDO A UM CANDIDATO / ASSUNTO É FAZENDO REFERÊNCIA AO MANIPULADOR CANDIDATO / ASSUNTO. POR EXEMPLO:

TWEET #1

@CIROGOMES CONCEDEU A UMA ENTREVISTA...

TWEET #2

@CIROGOMES TRAVA GRANDE LUTA....

TWEET #3

@JOÃODORIA DIZ QUE SÃO PAULO É...

NO EXEMPLO ACIMA, O SENTIMENTO PELOS TWEETS #1 E #2 SERÁ ATRIBUÍDO A CIRO GOMES, ENQUANTO O TWEET #3 SERÁ ATRIBUÍDO A JOÃO DORIA.

PASSOS DA ANÁLISE

- ETAPA 1: CRIA UM MAPEAMENTO SIMPLES PARA ROTULAR O NOME DO TWEET. COMO CANDIDATOS DIFERENTES SERÃO REFERENCIADOS NO TWEET DE MANEIRA DIFERENTE, É PRECISO MAPEAR CADA NOME DE CANDIDATO PARA OS DIFERENTES NOMES PELOS QUAIS ELES SÃO REFERIDOS.
- ETAPA 2: CRIA UM DICIONÁRIO DE PALAVRAS DE SENTIMENTO E SUAS PONTUAÇÕES ASSOCIADAS. ISSO SERÁ USADO PARA CALCULAR A PONTUAÇÃO GERAL DO SENTIMENTO DO TWEET.
- PASSO 3: PARA CADA TWEET, CALCULA A PONTUAÇÃO DO SENTIMENTO E SOMA A PONTUAÇÃO DE CADA CANDIDATO.

PASSOS DA ANÁLISE

A COMPUTAÇÃO PRINCIPAL OCORRE NOS TRECHOS DE CÓDIGO MOSTRADOS A SEGUIR:

```
sentimentTuple = tweets.rdd.map(lambda r: [r.id, r.text, r.name]) \  
    .map(lambda r: [sentiment(r[1]),r[2]]) \  
    .flatMapValues(lambda x: x) \  
    .map(lambda y: (y[1],y[0])) \  
    .reduceByKey(lambda x, y: x+y) \  
    .sortByKey(ascending=True)
```

```
scoreDF = sentimentTuple.join(candidates) \  
    .map(lambda (x, y): (y[1],y[0])) \  
    .reduceByKey(lambda a, b: a + b) \  
    .toDF()
```

PASSOS DA ANÁLISE

O RESULTADO DE:

```
.map(lambda r: [sentiment(r[1]),r[2]]) \  
[1.0012610959381487, [u'Jaime Soares']],  
[-13.599376521158035], [u'Jair Bolsonaro']],  
[-0.47868277536822768], [u'Jairo Jorge', u'Janaina Paschoal']]  
...
```

PASSOS DA ANÁLISE

O RESULTADO DE:

```
.flatMapValues(lambda x: x) \:
```

```
(0.10817623727073111, u'Luiz Fernando Velho')
```

```
(0.0, u'Luiz Muller')
```

```
(0.0, u'Luiza Beatriz')
```

```
(-1.1125864642379386, u'Lula pelo Brasil')
```

```
(1.7582520441032219, u'Lula2018. F\~emDeus')
```

```
...
```

PASSOS DA ANÁLISE

O RESULTADO DE:

```
.map(lambda y: (y[1],y[0])) \
```

```
(0.10817623727073111, u'Luiz Fernando Velho')
```

```
(u'Luiz Fernando Velho', 0.10817623727073111)
```

```
(u'Luiz Muller', 0.0)
```

```
(u'Luiza Beatriz', 0.0)
```

```
(u'Lula pelo Brasil', -1.1125864642379386)
```

```
(u'Lula2018. F\xe9emDeus', 1.7582520441032219)
```

```
...
```


PASSOS DA ANÁLISE

`reduceByKey` e `sortBykey` RESULTA EM:

(u'#1deAbril #DiaDoLula', -0.10222859593214292)

(u'#BolsonaroPresidente\U0001f1e7\U0001f1f7', -0.71155785495028201)

(u'#Givanildo \U0001f1e7\U0001f1f7', -1.0878299881551272)

(u'#LulaliderdoPT', -0.42560277789377526)

(u'#SomostodosMoro', 0.26211121699831136)

(u'#VotoImpresso #Direita #Bolsonaro', 0.43759497449368367)

...

PASSOS DA ANÁLISE

ÚLTIMO PASSO, FAZER LEFT JOIN COM O DATAFRAME DO CANDIDATO

```
Row(_1=u'CiroGomes',_2=1.3068750447223945)
```

```
Row(_1=u'Álvaro Dias',_2=0.20299071319474044)
```

```
Row(_1=u'JairBolsonaro',_2=-13.599376521158035)
```

PARA QUE O SPARK ENXERGUE AS TABELAS HIVE, DIGITE O COMANDO:

```
rm -r /etc/spark/conf/hive.xml
```

DEPOIS

```
sudo ln -s /etc/hive/conf/hive-site.xml /etc/spark/conf/hive-site.xml
```

PASSOS DA ANÁLISE

AGORA ENTRE NA PASTA ONDE O ARQUIVO SE ENCONTRA E DIGITE:

```
spark-submit SentimentAnalysis.py
```

SENTIMENTANALYSIS.PY PREENCHE A TABELA CANDIDATE_SCORE

ENTRE NO HIVE E FAÇA UMA CONSULTA A TABELA:

```
select * from candidate_score;
```

PASSOS DA ANÁLISE

RESULTADO: 4) 


U./Is default text 



```
1 select * from candidate_score;
```

INFO : Executing command(queryId=hive_20180403165858_3b8d707b-1436-4ac4-9306-f881b4491f49):
select * from candidate_score
INFO : Completed executing command(queryId=hive_20180403165858_3b8d707b-1436-4ac4-9306-f881b4491f49); Time taken: 0.001 seconds
INFO : OK

Query History  

Consultas guardadas 

Resultados (8)  

	candidate_score.candidate_name	candidate_score.sentiment_score
1	Jair Bolsonaro	200.95504893868048
2	Geraldo Alckmin	10.591474982280905
3	MarinaSilva	0.21442250696755896
4	Lula	0.83111787875596965
5	Ciro Gomes	2.1825637355220953
6	Alvaro Dias	0.42354956875818484
7	Rodrigo Maia	3.3498944246648508
8	Guilherme Boulos	-3.6000313083408551

Tabela:

Filter..

defa
can
sen

12. CONCLUSÕES FINAIS

- POR UMA QUESTÃO DE IMPARCIALIDADE, OS DADOS APRESENTADOS A RESPEITO DOS CANDIDATOS, FORAM APENAS ILUSTRATIVOS, NÃO CONDIZENDO COM A REALIDADE.
- O FLUME É REALMENTE UMA FERMENTA IMPORTANTE DO ECO SISTEMA HADOOP E ESTUDÁ-LO APROFUNDOU NOSSO CONHECIMENTO EM HDFS, HIVE E UMA INTRODUÇÃO AO SPARK UTILIZANDO PYSPARK, PARA FAZER O PROCESSAMENTO DISTRIBUÍDO.

The image features a minimalist design with the letters 'FIM' centered on a white background. The corners are decorated with realistic water droplets of various sizes, some showing highlights and shadows to give them a three-dimensional appearance.

FIM

The image features a white background with several realistic water droplets of varying sizes in the corners. The top-left corner has a large droplet and two smaller ones. The top-right corner has a medium droplet and a small one. The bottom-right corner has a large, irregular droplet and several smaller ones. The bottom-center has a medium droplet and a small one. The word "OBRIGADO" is centered in the middle of the page.

OBRIGADO