Big Data Analytics

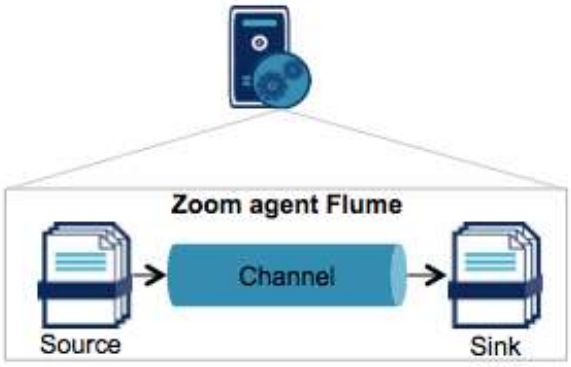
Workshop Apache flume & Hive

I. Introduction :

Dans ce workshop, nous allons voir comment importer des données non structurées dans hdfs avec flume.

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

Pour utiliser Flume, on doit lancer un agent Flume qui est un processus JVM qui héberge les composants via lesquels les événements circulent d'une source externe vers la destination.



1. Source :

Principaux types de sources :

ExecSource : commande bash simple,

AvroSource : écouter un port TCP et recevoir des logs au format Avro,

SyslogSource : router les logs d'un serveur syslog vers Flume,

SpoolingDirectorySource : récupérer le contenu des fichiers de log dans un répertoire,

2. Channel :

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

Les différents types de channels sont :

FileChannel : persister les logs sur le système de fichier pour garantir la non-perte de message en cas de panne et/ou redémarrage de l'agent

MemoryChannel : garder les logs en mémoire pour favoriser la performance

JDBCChannel : utiliser une base de données comme solution de stockage

3. **Sink :**

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

Les différents types de cible sont :

HDFSSink : écrire dans HDFS

HBaseSink et ElasticSearchSink, base NoSQL (Cassandra, MongoDB, etc.)

AvroSink : rediriger les logs au format Avro sur un port TCP distant,

FileRollSink : écrire dans le filesystem local

II. Scenario :

1. Dans ce premier scenario nous allons importer des données à partir d'un fichier **user-posts.txt** sous hdfs dans le répertoire **/user/cloudera/destinationLog**.

Le paramétrage de l'import sera effectué dans un fichier de configuration.

Dans ce qui suit, nous allons créer l'agent d'import flume **agent1** :

2. Créer le répertoire **destinationLog** sous hdfs.
3. Créer un fichier nommé **myconfig.conf** dans le dossier conf de flume avec :

```
gedit /usr/lib/flume-ng/conf
```

4. Ajouter les lignes suivantes :

```
agent1.sources=tail-source
agent1.sinks=hdfs-sink
agent1.channels=memory-channel

#source configure
agent1.sources.tail-source.type=exec
agent1.sources.tail-source.command=tail -F /home/cloudera/user-posts.txt
agent1.sources.tail-source.channels=memory-channel

#hdfs sink configure
agent1.sinks.hdfs-sink.channel=memory-channel
agent1.sinks.hdfs-sink.type=hdfs
agent1.sinks.hdfs-sink.hdfs.path=hdfs:///user/cloudera/destinationLog
agent1.sinks.hdfs-sink.hdfs.fileType=DataStream

# chanel configure
agent1.channels.memory-channel.type=memory
agent.channels.memory-channel.capacity = 1000
agent.channels.memory-channel.transactionCapacity = 100
```

5. Démarrer l'agent créer avec :

```
flume-ng agent -n agent1 -f //usr/lib/flume-ng/conf/myconfig.conf
```

6. Ajouter les lignes suivantes dans le fichier user-posts.txt

```
User3,good work,1343182133839
User2,love it,1343182154633
```

7. Vérifier le contenu du répertoire /user/cloudera/destinationLog