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请参见本博客的《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node1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：

```
[wyp@master conf]$ vim example.conf
agent.sources = baksrc
agent.channels = memoryChannel
agent.sinks = remotesink

agent.sources.baksrc.type = exec
agent.sources.baksrc.command = tail -F /home/wyp/Documents/data/data.txt
agent.sources.baksrc.checkperiodic = 1000

agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.keep-alive = 30
agent.channels.memoryChannel.capacity = 10000
agent.channels.memoryChannel.transactionCapacity = 10000

agent.sinks.remotesink.type = avro
agent.sinks.remotesink.hostname = node1
agent.sinks.remotesink.port = 23004
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：

```
[wyp@node1 conf]$ vim example.conf
agent.sources = avrosrc
agent.channels = memoryChannel
agent.sinks = fileSink

agent.sources.avrosrc.type = avro
agent.sources.avrosrc.bind = node1
agent.sources.avrosrc.port = 23004
agent.sources.avrosrc.channels = memoryChannel

agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.keep-alive = 30
agent.channels.memoryChannel.capacity = 10000
agent.channels.memoryChannel.transactionCapacity =10000
```

```
agent.sinks.fileSink.type = hbase
agent.sinks.fileSink.table = wyp
agent.sinks.fileSink.columnFamily = cf
agent.sinks.fileSink.column = charges
agent.sinks.fileSink.serializer =
        org.apache.flume.sink.hbase.RegexHbaseEventSerializer
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：

```
[wyp@master apache-flume-1.4.0-bin]$ bin/flume-ng agent
                        --conf conf
                        --conf-file conf/example.conf
                        --name agent
                        -Dflume.root.logger=INFO,console

[wyp@node1 apache-flume-1.4.0-bin]$ bin/flume-ng agent
                        --conf conf
                        --conf-file conf/example.conf
                        --name agent
                        -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：

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

在master机器上将会输出以下的信息：

```
2014-01-20 22:42:16,625 (lifecycleSupervisor-1-0)
[INFO - org.apache.flume.sink.AbstractRpcSink.
createConnection(AbstractRpcSink.java:205)]
Rpc sink remotesink: Building RpcClient with hostname: node1, port: 23004
2014-01-20 22:42:16,625 (lifecycleSupervisor-1-0)
[INFO - org.apache.flume.sink.AvroSink.initializeRpcClient(AvroSink.java:126)]
Attempting to create Avro Rpc client.
2014-01-20 22:42:19,639 (lifecycleSupervisor-1-0)
[INFO - org.apache.flume.sink.AbstractRpcSink.start(AbstractRpcSink.java:300)]
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：

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