# 发送消息至kafka

目录

[一、说明 2](#_Toc452112424)

[二、配置JVM 2](#_Toc452112425)

[1、配置FLUME-JVM 2](#_Toc452112426)

[三、配置CONF文件 3](#_Toc452112427)

[1、新建并配置zhouwh.conf(一份数据源发送到kafka集群的 一个topic): 3](#_Toc452112428)

[2、新建并配置zhouwh.conf(一份数据源发送到kafka集群的 两个topic): 4](#_Toc452112429)

[3、新建并配置zhouwh\_http.conf 5](#_Toc452112430)

[3.1、配置内容 6](#_Toc452112431)

[3.2、自定义HTTPSourceHandler 7](#_Toc452112432)

[四、启动FLUME 7](#_Toc452112433)

[1、默认日志级别启动： 7](#_Toc452112434)

[2、推荐使用-----修改日志级别为ERROR启动（提升发送消息速度）： 8](#_Toc452112435)

[五、测试post 8](#_Toc452112436)

[5.1、shell模拟post消息 8](#_Toc452112437)

[六、运行工程发送本地消息至Kafka 9](#_Toc452112438)

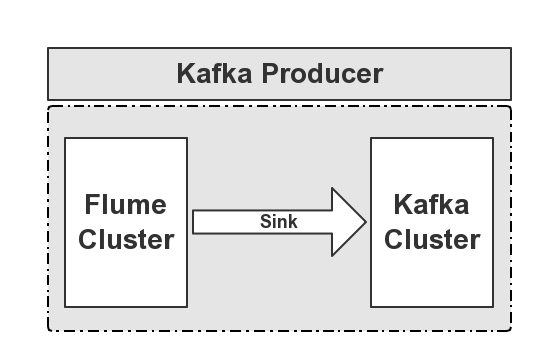
# 一、说明

本文档用于运行flume集群，读取本地目录下文件并发送消息到kafka集群、接收http

的post消息发送到kafka集群的配置介绍和运行本地工程发送消息至kafka集群。

（本文在172.16.71.63的CDH 5.5.2中flume、kafka集群测试，运行自己配置的zhouwh.conf、zhouwh\_http.conf，其中为了满足目前需求，使用zhouwh\_http.conf需要自定义一个HTTPSourceHandler并打包放入服务器）

下图表达了从Flume到Kafka的输送工程：



我们使用Flume作为日志收集系统，将收集到的数据输送到Kafka中间件，以供Storm、spark等去实时消费计算，整个流程从各个节点上，通过Flume的Agent代理收集日志，然后汇总到Flume集群，在由Flume的Sink将日志输送到Kafka集群，完成数据的生产流程。

# 二、配置JVM

## 1、配置FLUME-JVM

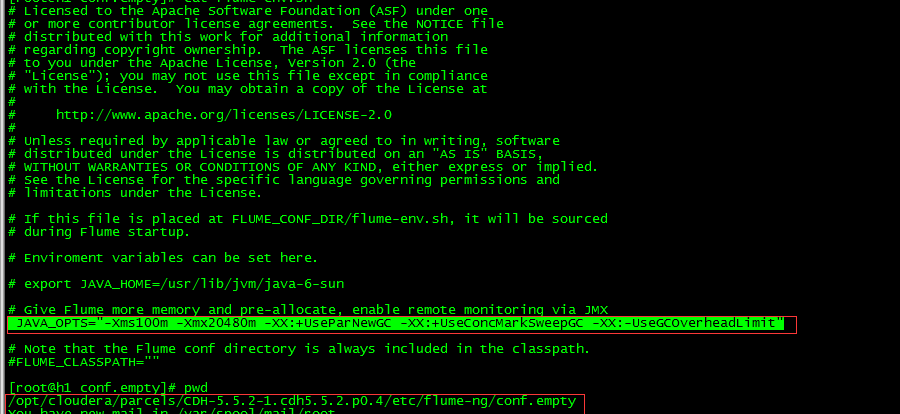
配置文件路径：

/opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/flume-ng/conf



配置flume-env.sh中的JVM参数，避免内存溢出：

添加 JAVA\_OPTS="-Xms100m -Xmx20480m -XX:+UseParNewGC -XX:+UseConcMarkSweepGC -XX:-UseGCOverheadLimit"



ps: 在启动 agent 的时候一定要带上 -c conf 选项，否则 flume-env.sh 里配置的环境变量不

会被加载生效

# 三、配置CONF文件

## 1、新建并配置zhouwh.conf(一份数据源发送到kafka集群的 一个topic):

CONF文件路径地址：

/opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/flume-ng/conf



作用：

flume读取本地目录下文件并发送消息到kafka集群的一个topic中

配置如下：

a1.sources=r1 （Source从Client收集数据，传递给Channel）

a1.sinks=k1 （Sink从Channel收集数据，运行在一个独立线程）

a1.channels=c1 （Channel连接 sources 和 sinks ，这个有点像一个队列）

a1.sources.r1.channels=c1 （固定配置，配置sources的channel）

a1.sources.r1.type=spooldir （定义源数据获取类型）

a1.sources.r1.spoolDir=/home/zhouwh （定义源数据目录）

a1.sinks.k1.type=org.apache.flume.plugins.KafkaSink （发送数据类型）

a1.sinks.k1.metadata.broker.list=h1.hadoop:9092,h2.hadoop:9092,h3.hadoop:9092,h4.hadoop:9092 （kafka集群地址，集群中输入"cat /etc/hosts"可查询）

a1.sinks.k1.serializer.class=kafka.serializer.StringEncoder （序列化处理类）

a1.sinks.k1.request.required.acks=1 （producer接收消息ack的时机.默认为0 ，配置1意味着在leader replica已经接收到数据后，producer会得到一个ack，这个选项提供了更好的持久性）

a1.sinks.k1.max.message.size=1000000 （最大消息）

a1.sinks.k1.producer.type=async （决定消息是否应在一个后台线程异步发送。合法的值为async，表示异步发送;sync表示同步发送）

a1.sinks.k1.custom.encoding=UTF-8 （编码格式）

a1.sinks.k1.custom.topic.name=zhouwh （kafka集群的topic）

a1.sinks.k1.channel=c1 (固定配置，配置sink的channel)

a1.channels.c1.type=memory （通道类型：memory，表示存内存；file，表示持久化本地）

a1.channels.c1.capacity=1000000 （通道容量）

a1.channels.c1.transactionCapacity=1000000 （通道事务容量）

## 2、新建并配置zhouwh.conf(一份数据源发送到kafka集群的 两个topic):

CONF文件路径地址：

/opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/flume-ng/conf



作用：

flume读取本地目录下文件并发送消息到kafka集群的两个topic中

配置如下：

a1.sources=r1

a1.sinks=k1 k2 （配置两个sink）

a1.channels=c1 c2 （配置两个channel）

a1.sources.r1.channels=c1 c2 （sources定义两个channel）

a1.sources.r1.type=spooldir

a1.sources.r1.spoolDir=/home/zhouwh

a1.sinks.k1.type=org.apache.flume.plugins.KafkaSink

a1.sinks.k1.metadata.broker.list=h1.hadoop:9092,h2.hadoop:9092,h3.hadoop:9092,h4.hadoop:9092

a1.sinks.k1.serializer.class=kafka.serializer.StringEncoder

a1.sinks.k1.request.required.acks=1

a1.sinks.k1.max.message.size=1000000

a1.sinks.k1.producer.type=async

a1.sinks.k1.custom.encoding=UTF-8

a1.sinks.k1.custom.topic.name=zhouwh

a1.sinks.k1.channel=c1

/////////以下定义第二个sink和第二个kafka的topic

a1.sinks.k2.type=org.apache.flume.plugins.KafkaSink

a1.sinks.k2.metadata.broker.list=h1.hadoop:9092,h2.hadoop:9092,h3.hadoop:9092,h4.hadoop:9092

a1.sinks.k2.serializer.class=kafka.serializer.StringEncoder

a1.sinks.k2.request.required.acks=1

a1.sinks.k2.max.message.size=1000000

a1.sinks.k2.producer.type=async

a1.sinks.k2.custom.encoding=UTF-8

a1.sinks.k2.custom.topic.name=zhouwh2 （第二个topic）

a1.sinks.k2.channel=c2 （第二个channel）

a1.channels.c1.type=memory

a1.channels.c1.capacity=1000000

a1.channels.c1.transactionCapacity=1000000

/////////以下定义第二个channel

a1.channels.c2.type=memory

a1.channels.c2.capacity=1000000

a1.channels.c2.transactionCapacity=1000000

## 3、新建并配置zhouwh\_http.conf

ps：为了满足目前需求，使用zhouwh\_http.conf需要自定义一个HTTPSourceHandler并打包放入服务器（地址：/opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/flume-ng/lib）



CONF文件路径地址：

/opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/flume-ng/conf



作用：

flume接收post消息并发送消息到kafka集群的一个topic中

### 3.1、配置内容

配置如下：

a1.sources=r1

a1.sinks=k1

a1.channels=c1

a1.sources.r1.channels=c1

a1.sources.r1.type=http （http：接收post数据的类型）

a1.sources.r1.bind=172.16.71.63 (flume接收消息的ip)

a1.sources.r1.port=4787 (flume接收消息的port)

a1.sources.r1.handler=org.pq.flumeDemo.sources.HTTPSourceXMLHandler (自定义数据类型)

a1.sinks.k1.type=org.apache.flume.plugins.KafkaSink

a1.sinks.k1.metadata.broker.list=h1.hadoop:9092,h2.hadoop:9092,h3.hadoop:9092,h4.hadoop:9092

a1.sinks.k1.serializer.class=kafka.serializer.StringEncoder

a1.sinks.k1.request.required.acks=1

a1.sinks.k1.max.message.size=1000000

a1.sinks.k1.producer.type=async

a1.sinks.k1.custom.encoding=UTF-8

a1.sinks.k1.custom.topic.name=zhouwh

a1.sinks.k1.channel=c1

a1.channels.c1.type=file （通道类型：本地持久化）

a1.channels.c1.checkpointDir=/mnt/data/flume/checkpoint

a1.channels.c1.dataDirs=/mnt/data/flume/data

a1.channels.c1.capacity=1000000

a1.channels.c1.transactionCapacity=1000000

### 3.2、自定义HTTPSourceHandler

HTTPSourceHandler作用：

取出flume接收的数据流数据，处理后存入固定的List<Event>，flume使用该集合并发送到kafka集群。

HTTPSourceHandler的maven项目工程：



eclipse打包HTTPSourceHandler的flume-demo.jar：



flume-demo.jar放入服务器flume服务的lib下即可

172.16.71.63上：/opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/flume-ng/lib



# 四、启动FLUME

在flume-ng目录下输入以下命令

## 1、默认日志级别启动：

bin/flume-ng agent --conf conf --conf-file conf/zhouwh.conf --name a1 -Dflume.root.logger=INFO,console -c conf

**-c/--conf 后跟配置目录，-f/--conf-file 后跟具体的配置文件，-n/--name 指定agent的名称**

## 2、推荐使用-----修改日志级别为ERROR启动（提升发送消息速度）：



bin/flume-ng agent --conf conf --conf-file conf/zhouwh.conf --name a1 -Dflume.root.logger=ERROR,console -c conf

# 五、测试post

测试flume接收消息并发送至kafka，需要先启动flume和kafka服务，flume启动选择日志级别为INFO查看接收的消息

在flume-ng目录下输入命令：

目录如下：

/opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/flume-ng



命令如下：

bin/flume-ng agent --conf conf --conf-file conf/zhouwh\_http.conf --name a1 -Dflume.root.logger=INFO,console

（说明：1、--conf conf --conf-file conf/zhouwh\_http.conf，运行哪个配置文件；2、--name a1，配置文件中agent的名称为a1；3、-Dflume.root.logger=INFO,console，日志级别）



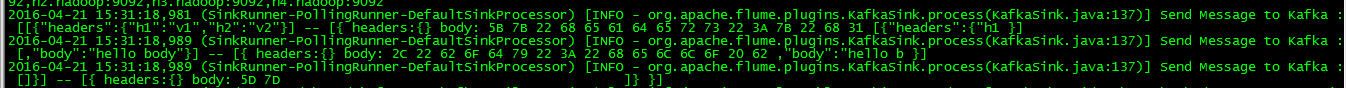
## 5.1、shell模拟post消息



curl -X POST -d'[{"headers":{"h1":"v1","h2":"v2"},"body":"hello body"}]'  http://172.16.71.63:4787

ps：shell命令中http的ip及port同配置文件中sources对应

flume日志输出：



即表示发送消息至flume并传入kafka成功

# 六、运行工程发送本地消息至Kafka

说明：本地运行KafkaProducer (Kafka生产)工程，该工程是遍历本地目录下所有txt文件向Kafka集群发送消息的工具。

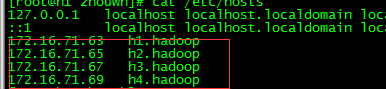
ps：工程在svn上



路径：https://172.16.71.61:8443/svn/Code/trunk/KafKaTest

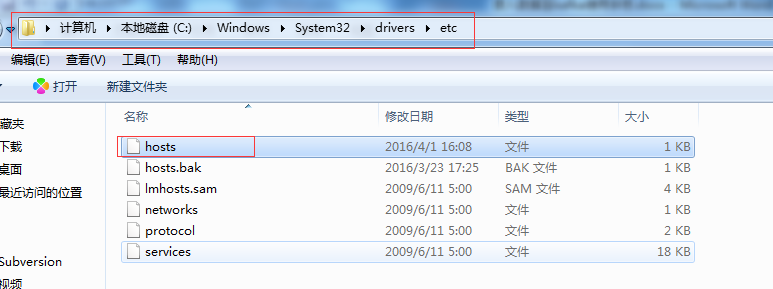
1、添加映射：

查看服务器kafka集群的映射

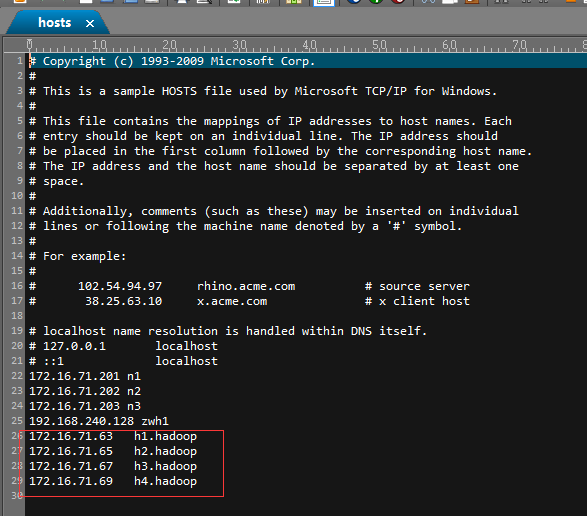


本地主机的映射（hosts）：

hosts地址：

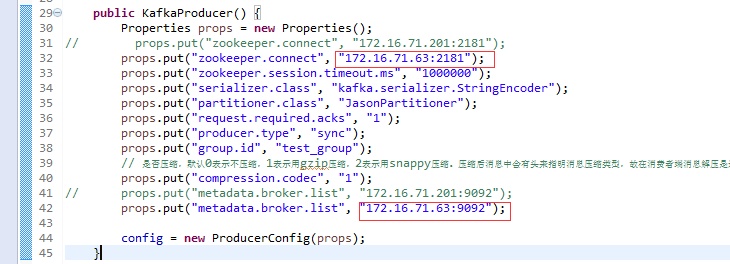


hosts添加内容：



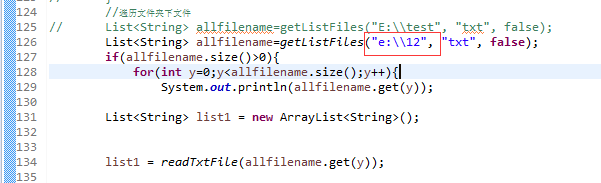
2、修改KafkaProducer中连接配置：

ps：一般情况下部署kafka集群的端口号都是统一的，所以只需要改动IP地址就行

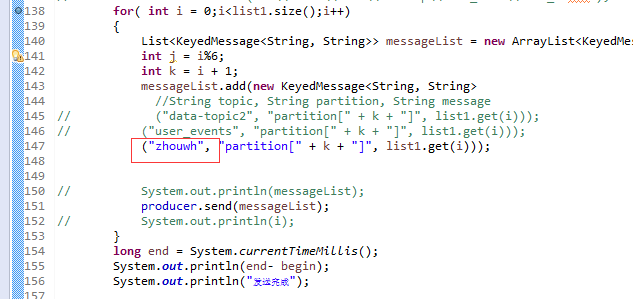


3、修改发送消息文件的本地目录：

ps:目录到文件夹，例截图中是遍历E盘下名称为“12”文件夹下的所有txt文件



4、修改发送到kafka集群中topic名称：



5、eclipse运行工程即可发送消息至kafka

周文辉

2016.5.27