目录

[FusionInsight HD平台安装 2](#_Toc524424935)

[Flume服务端配置 2](#_Toc524424936)

[Flume客户端安装配置 7](#_Toc524424937)

[Flume客户端下载 7](#_Toc524424938)

[Flume客户端安装 10](#_Toc524424939)

[Flume客户端配置 11](#_Toc524424940)

[Kafka配置 12](#_Toc524424941)

[数据库配置 14](#_Toc524424942)

[Spark程序提交运行 17](#_Toc524424943)

[监控程序运行 18](#_Toc524424944)

[程序结果验证 20](#_Toc524424945)

# FusionInsight HD平台安装

参考华为官方文档：“[FusionInsight HD V100R002C70SPC200 快速安装指南 01.pdf](http://support.huawei.com/enterprise/zh/doc/DOC1000175679)”

后续操作需要登录集群的FusionInsight Manager控制台。

# Flume服务端配置

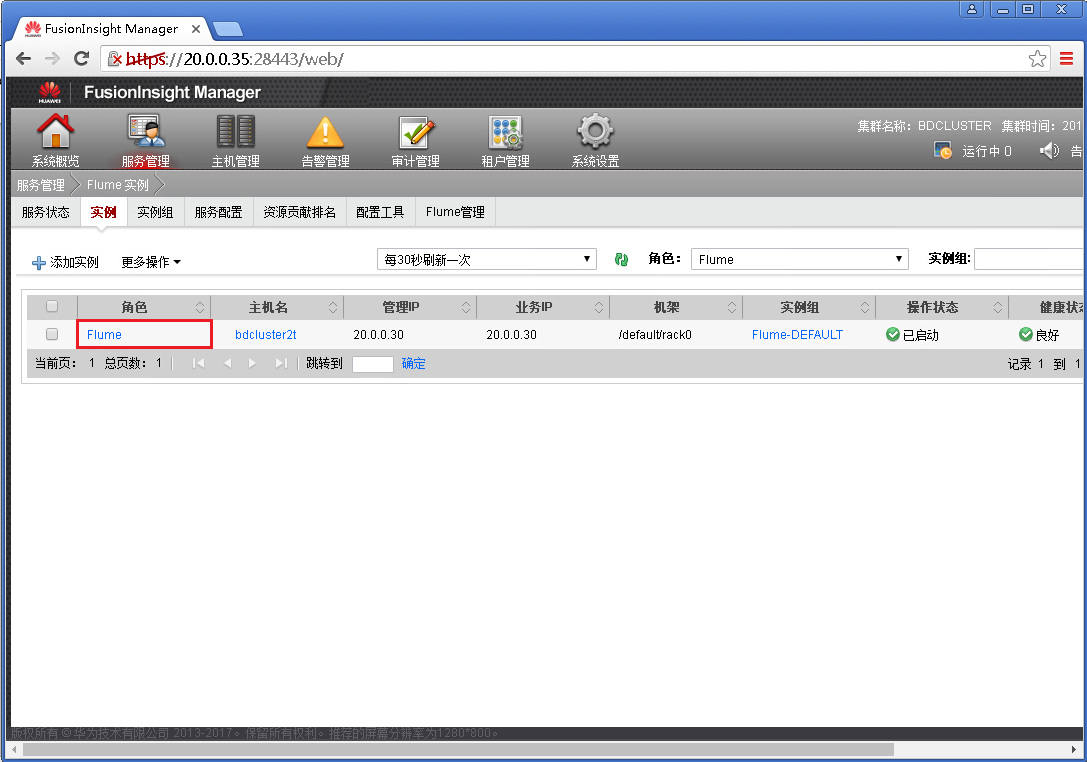
登录到集群环境的FusionInsight Manager，从“服务管理”中选择Flume。



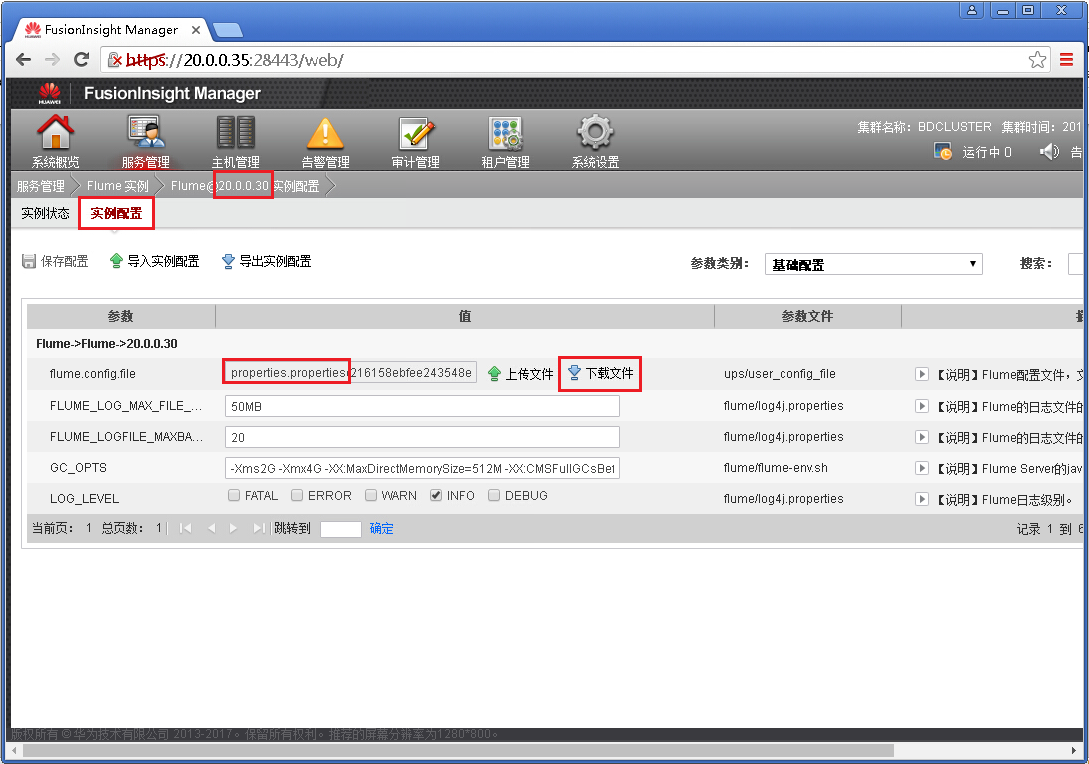
点击“Flume”角色



选择其中一台服务器。



下载服务器实例的配置文件。

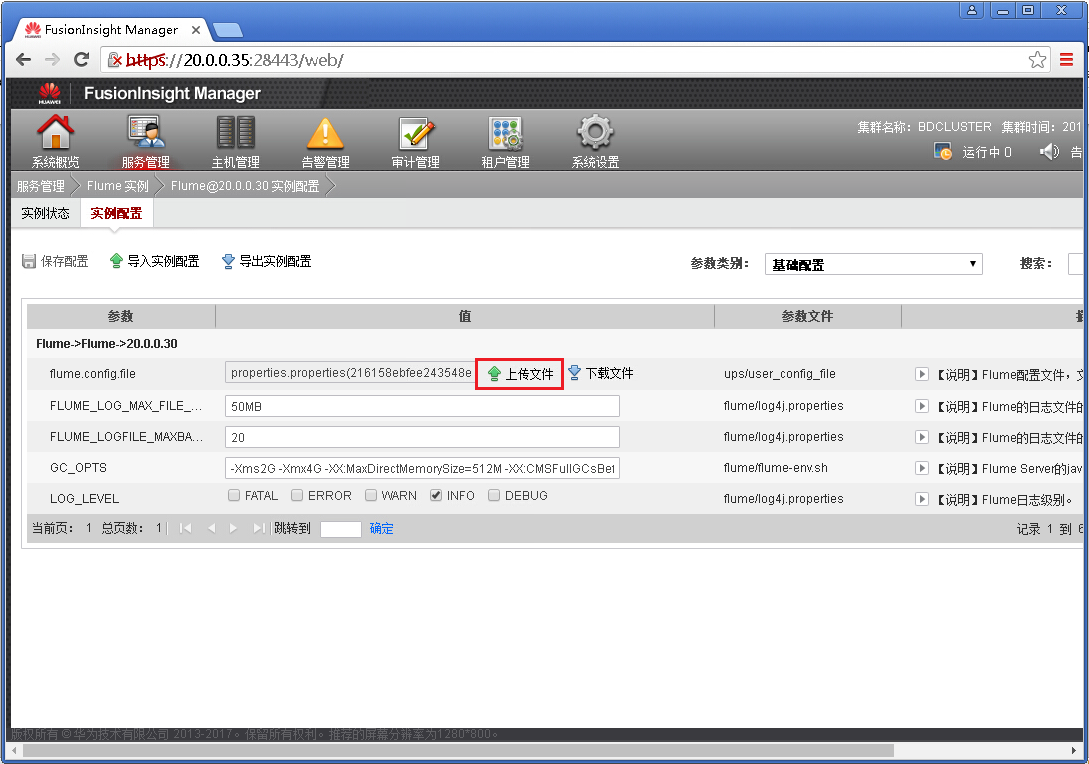


对服务器配置文件进行编辑。

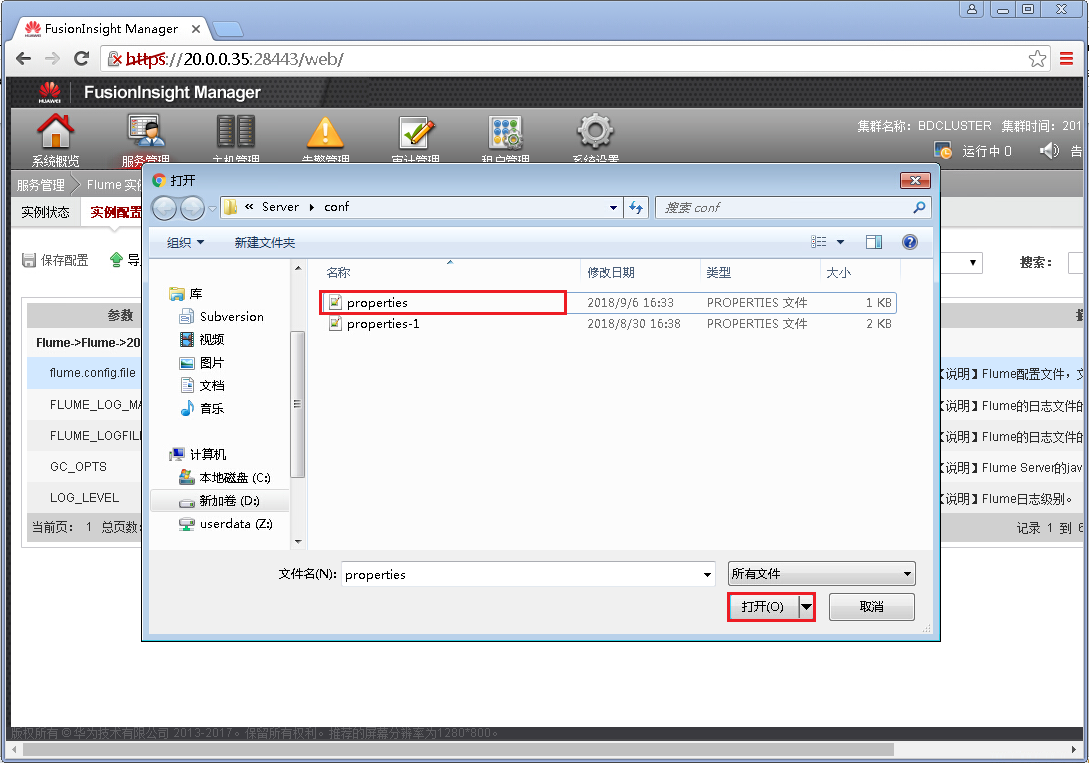
|  |
| --- |
| server.sources = source  server.channels = channel  server.sinks = sink  server.sources.source.type = avro  server.sources.source.bind = **20.0.0.30**  server.sources.source.port = **44444**  server.sources.source.threads = 8  server.sources.source.compression-type = none  server.sources.source.compression-level = 6  server.sources.source.ssl = false  server.sources.source.channels = channel  server.channels.channel.type = file  server.channels.channel.dataDirs = **/srv/BigData/hadoop/data1/flume/data**  server.channels.channel.checkpointDir = **/srv/BigData/hadoop/data1/flume/checkpoint**  server.sinks.sink.type = org.apache.flume.sink.kafka.KafkaSink  server.sinks.sink.kafka.security.protocol = SASL\_PLAINTEXT  server.sinks.sink.channel = channel |

配置文件中的红色部分为服务端的地址及接收端口，绿色部分为集群中存在的目录，如果不存在则需要创建或者指定其他目录。

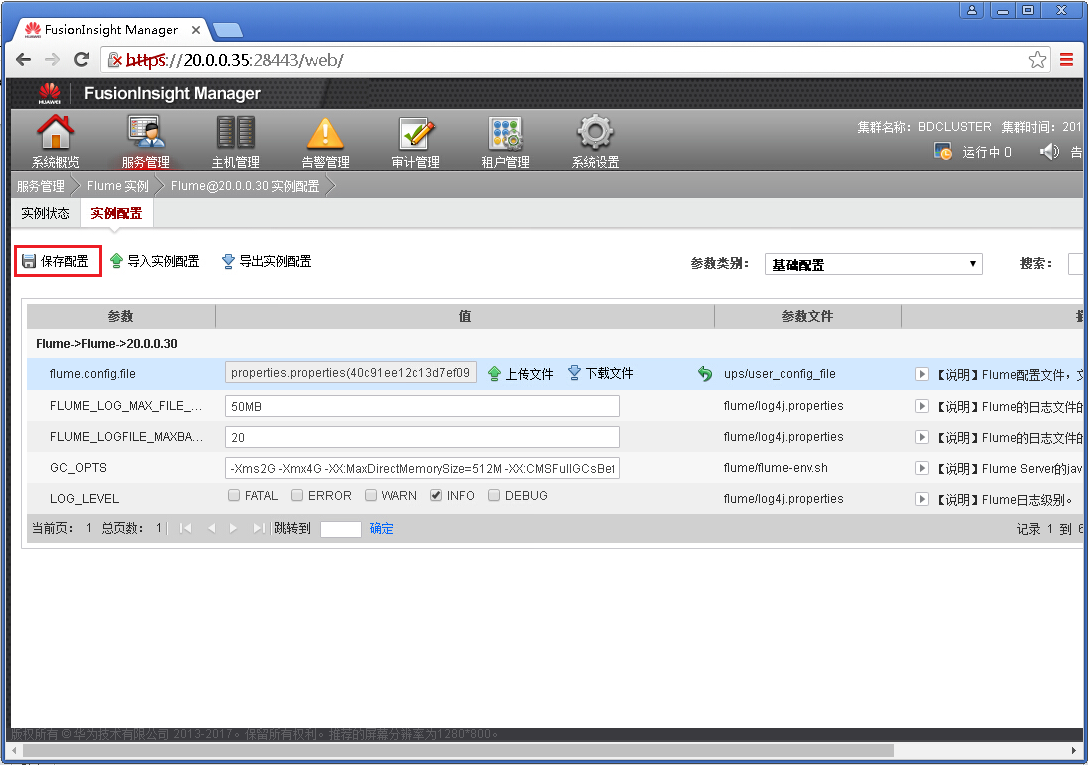
上传配置文件。



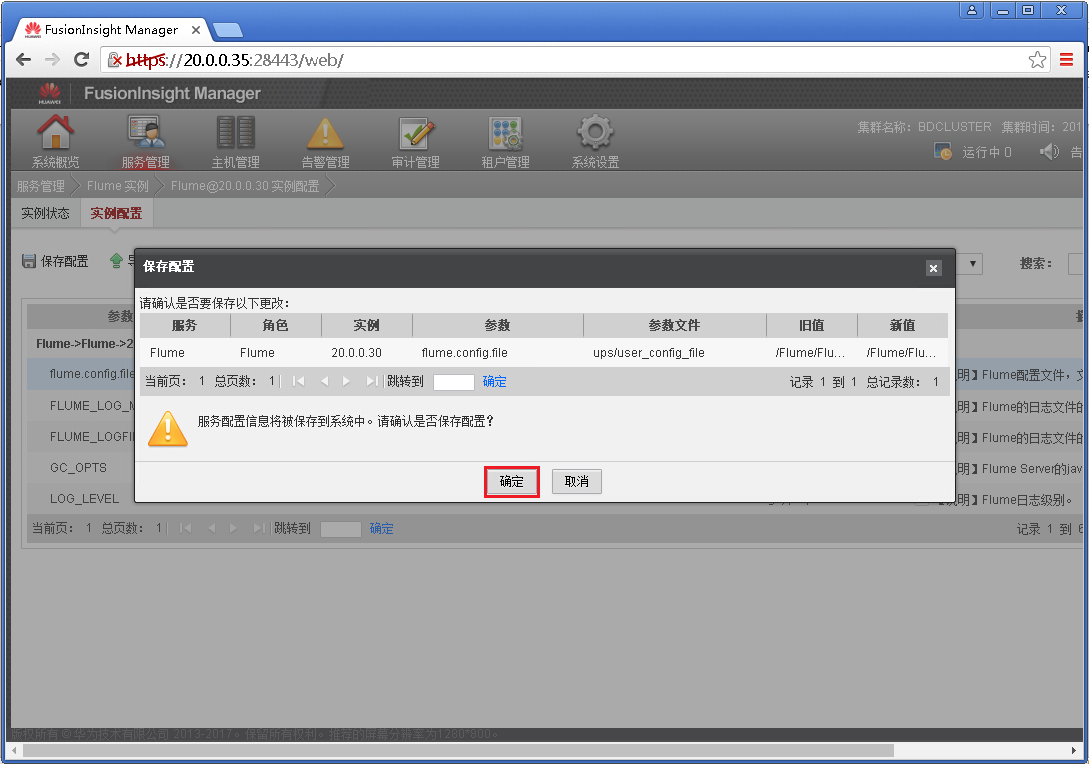
选择刚编辑的配置文件。



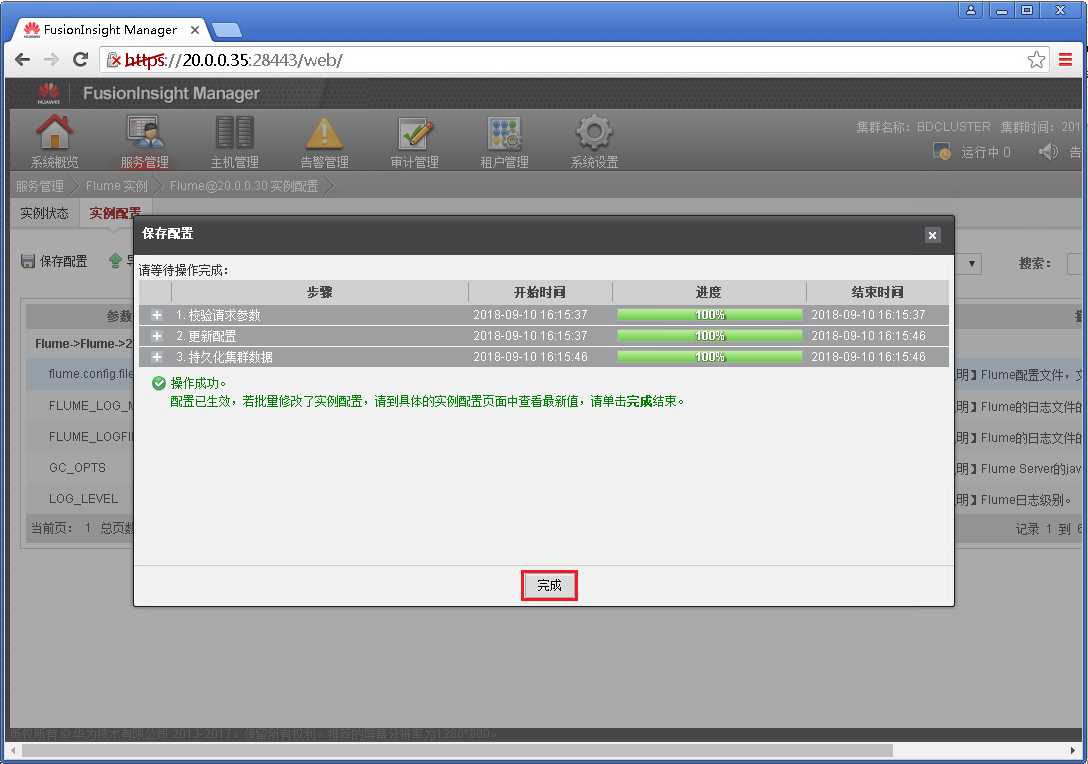
上传完成后保存配置。



选择确定。



等待配置更新并生效。



这样就完成了Flume客户端的配置，后续如果需要进行优化配置文件，按照以上步骤修改后上传即可。

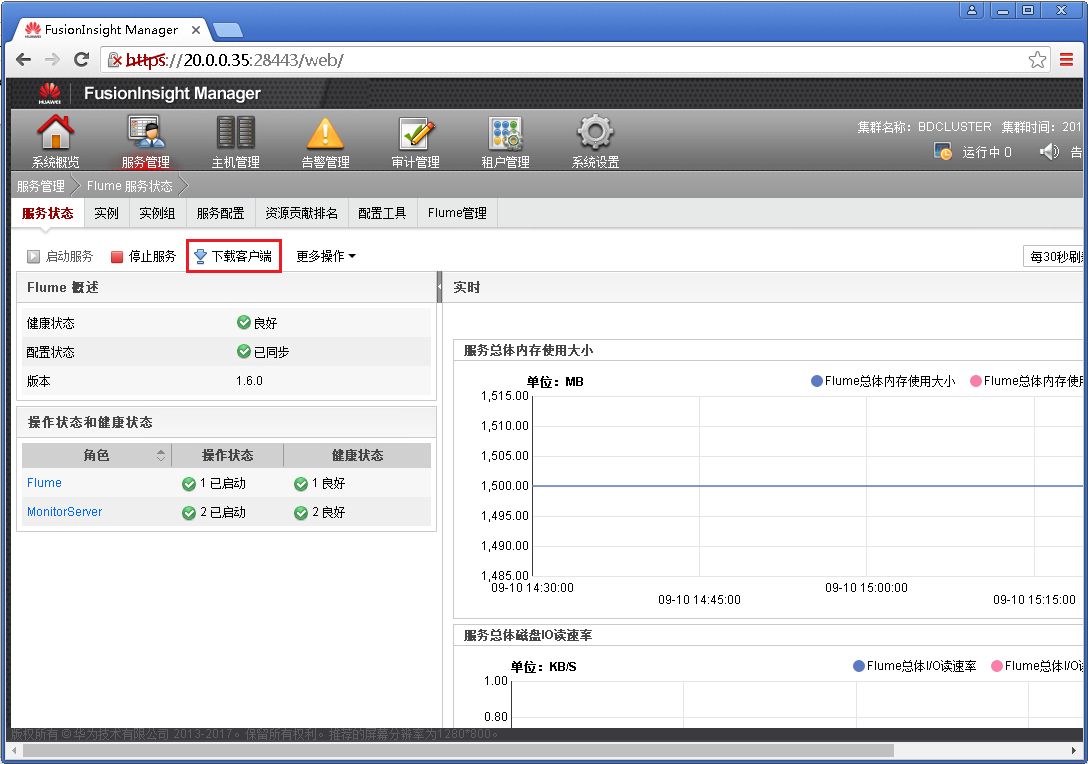
# Flume客户端安装配置

## Flume客户端下载

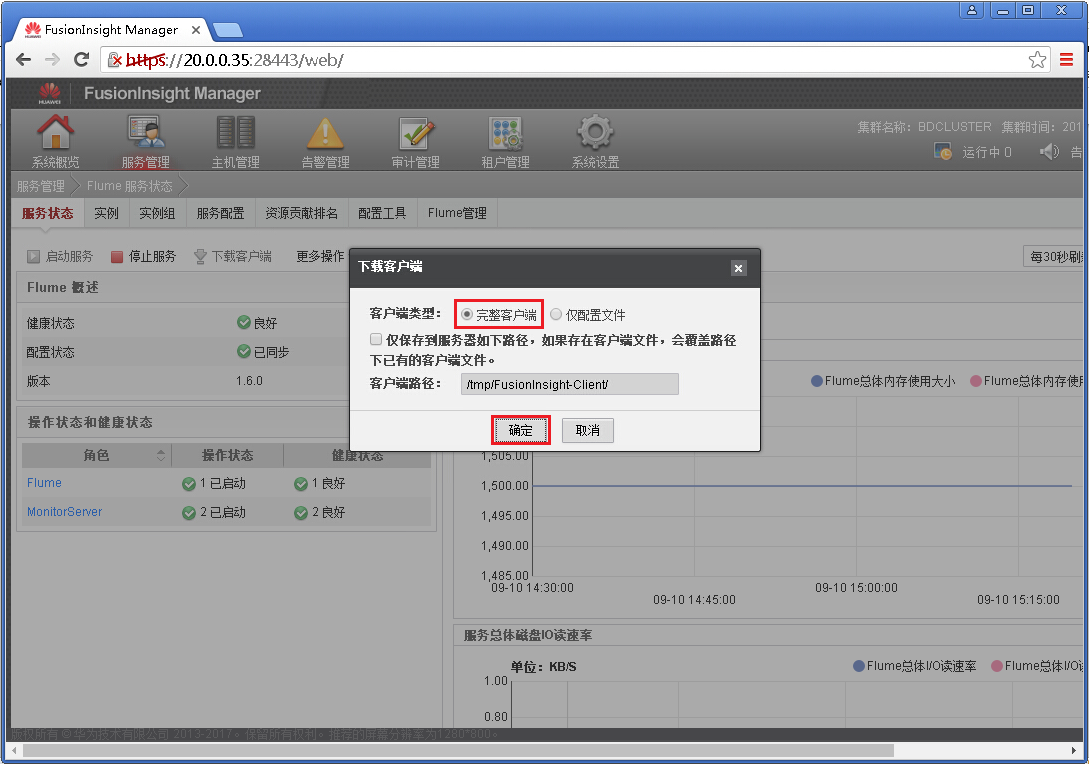
登录到集群环境的FusionInsight Manager，从“服务管理”中选择Flume。



点击“下载客户端”



选择“完整客户端”，确定。



得到客户端安装文件：FusionInsight\_Flume\_Client.tar文件。



## Flume客户端安装

1、从FusionInsight HD控制台下载Flume客户端，将客户端上传到日志采集的服务器

2、解压客户端压缩包

tar -xf FusionInsight\_Flume\_Client.tar

3、校验客户端

sha256sum -c FusionInsight\_Flume\_ClientConfig.tar.sha256

4、解压客户端安装包

tar -xf FusionInsight\_Flume\_ClientConfig.tar

5、根据实际情况创建文件夹

mkdir -p ${HOME}/FlumeClient/client

mkdir -p ${HOME}/FlumeClient/logs

mkdir -p ${HOME}/FlumeClient/works

mkdir -p ${HOME}/FlumeClient/works/data

mkdir -p ${HOME}/FlumeClient/works/data/data1

mkdir -p ${HOME}/FlumeClient/works/data/data2

mkdir -p ${HOME}/FlumeClient/works/checkpoint

mkdir -p ${HOME}/FlumeClient/works/checkpoint/checkpoint1

mkdir -p ${HOME}/FlumeClient/works/checkpoint/checkpoint2

mkdir -p ${HOME}/FlumeClient/works/position

mkdir -p ${HOME}/FlumeClient/works/monitor

6、设置环境变量

export JAVA\_HOME=/usr/lib/jvm/java-1.7.0-openjdk-1.7.0.45.x86\_64

7、切换到软件解压目录

cd FusionInsight\_Flume\_ClientConfig/Flume/FlumeClient

8、执行安装

sh install.sh -d ${HOME}/FlumeClient/client -f 20.0.0.31,20.0.0.33 -c flume/conf/properties.properties -l ${HOME}/FlumeClient/logs -e 20.0.0.30 -n FlumeClientName

-d 指定客户端安装目录

-f 指定MonitorServer角色的服务器地址

-c 指定客户端安装使用的配置文件

-l 指定客户端运行的日志记录目录

-e 指定Flume角色的服务端地址

-n 指定Flume角色的客户端名称

9、到FusionInsight HD控制台查看Flume客户端列表

## Flume客户端配置

客户端的配置需要根据实际情况进行配置，其中：日志文件路径、名称、个数、对应哪个Topic等信息需要按照实际情况进行修改。

|  |
| --- |
| client.sources = source  client.channels = channel  client.sinks = sink  client.sources.source.type = TAILDIR  client.sources.source.channels = channel  client.sources.source.positionFile = **${HOME}/FlumeClient/works/position/position**  client.sources.source.filegroups = **f1 f2**  client.sources.source.filegroups.f1 = **${HOME}/FlumeClient/works/monitor/sys1-log1.log**  client.sources.source.headers.f1.topic = **demo\_topic\_1**  client.sources.source.filegroups.f2 = **${HOME}/FlumeClient/works/monitor/sys1-log2.log**  client.sources.source.headers.f2.topic = **demo\_topic\_2**  client.channels.channel.type = file  client.channels.channel.dataDirs = **${HOME}/FlumeClient/works/data**  client.channels.channel.checkpointDir = **${HOME}/FlumeClient/works/checkpoint**  client.sinks.sink.type = avro  client.sinks.sink.hostname = **20.0.0.30**  client.sinks.sink.port = **44444**  client.sinks.sink.channel = channel |

配置文件中，绿色加粗部分，是“Flume客户端安装”步骤中规划的目录，修改成与安装步骤中创建的目录一致即可。

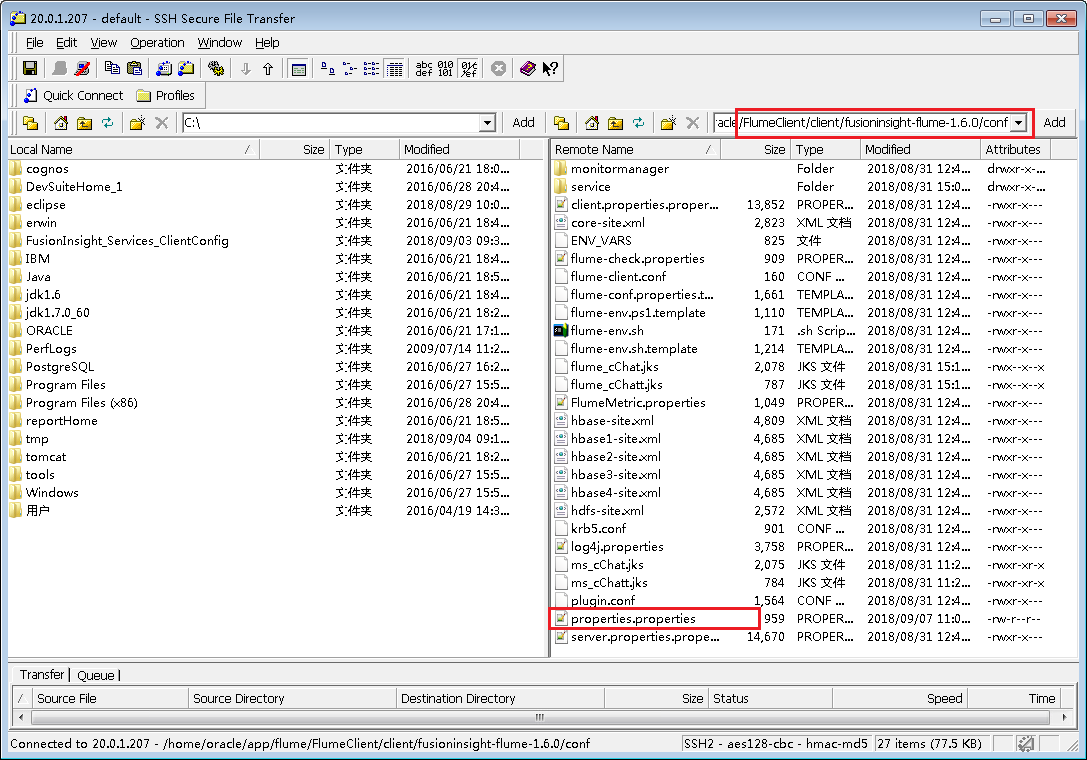
配置文件中，红色加粗部分，是需要采集的日志文件分组配置，f1 f2表示需要采集两个日志文件数据，并且分别指定f1和f2的日志文件绝对路径、对应的Topic信息。如果指采集1个日志文件，则去掉f2的配置；如果需要采集多个日志文件，则增加配置f1 f2 f3 … fn，并且配置f3 … fn的日志文件的绝对路径、对应的Topic。

**注意：配置文件中的${HOME}变量要替换成具体的绝对路径，比如：/home/bigdata**

**最后的红色加粗的服务器IP、端口必须要与“Flume服务端配置”中指定的IP、端口一致。**

**配置完成后，将文件进行保存，名称必须是“properties.properties”。**

然后使用FTP客户端将配置文件上传到Flume客户端的conf目录下。



**注意：这里Flume客户端的安装路径是指前面的安装过程中 -d 参数指定的路径，并非安装包的解压路径。按照以上步骤安装，则实际路径应该为：${HOME}/FlumeClient/client/fusioninsight-flume-1.6.0/conf。**

如果需要在多个系统、多台服务器上进行日志采集，在每台服务器上重复执行“Flume客户端安装配置”步骤，完成客户端的安装及配置。

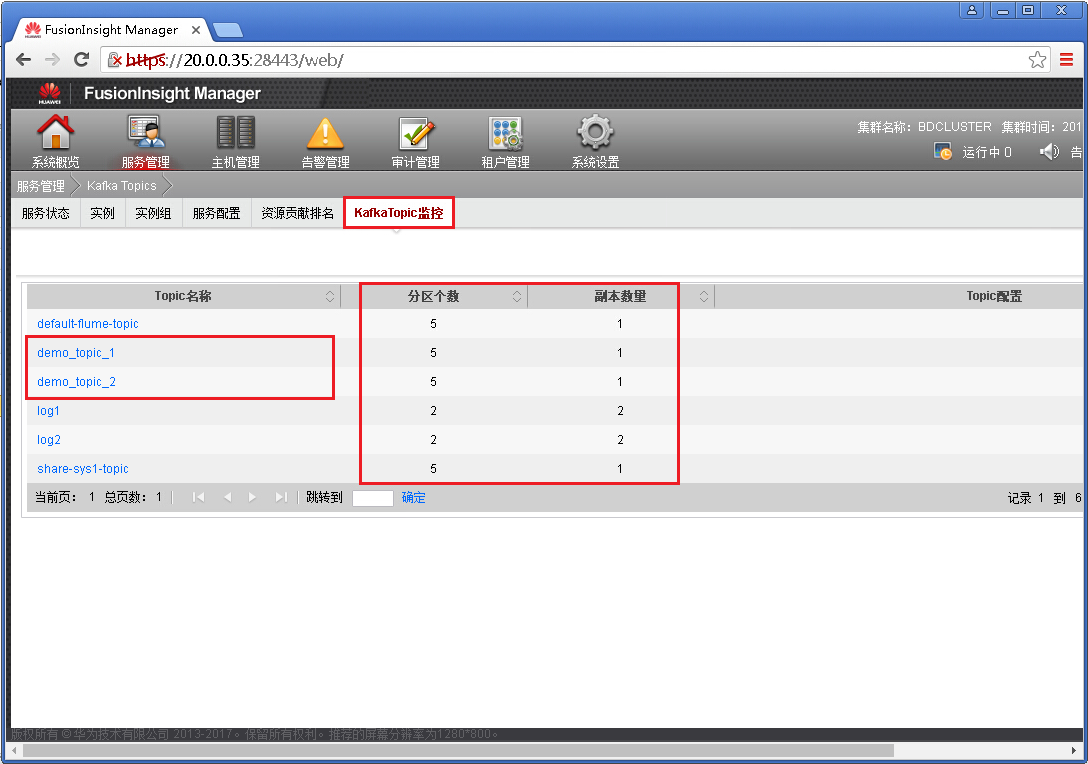
# Kafka配置

通过以上Flume的配置，会自动在Kafka创建Flume客户端配置文件中配置的Topic。

可以通过“服务管理”中点击Kafka。



在“Kafka Topic监控”中进行查看。



虽然能够自动创建Topic，但是这是不建议的方式，因为自动创建的Topic的分区个数、副本数量的设置采用了默认的方式，并非根据集群中Broker的数量进行设置，也就没法充分利用Kafka的性能，所以建议提前手工创建Topic。

登录到“FusionInsight HD平台安装”中规划的安装有FusionInsight客户端的服务器，在Kafka客户端安装路径的bin目录下执行：

kafka-topics.sh --create --topic topic\_name --partitions 5 --replication-factor 1 --zookeeper 20.0.0.29:24002,20.0.0.32:24002,20.0.0.33:24002/kafka

--partitions 指定分区数

--replication-factor 指定副本数

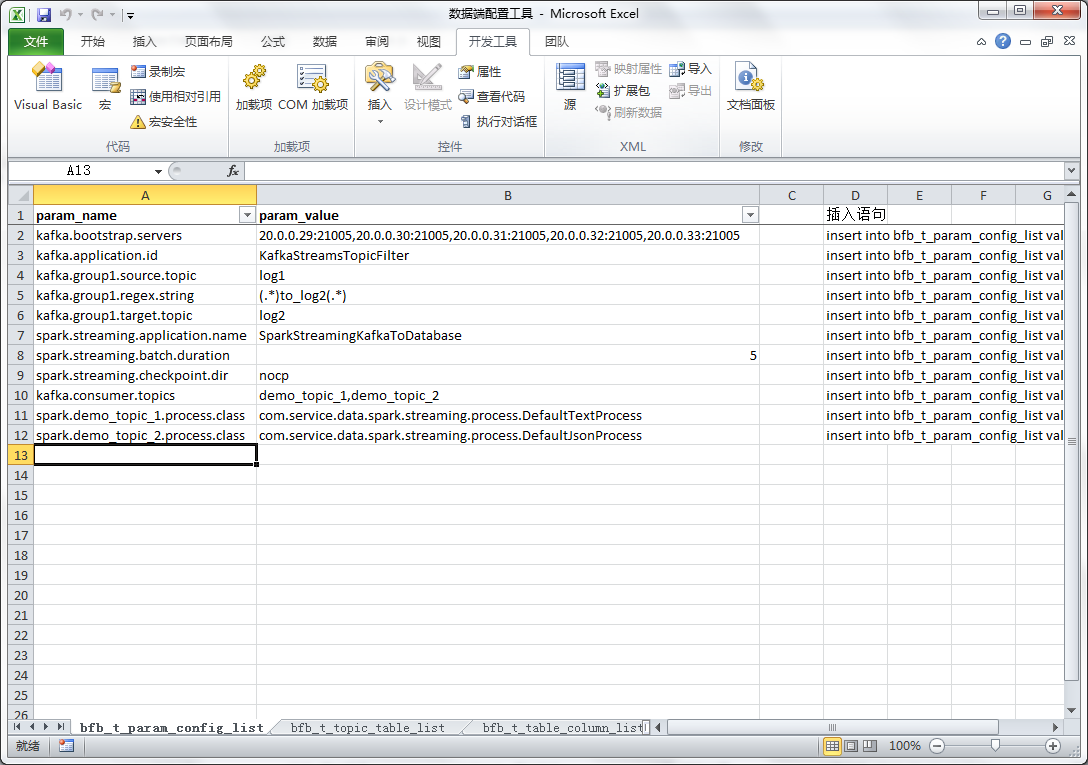
--zookeeper 指定Zookeeper地址

# 数据库配置

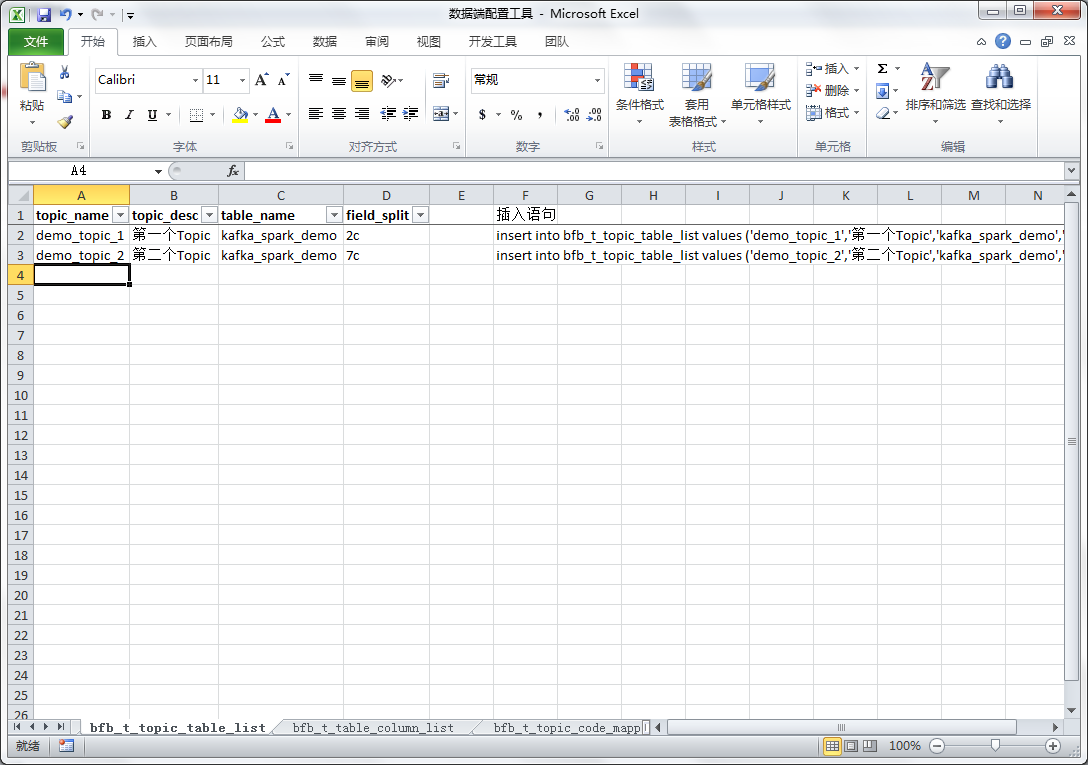
执行sqls文件夹下的CreateTable.sql以创建配置表结构。

根据实际情况修改“数据端配置工具.xlsx”文件中的内容，并生成SQL语句。

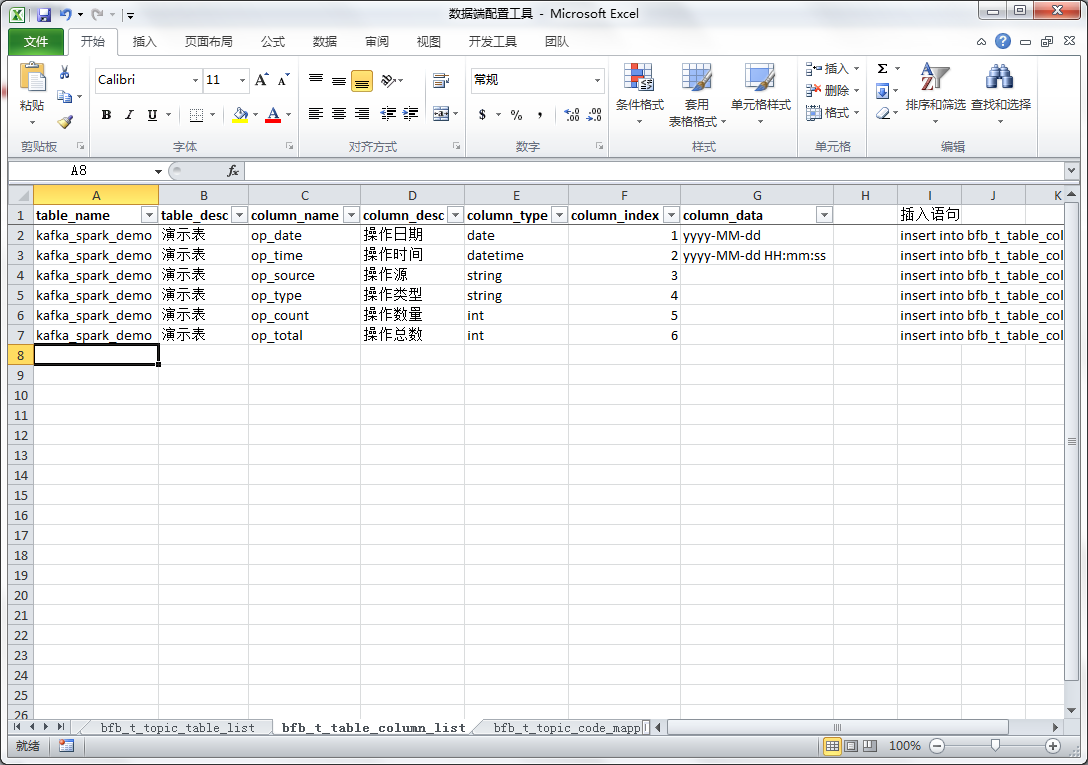
1、程序运行的配置文件，bfb\_t\_param\_config\_list，具体参考InitData.sql中的注释进行修改及SQL语句生成。



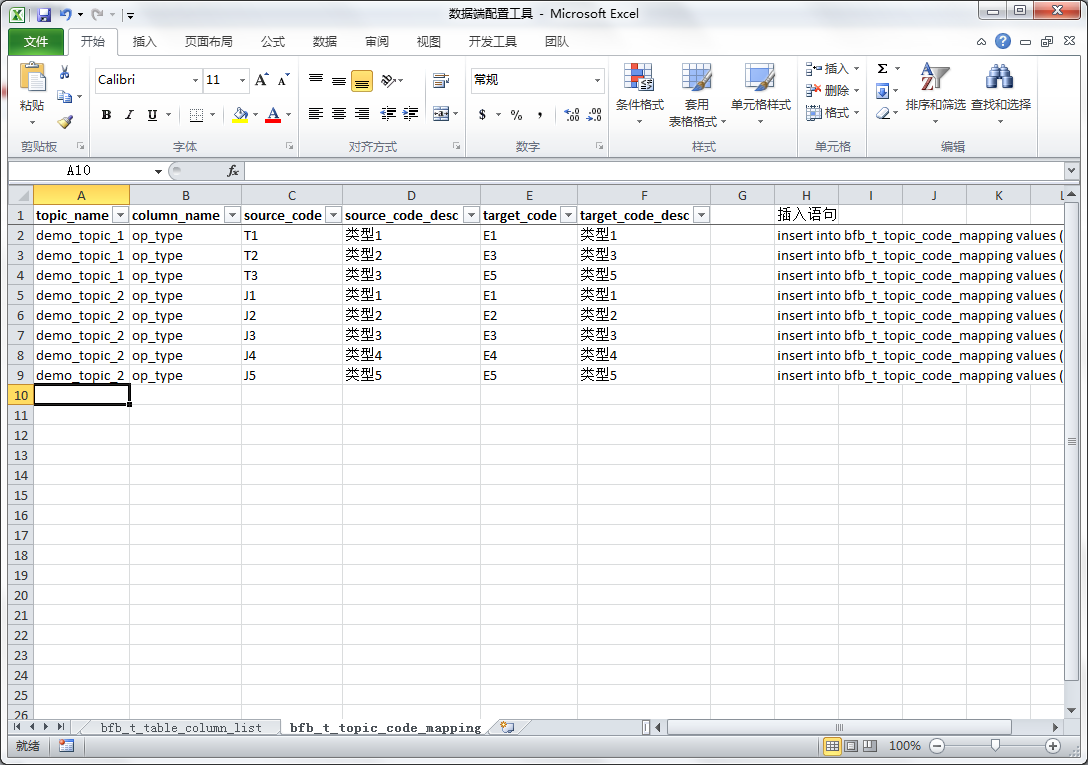
2、Topic对应的表配置，bfb\_t\_topic\_table\_list，每个Topic需要配置一条记录，分别配置哪个Topic的数据处理到哪张表中，以及该Topic的数据是用什么分隔符分隔的，为了保障特殊字符做分隔符的时候能够配置进数据库，这里需要将分隔符转换成16进制数据进行填写，比如：逗号“,”分隔填2c，竖线“|”分隔填7c。不同Topic的数据可以处理到同一张表，但是一个Topic的数据不能处理到两张表。



3、表字段配置，bfb\_t\_table\_column\_list，根据采集的日志的情况，确定每张表的表结构，与日志格式对应，需要配置字段名称、类型、顺序等，如果是日期、时间戳格式的数据，最后一列配置数据的格式。

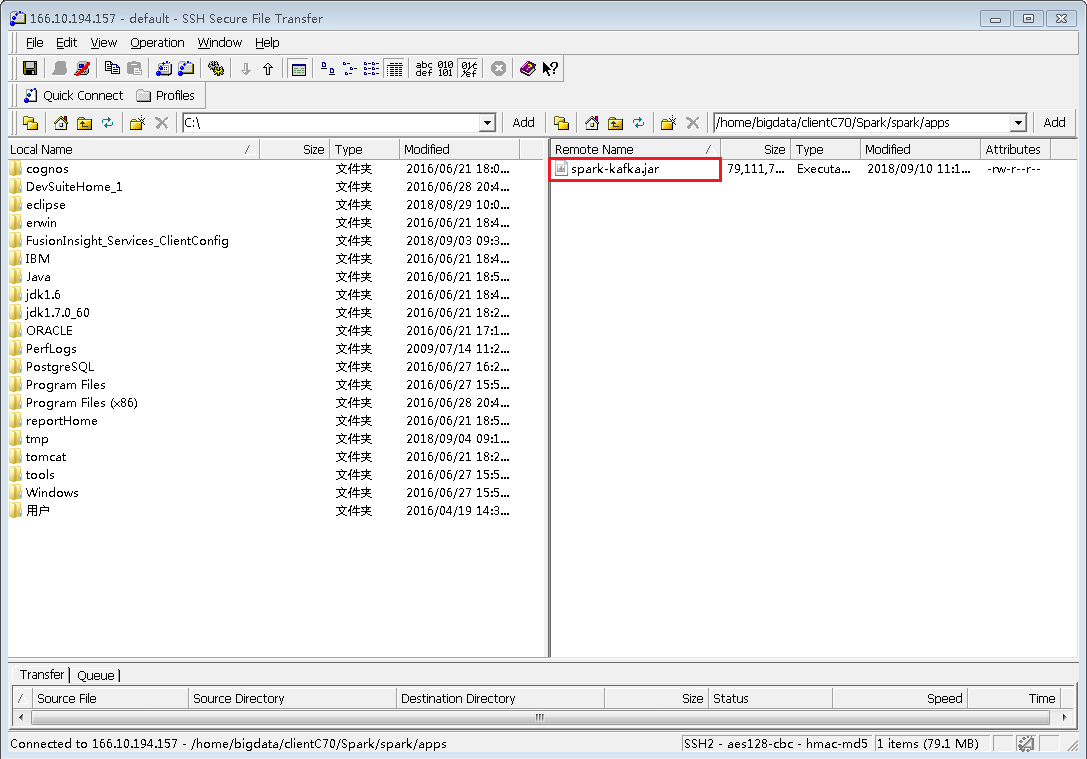


4、码值标准化转换配置，bfb\_t\_topic\_code\_mapping，业务系统按照自己系统的代码值记录到日志文件，但是在自己的系统中相同含义的代码可能码值不一样，这里配置码值的映射关系，需要配置Topic中的哪个字段的源代码值是什么以及对应的目标代码值是什么。



# Spark程序提交运行

登录到“FusionInsight HD平台安装”中规划的安装有FusionInsight客户端的服务器，将spark-kafka.jar程序包上传到Spark客户端安装目录下的apps（如果没有就创建该目录）目录下。



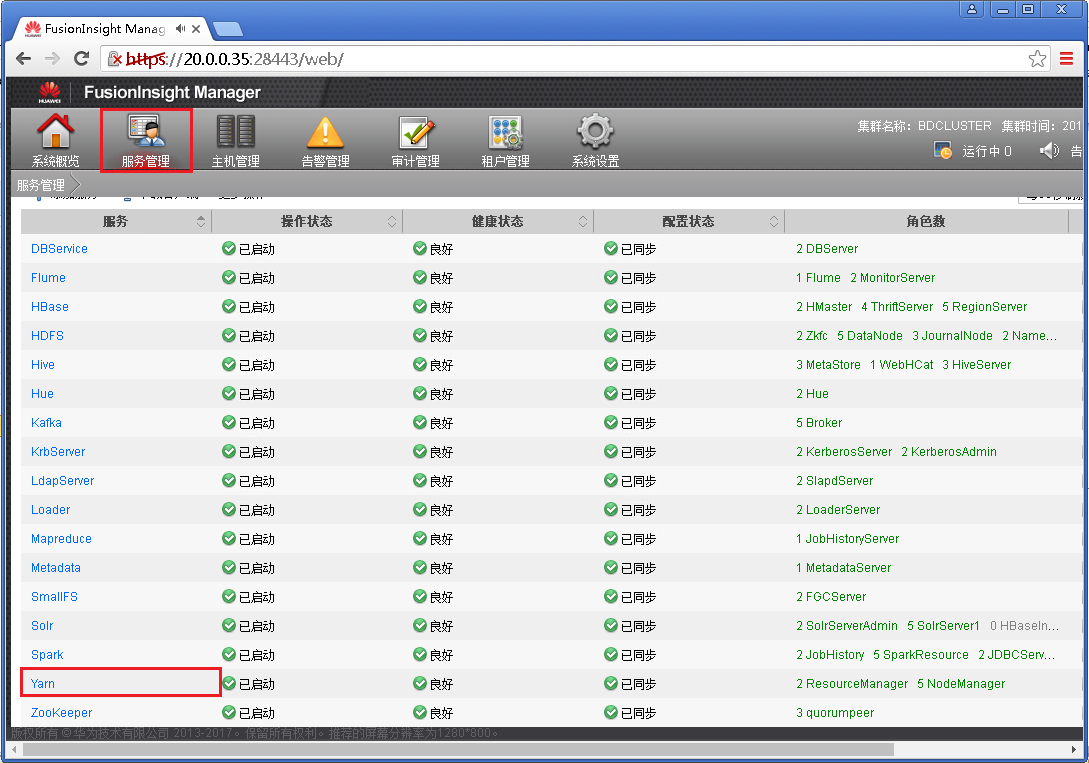
并且在Spark安装目录下执行：

bin/spark-submit --class com.service.data.spark.streaming.StreamingFromKafka --master yarn-client apps/spark-kafka.jar

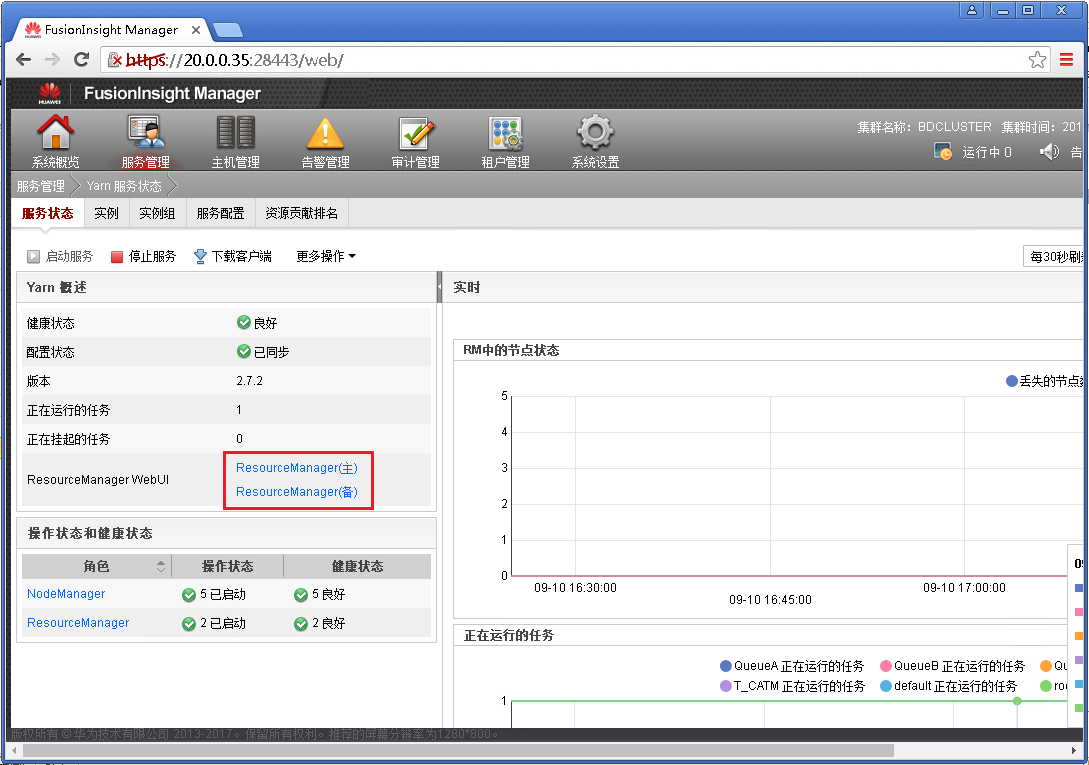
将程序提交到Yarn上运行。

# 监控程序运行

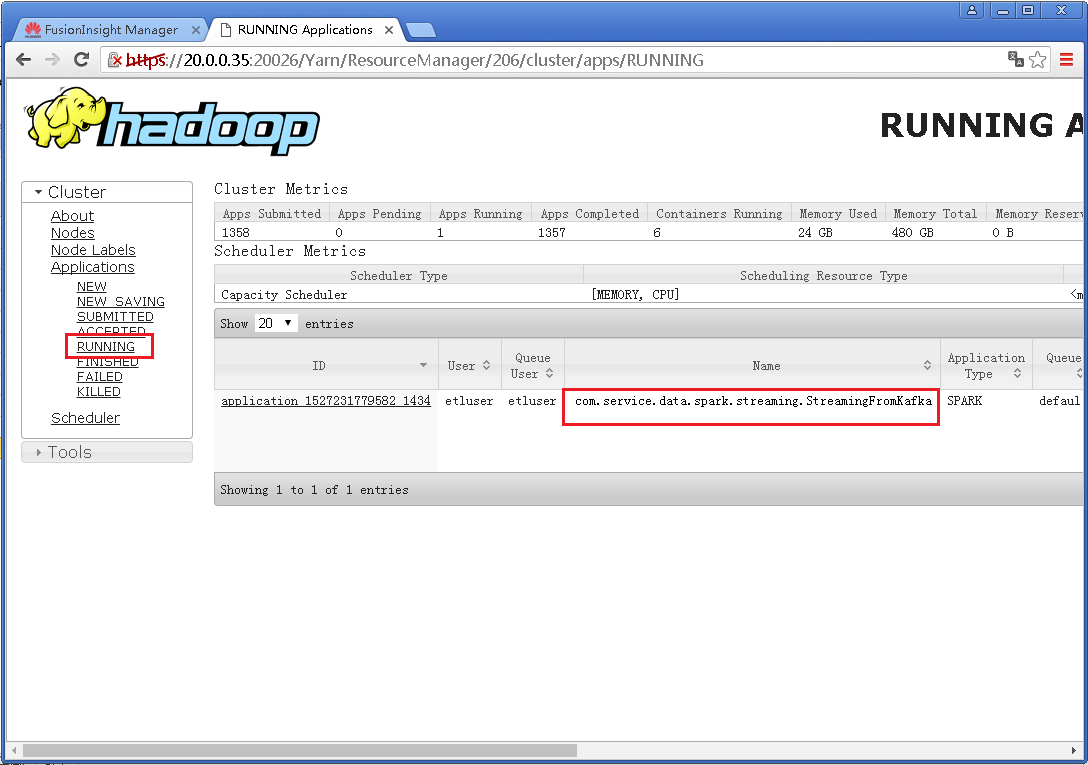
通过“服务管理”，点击Yarn。



点击“ResourceManager”，正常情况下点击“主”，也有可能在“备”上。



在打开的UI界面可以看到正在运行的应用程序。



# 程序结果验证

为了模拟实时数据的产生，编写Shell脚本进行数据的生成，并将生成的数据写入到Flume客户端配置文件中指定的日志文件中。Shell目录下提供了相应的脚本。

config.cnf 配置目标表spark\_kafka\_demo的表结构，生成的数据会根据其中的字段个数及类型进行数据的生成。

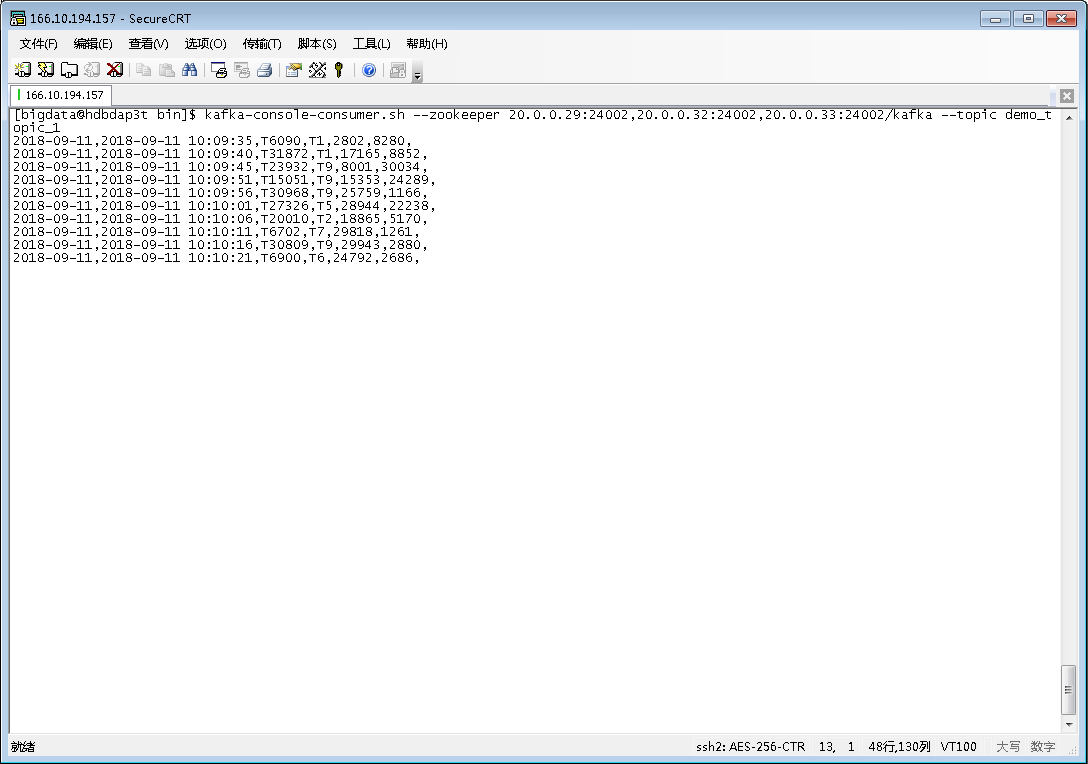
genText.sh 根据config.cnf中的字段个数、顺序、类型产生以逗号分隔的字符串数据。

genJson.sh 根据config.cnf中的字段个数、顺序、类型产生JSON格式的数据。

执行以下命令以产生两种类型的数据到日志文件。

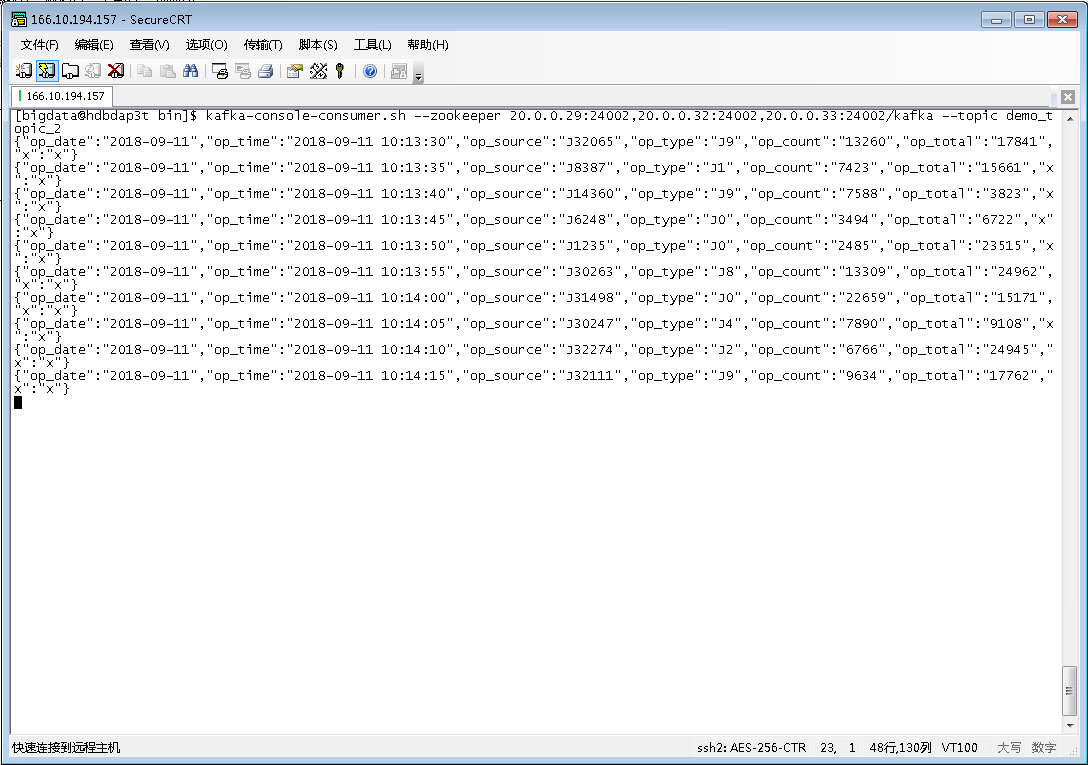
sh genText.sh >> ${HOME}/FlumeClient/works/monitor/sys1-log1.log

通过Kafka的控制台消费者可以消费产生的数据，以验证Flume采集日志并放入Kafka是正常的。



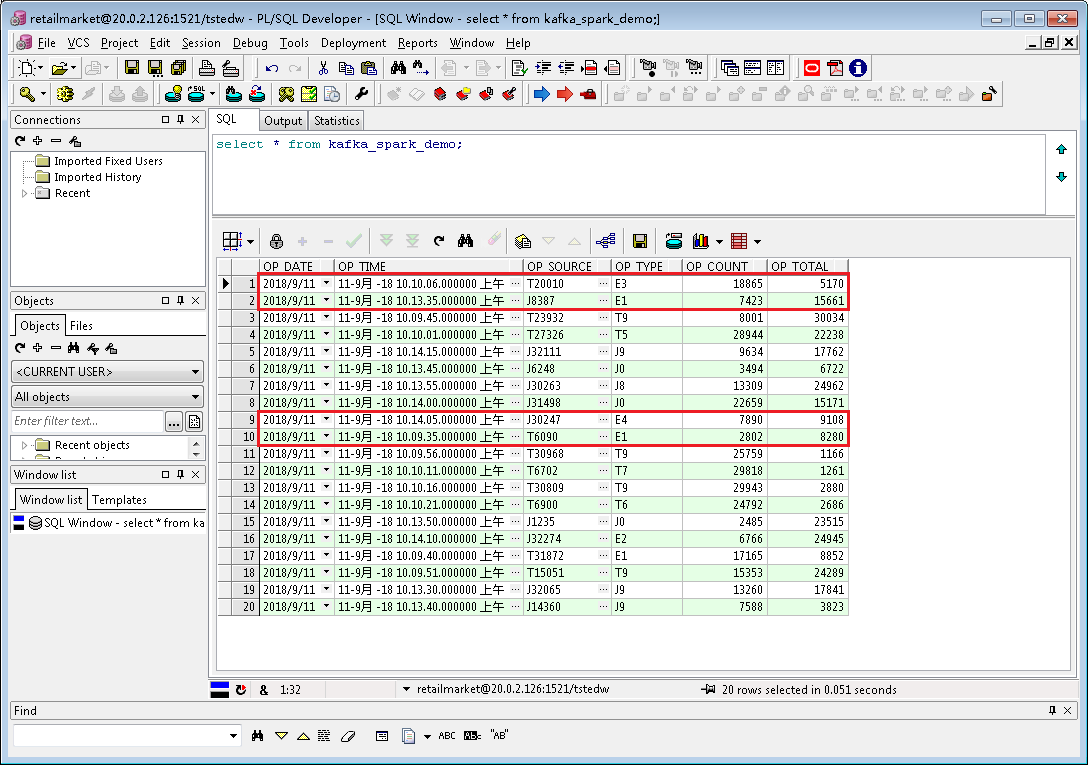
sh genJson.sh >> ${HOME}/FlumeClient/works/monitor/sys1-log2.log

通过Kafka的控制台消费者可以消费产生的数据，以验证Flume采集日志并放入Kafka是正常的。

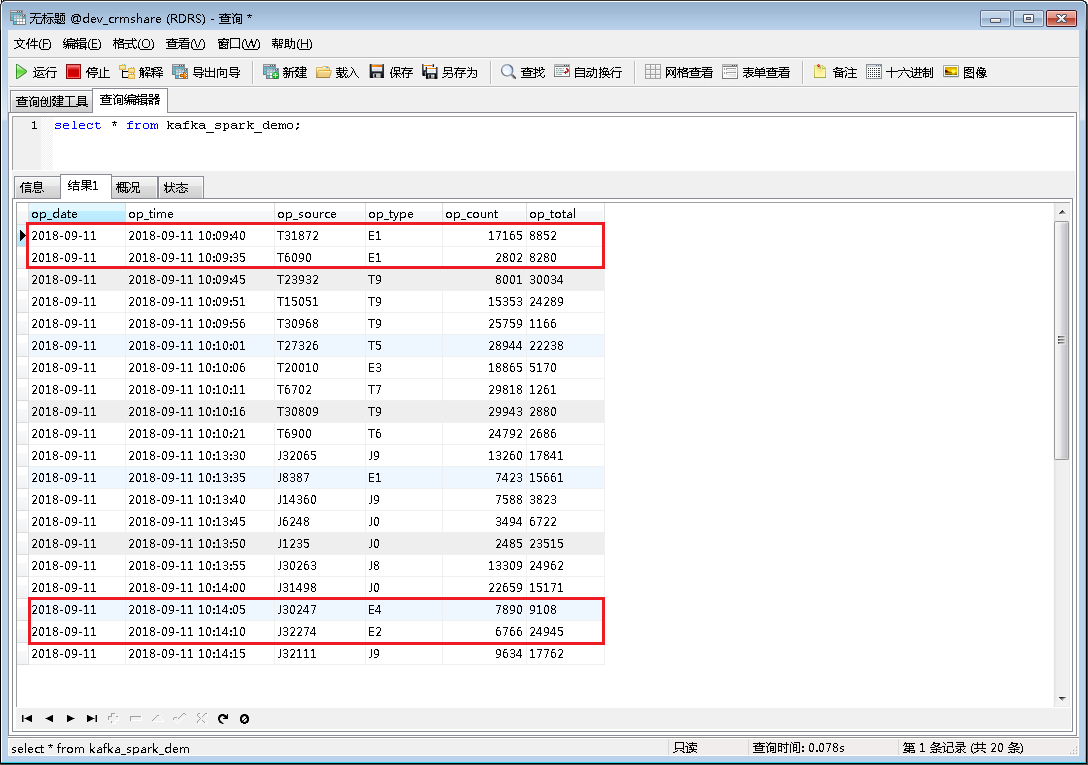


通过查询数据库的目标表的数据，以验证SparkStreaming程序正常消费了kafka的数据并正确处理到了数据库中。并且，可以对比产生的数据是否按照码值标准化处理配置中的映射规则进行了标准化处理。

处理到Oracle数据库。



处理到MySQL数据库。



整个流程完成。