Flume-1.4.0 和 Hbase-0.96.0 整合

最近由于项目需要把 Flume 收集到的日志信息插入到 Hbase 中，由于第一次接触这些，在整合的过程中，我遇到了许多问题，我相信很多人也应该会遇到这些问题的，于是我把整个整合的过程写出来，希望给那些同样遇到这样问题的朋友帮助。

在使用 Flume 的时候，请确保你电脑里面已经搭建好 Hadoop、Hbase、Zookeeper 以及 Flume。本文将以最新版的 Hadoop-2.2.0、Hbase-0.96.0、Zookeeper-3.4.5 以及 Flume-1.4.0 为例进行说明。如何安装分布式的 Hadoop、Hbase、Zookeeper 请参见本论坛的《Hadoop2.2.0 完全分布式集群平台安装与设置》、《Hbase 0.96.0 分布式安装手册》、《Zookeeper 3.4.5 分布式安装手册》；如何安装分布式 Flume 本博客将在以后的文章中介绍。

1、本程序一共用了三台集群搭建集群，这三台机器的 Hostname 分别为 master、node1、node2；master 机器是 Hadoop 以及 Hbase 集群的 master。三台机器上分别启动的进程如下：

```
[wyp@master ~]$ jps
2973 HRegionServer
4083 Jps
2145 DataNode
3496 HMaster
2275 NodeManager
1740 NameNode
2790 QuorumPeerMain
1895 ResourceManager

[wyp@node1 ~]$ jps
7801 QuorumPeerMain
11669 DataNode
29419 Jps
11782 NodeManager
29092 HRegionServer

[wyp@node2 ~]$ jps
2310 DataNode
2726 HRegionServer
2622 QuorumPeerMain
3104 Jps
2437 NodeManager
```

2、以 master 机器作为 flume 数据的源、并将数据发送给 node1 机器上的 flume，最后 node1 机器上的 flume 将数据插入到 Hbase 中。master 机器上的 flume 和 node

1 机器上的 flume 中分别做如下的配置：
在 master 的$FLUME_HOME/conf/目录下创建以下文件（文件名随便取），并做如下配置，这是数据的发送端：

```
01 [wyp@master conf]$ vim example.conf
02 agent.sources = baksrc
03 agent.channels = memoryChannel
04 agent.sinks = remotesink
05
06 agent.sources.baksrc.type = exec
07 agent.sources.baksrc.command = tail -F /home/wyp/Documents/data/data.txt
08 agent.sources.baksrc.checkperiodic = 1000
09
10 agent.channels.memoryChannel.type = memory
11 agent.channels.memoryChannel.keep-alive = 30
12 agent.channels.memoryChannel.capacity = 10000
13 agent.channels.memoryChannel.transactionCapacity = 10000
14
15 agent.sinks.remotesink.type = avro
16 agent.sinks.remotesink.hostname = node1
17 agent.sinks.remotesink.port = 23004
18 agent.sinks.remotesink.channel = memoryChannel
```

在 node1 的$FLUME_HOME/conf/目录下创建以下文件（文件名随便取），并做如下配置，这是数据的接收端：

```
01 [wyp@node1 conf]$ vim example.conf
02 agent.sources = avrosrc
03 agent.channels = memoryChannel
04 agent.sinks = fileSink
05
06 agent.sources.avrosrc.type = avro
07 agent.sources.avrosrc.bind = node1
08 agent.sources.avrosrc.port = 23004
09 agent.sources.avrosrc.channels = memoryChannel
10
```

```
11 agent.channels.memoryChannel.type = memory
12 agent.channels.memoryChannel.keep-alive = 30
13 agent.channels.memoryChannel.capacity = 10000
14 agent.channels.memoryChannel.transactionCapacity =10000
15
16 agent.sinks.fileSink.type = hbase
17 agent.sinks.fileSink.table = wyp
18 agent.sinks.fileSink.columnFamily = cf
19 agent.sinks.fileSink.column = charges
20 agent.sinks.fileSink.serializer =
21 org.apache.flume.sink.hbase.RegexHbaseEventSerializer
22 agent.sinks.fileSink.channel = memoryChannel
```

这两个文件配置的含义我就不介绍了，自己 google 一下吧。

3、在 master 机器和 node1 机器上分别启动 flume 服务进程：

```
01 [wyp@master apache-flume-1.4.0-bin]$ bin/flume-ng agent
02 --conf conf
03 --conf-file conf/example.conf
04 --name agent
05 -Dflume.root.logger=INFO,console
06
07 [wyp@node1 apache-flume-1.4.0-bin]$ bin/flume-ng agent
08 --conf conf
09 --conf-file conf/example.conf
10 --name agent
11 -Dflume.root.logger=INFO,console
```

当分别在 node1 和 master 机器上启动上面的进程之后，在 node1 机器上将会输出以下的信息：

```
01 2014-01-20 22:41:56,179 (pool-3-thread-1)
02 [INFO - org.apache.avro.ipc.NettyServer$NettyServerAvroHandler.
03 handleUpstream(NettyServer.java:171)]
04 [id: 0x16c775c5, /192.168.142.161:42201 => /192.168.142.162:23004] OPEN
05 2014-01-20 22:41:56,182 (pool-4-thread-1)
```

```
06 [INFO - org.apache.avro.ipc.NettyServer$NettyServerAvroHandler.
07 handleUpstream(NettyServer.java:171)]
08 [id: 0x16c775c5, /192.168.142.161:42201 => /192.168.142.162:23004]
09 BOUND: /192.168.142.162:23004
10 2014-01-20 22:41:56,182 (pool-4-thread-1)
11 [INFO - org.apache.avro.ipc.NettyServer$NettyServerAvroHandler.
12 handleUpstream(NettyServer.java:171)]
13 [id: 0x16c775c5, /192.168.142.161:42201 => /192.168.142.162:23004]
14 CONNECTED: /192.168.142.161:42201
```

在 master 机器上将会输出以下的信息：

```
01 2014-01-20 22:42:16,625 (lifecycleSupervisor-1-0)
02 [INFO - org.apache.flume.sink.AbstractRpcSink.
03 createConnection(AbstractRpcSink.java:205)]
04 Rpc sink remotesink: Building RpcClient with hostname: node1, port: 23004
05 2014-01-20 22:42:16,625 (lifecycleSupervisor-1-0)

06  [INFO - org.apache.flume.sink.AvroSink.initializeRpcClient(AvroSink.java:126)]

07 Attempting to create Avro Rpc client.
08 2014-01-20 22:42:19,639 (lifecycleSupervisor-1-0)

09  [INFO - org.apache.flume.sink.AbstractRpcSink.start(AbstractRpcSink.java:300)]

10 Rpc sink remotesink started.
```

这样暗示 node1 上的 flume 和 master 上的 flume 已经连接成功了。

4、如何测试？可以写一个脚本往/home/wyp/Documents/data/data.txt（见上面 master 机器上 flume 上面的配置）文件中追加东西：云凡教育大数据学院 www.cloudyhadoop.com

```
1 for i in {1..1000000}; do
2 echo "test flume to Hbase $i" >>
3 /home/wyp/Documents/data/data.txt;
4 sleep 0.1;
5 done
```

运行上面的脚本，这样将每隔 0.1 秒往/home/wyp/Documents/data/data.txt 文件中添加内容，这样 master 上的 flume 将会接收到 /home/wyp/Documents/data/data.txt 文件内容的变化，并变化的内容发送到 node1 机器上的 flume，node1 机器上的 flume 把接收到的内容插入到 Hbase 的 wyp 表中的 cf:charges 列中（见上面的配置）。

本文是以最新版的 Flume 和最新办的 Hbase 进行整合，在整合的过程中将会出现 flume 依赖包版本问题，解决方法是用 $HADOOP_HOME/share/hadoop/common/lib/guava-11.0.2.jar 替换 $FLUME_HOME/lib/guava-10.0.1.jar 包；
用$HADOOP_HOME/share/hadoop/common/lib/protobuf-java-2.5.0.jar 替换 $HBASE_HOME/lib/protobuf-java-2.4.0.jar 包。然后再启动步骤三的两个进程。