

Tutoriel : Analyser les données Twitter avec Flume et Hive

21 avril 2015

Stéphane WALTER



Partager



L'objectif de ce tutoriel est de vous montrer comment utiliser [Flume](#) et [Hive](#) pour analyser des données en provenance de [Twitter](#).

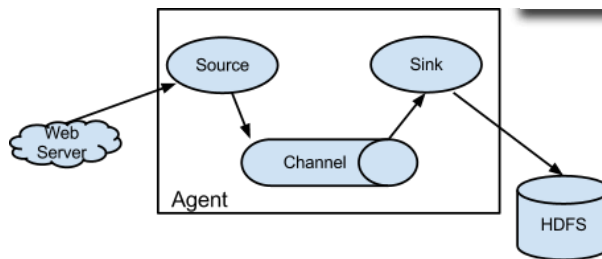
Il a également pour objectif de mettre en évidence les difficultés que l'on rencontre actuellement avec des plateformes Big Data en évolution rapide mais pas toujours stabilisées, d'où l'importance de disposer d'une expertise suffisante dans le domaine.

Ce tutoriel a été élaboré à partir de la version [sandbox 2.1 de la distribution Hortonworks](#).

Présentation de Flume

[Flume](#) a été initialement développé par [Cloudera](#) avant d'être reversé à la communauté Apache. Il porte maintenant l'appellation Flume NG (Next Génération). C'est un outil relativement simple faisant aujourd'hui parti de l'éco-système [Hadoop](#).

Flume fonctionne comme un service distribué pour assurer la collecte de données en temps réel, leur stockage temporaire et leur diffusion vers une cible.



Techniquement, un agent Flume permet de créer des routes pour relier une source à une cible via un canal d'échange.

La « source » Flume a pour but de récupérer les messages à partir de différentes sources, en particulier des fichiers de logs mais aussi comme nous le verrons des données Twitter.

Le « canal » Flume est une zone tampon qui permet de stocker les messages avant qu'ils soient consommés. On utilise généralement un stockage en mémoire.

Le « cible » Flume consomme par lot les messages en provenance du « canal » pour les écrire sur une destination comme HDFS par exemple.

Lorsque la vitesse d'intégration des nouveaux messages est plus rapide que celle d'écriture vers la cible, la taille du « canal » augmente afin de garantir qu'aucun message ne soit perdu.

Installation de Flume

On se connecte tout d'abord à la machine virtuelle à partir d'un terminal distant:

```
ssh root@127.0.0.1 -p 2222
```

Puis on installe le package via la commande:

```
yum install -y flume
```

```
swal4u — root@sandbox:~ — ssh — 80x15
bash root@sandbox:~
Last login: Thu Jan 8 12:45:14 on ttys000
MacBook-Pro-de-Stephane:~ swal4u$ ssh root@127.0.0.1 -p 2222
The authenticity of host '[127.0.0.1]:2222 ([127.0.0.1]:2222)' can't be established.
RSA key fingerprint is 19:50:65:c6:dc:1b:9a:db:15:6f:c0:a3:21:12:0d:9c.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '[127.0.0.1]:2222' (RSA) to the list of known hosts.
root@127.0.0.1's password:
Last login: Thu Jan 8 02:17:52 2015
[root@sandbox ~]# yum install -y flume
```

Données personnelles

Le package s'installe et à la fin, votre écran devrait ressembler à ceci.

```
swal4u -- root@sandbox:~ -- ssh -- 96x24
bash root@sandbox:~
Installing:
flume noarch 1.4.0.2.1.1.0-385.el6 HDP-2.1 64 M
Transaction Summary
=====
Install      1 Package(s)
Total download size: 64 M
Installed size: 71 M
Downloading Packages:
flume-1.4.0.2.1.1.0-385.el6.noarch.rpm | 64 MB 00:23
Running rpm_check_debug
Running Transaction Test
Transaction Test Succeeded
Running Transaction
Installing : flume-1.4.0.2.1.1.0-385.el6.noarch 1/1
Verifying : flume-1.4.0.2.1.1.0-385.el6.noarch 1/1
Installed:
flume.noarch 0:1.4.0.2.1.1.0-385.el6
Complete!
[root@sandbox ~]#
```

Configuration pour accéder à Twitter

Vous devez tout d'abord créer une application Twitter en allant sur le site des développeurs Twitter (<https://dev.twitter.com/apps/>) afin de générer vos identifiants.

swal4u

Details Settings Keys and Access Tokens Permissions

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	[REDACTED]
Consumer Secret (API Secret)	[REDACTED]
Access Level	Read-only (modify app permissions)
Owner	[REDACTED]
Owner ID	[REDACTED]

Avec ces identifiants, on va pouvoir mettre à jour le fichier de configuration flume.conf.

```
58 TwitterAgent.sources = Twitter
59
60 TwitterAgent.channels = MemChannel
61
62 TwitterAgent.sinks = HDFS
63
64 TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
65
66 TwitterAgent.sources.Twitter.channels = MemChannel
67
68 TwitterAgent.sources.Twitter.consumerKey = [REDACTED]
69
70 TwitterAgent.sources.Twitter.consumerSecret = X [REDACTED]
71
72 TwitterAgent.sources.Twitter.accessToken = [REDACTED]
73
74 TwitterAgent.sources.Twitter.accessTokenSecret [REDACTED]
75
76 TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientist
77
78
79 TwitterAgent.sinks.HDFS.channel = MemChannel
80
81 TwitterAgent.sinks.HDFS.type = hdfs
82
83 TwitterAgent.sinks.HDFS.hdfs.path = hdfs://sandbox.hortonworks.com:8020/user/hue/tweets
84
85 TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
86
87 TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
88
89 TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
90
91 TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
92
93 TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
94
95 TwitterAgent.channels.MemChannel.type = memory
96
97 TwitterAgent.channels.MemChannel.capacity = 10000
98
99 TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Outre les identifiants Twitter, on va mettre à jour le paramètre keywords pour filtrer les tweets sur les éléments qui nous intéressent (dans mon exemple, je m'intéresse aux tweets sur le big data).

Données personnelles

On remarque que tous les paramètres sont préfixés par le nom de l'agent, ici **TwitterAgent**.

On va également préciser l'emplacement HDFS où seront stockés les tweets récupérés.

```
TwitterAgent.sinks.HDFS.hdfs.path =  
hdfs://sandbox.hortonworks.com:8020/user/hue/twitter/tweets
```

Vous trouverez une information plus exhaustive du paramétrage sur <http://flume.apache.org>.

Si vous avez modifié le fichier sur votre machine, il ne reste plus qu'à le copier sur la VM.

```
scp -P 2222 flume.conf root@127.0.0.1:/etc/flume/conf/
```

Il faut maintenant télécharger [flume-sources-1.0-SNAPSHOT.jar](#). Ce package contient les fonctions nécessaires pour accéder et récupérer les données de Twitter.

On va également le copier sur la VM, dans le répertoire des fichiers jars de Flume.

```
scp -P 2222 flume-sources-1.0-SNAPSHOT.jar  
root@127.0.0.1:/usr/lib/flume/lib/
```

J'ai rencontré pas mal de soucis pour savoir où positionner ce fichier. :

Démarrer Flume

On démarre flume avec la commande suivante:

```
flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume.conf -n Twi
```

On précise le répertoire et le fichier de configuration de Flume. On indique également l'agent que l'on démarre, ici TwitterAgent. On redirige la sortie dans un fichier pour analyse si besoin.

Données personnelles

rint flume sh dans lequ

vos recherches les don



transmises dans les

ne peut pas les traiter nativement

es iar

nombreux lecteurs de

Données personnelles

Là encore, après de multiples recherches et tentatives, j'ai finalement trouvé une solution en récupérant la dernière version de ce fichier sur <https://github.com/rcongiu/Hive-JSON-Serde>. Je n'ai pas pris les sources mais directement le binaire suivant:

`http://www.congiu.net/hive-json-serde/1.3/cdh5/
json-serde-1.3-jar-with-dependencies.jar`

J'ai simplement copié ce fichier sur /root/.

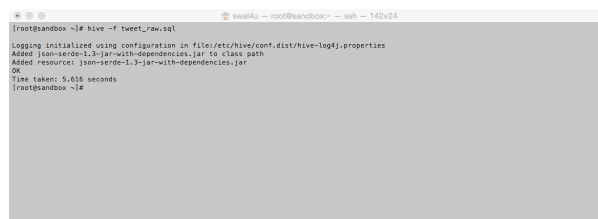
On va ensuite créer une table Hive. Une table externe car nous n'avons pas besoin de déplacer physiquement les données. Celles-ci resteront toujours sur HDFS et ne seront pas dupliquées.

On n'oublie pas au début du script de rajouter en début de script une commande pour prendre en compte la librairie java évoquée ci-dessus.

```
1 ADD JAR json-serde-1.3-jar-with-dependencies.jar;
2
3 --create the tweets_raw table containing the records as received from Twitter
4
5 CREATE EXTERNAL TABLE tweets_raw (
6   id BIGINT,
7   created_at STRING,
8   source STRING,
9   favorited BOOLEAN,
10  retweet_count INT,
11  retweeted_status STRUCT<
12    text:STRING,
13    user:STRUCT<screen_name:STRING,name:STRING>>,
14  entities STRUCT<
15    urls:ARRAY<STRUCT<expanded_url:STRING>>,
16    user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
17    hashtags:ARRAY<STRUCT<text:STRING>>>,
18  text STRING,
19  user STRUCT<
20    screen_name:STRING,
21    name:STRING,
22    friends_count:INT,
23    followers_count:INT,
24    statuses_count:INT,
25    verified:BOOLEAN,
26    utc_offset:STRING, -- was INT but nulls are strings
27    time_zone:STRING>,
28  in_reply_to_screen_name STRING,
29  year int,
30  month int,
31  day int,
32  hour int
33  )
34 ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
35 --ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
36 LOCATION '/user/hue/twitter/tweets'
37 ;
```

Il suffit ensuite de lancer le fichier via la commande suivante:

`hive -f tweet_raw.sql`



```
[root@sandbox ~]# hive -f tweet_raw.sql
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-logging.properties
Added resource: json-serde-1.3-jar-with-dependencies.jar
OK
Time taken: 5.616 seconds
[root@sandbox ~]#
```

Query Editor
My Queries
Saved C

DATABASE

default

SETTINGS

Add

FILE RESOURCES

Type

jar

x

Path

/apps/hive/json-se

..

Add

USER-DEFINED FUNCTIONS

Add

PARAMETERIZATION

☒ Enable Parameterization

EMAIL NOTIFICATION

☐ Email me on completion

Si l'on veut utiliser l'interface web Hue pour visualiser la table, il faut au préalable copier le fichier jar json sur /apps/hive puis le charger dans l'interface en rajoutant ce fichier jar comme un fichier de ressource dans l'onglet Query Editor (Beewax).

On peut ensuite visualiser la table dans l'interface Hue directement.

Query Results: tweets_raw

DOWNLOADS

Download as CSV
Download as XLS
☐ Enable visualization
Save

Did you know? If the result contains a large number of columns, click a row to select a column to jump to. As you type into the field, a drop-down list displays column names that match the string.

	Results	Query	Log	Columns
	default.tweets_raw.id	default.tweets_raw.created_at	default.tweets_raw.source	
0	553242207009116180	Thu Jan 08 17:29:55 +0000 2015	<a href="http://twitter.com/download/iphone" rel="nofollow"	
1	553242219940151296	Thu Jan 08 17:29:58 +0000 2015	Paper.li	
2	553242224540929853	Thu Jan 08 17:29:59 +0000 2015	LinkedIn	
3	553242235312308225	Thu Jan 08 17:30:02 +0000 2015	Sprinklr	
4	553242241935097857	Thu Jan 08 17:30:03 +0000 2015	Hootsuit	
5	55324224455832327169	Thu Jan 08 17:30:04 +0000 2015	TWDT	
6	553242245588320256	Thu Jan 08 17:30:04 +0000 2015	JFTTT	
7	553242246263983808	Thu Jan 08 17:30:04 +0000 2015	Java Retweet bot	
8	553242247035387904	Thu Jan 08 17:30:05 +0000 2015	Twitter Web Clie	
9	553242248629211137	Thu Jan 08 17:30:05 +0000 2015	Hootsuit	
10	553242254610276352	Thu Jan 08 17:30:06 +0000 2015	Murph	
11	553242254622867456	Thu Jan 08 17:30:06 +0000 2015	Buffer	

On peut évidemment faire des requêtes.

Pour illustrer ce point, je vais classer les utilisateurs de tweets par ordre décroissant en fonction du nombre de followers. A noter l'accès au champ followers_count qui n'est pas un champs simple de la table mais un champs in Données personnelles pure.

```
select user.screen_name, user.followers_count c from tweets_raw
order by c desc
```

Et voici le résultat:

Query Results: Followers

DOWNLOADS

Download as CSV

Download as XLS

☐ Enable visualization

Save

MIN JOB ID:

1420712212014_0010

	screen_name	c
0	Slate	1256498
1	AnnTran	439254
2	NancyGraceHLN	436483
3	socialmedia2day	406127
4	hphelioncloud	281744
5	FoxBusiness	230441
6	overnightprints	179166
7	overnightprints	179166
8	TwitterOSS	178067
9	Computerworld	156102
10	ClementCharles	140554
11	EricTTung	135195

Next Page →

Merci de m'avoir suivi et a bientôt pour un tutoriel sur la visualisation des données analysées de Twitter avec [Qlik](#).



STÉPHANE WALTER
Manager Conseil & Expertise Big Data
Business & Decision



17 ans d'expérience dans la mise en place d'architectures dédiées à la valorisation de vos données grâce aux technologies BI et Big Data.

[En savoir plus](#)

COMMENTAIRES (5)

Stéphane Raynal *Le 16 août 2016 à 17h16*

Bonjour Stéphane,

Une info pour ceux qui voudraient essayer de lire des flux tweeter via Cloudera.

Dans la dernière version de Cloudera avec Flume intégré, il n'est pas nécessaire d'importer le "flume-sources-1.0-SNAPSHOT.jar".
Ce jar, ou son équivalent mis à jour, est déjà dans le \$FLUME_HOME/lib.

Par contre il est nécessaire de changer :

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

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

Une autre note utile, pour démarrer Flume en mode debug il faut
rajouter la ligne suivante lors du lancement de flume-ng :
-Dflume.root.logger=DEBUG,console

Stéphane WALTER *Le 01 août 2016 à 10h08*

Bonjour Mélissa,

Non ce n'est absolument pas normal.

La seule explication à te donner serait que ces tweets ont été générés de manière automatique par un logiciel. Le contenu serait identique donnant l'illusion que les tweets sont les mêmes. Mais ils devraient avoir chacun un identifiant distinct.

Bon courage pour ton projet.

Foloxflou *Le 27 juillet 2016 à 15h08*

Bonjour,

Après avoir suivi votre tutoriel (merci d'ailleurs), je m'aperçois que les multiples fichiers inscrits dans le HDFS contiennent les mêmes données (ie les mêmes tweets) lorsque je lance un agent flume.

Est ce normal ?

Données personnelles

Je dois à chaque fois relancer l'agent si je veux que les fichiers contiennent des tweets "nouveaux" ...

Stéphane WALTER *Le 30 avril 2015 à 11h44*

Bonjour Pierre-Henri,

Effectivement, les distributions évoluent très rapidement et on constate des différences de comportement dès qu'on change de version. C'est l'une des difficultés de ce type de tutoriel. Merci beaucoup pour votre complément.

PH Brunelle *Le 30 avril 2015 à 11h37*

Bonjour,

Merci pour cet excellent tutoriel. J'ai suivi vos instructions en utilisant la version 2.2 de la sandbox Hortonworks, et il semble qu'il y ait quelques différences.

Le répertoire de flume semble désormais se situer dans `"/usr/hdp/2.2.0.0-2041/flume"`. Malgré cette modification, l'exécution échoue avec des message de type `"java.lang.NoSuchMethodError: twitter4j.conf.Configuration.getRequestHeaders()Ljava/util/Map;"`. Cela vient d'un problème de version de twitter4j dans le repertoire flume/lib. En remplaçant la librairie existante par `twitter4j-core-3.0.6-SNAPSHOT-sources.jar`, cela fonctionne parfaitement.

Bonne continuation.

LAISSEZ UN COMMENTAIRE

Commentaire *

Nom *

S'affichera sur le site

Adresse e-mail *

LAISSER UN COMMENTAIRE

Votre adresse de messagerie est uniquement utilisée par Business & Decision, responsable de traitement, aux fins de traitement de votre demande et d'envoi de toute communication de Business & Decision en relation avec votre demande uniquement. [En savoir plus](#) sur la gestion de vos données et vos droits.

DÉCOUVREZ AUSSI...

TRANSFORMATION DIGITALE



Comment la transformation digitale peut aider les organismes associatifs à changer le monde ?



INTÉGRER L'IA ET LA DATA SCIENCE



Intelligence Artificielle et Data Science avec Python NumPy et Pandas [REPLAY #10]

Données personnelles



Revue de presse Data & Digital - Octobre 2019

★
PREMIUM



GDPR : en route vers la conformité

Le règlement 679/2016 appelé GDPR est donc un sujet majeur et implique de l'apprécier sous de multiples angles : juridique, IT, usages, changement...

TÉLÉCHARGER LE LIVRE BLANC

NEWSLETTER

Recevoir notre actualité par mail

E-mail *

OK

GARDONS LE CONTACT

Données personnelles

Retrouvez-nous sur les réseaux sociaux



ÉVÉNEMENTS



NEWSLETTER



CONTACTEZ-NOUS

le blog

Mentions légales

© Copyright - 2019
> www.businessdecision.com

Données personnelles