# 日志采集框架Flume

## Flume介绍

*官网文档：http://flume.apache.org/*

Flume是Cloudera提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统。

Flume可以采集文件，socket数据包等各种形式源数据，又可以将采集到的数据输出到HDFS、hbase、hive、kafka等众多外部存储系统中。

一般的采集需求，通过对flume的简单配置即可实现。

Flume针对特殊场景也具备良好的自定义扩展能力，因此，flume可以适用于大部分的日常数据采集场景。

Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

Flume属于生产者与消费者模式；

当前Flume有两个版本Flume 0.9X版本的统称Flume-og，Flume1.X版本的统称Flume-ng。由于Flume-ng经过重大重构，与Flume-og有很大不同，使用时请注意区分。

Flume目前不支持高可用，所以需要编写相应脚本来监听，当agent挂掉时，脚本自动启动agent。Agent会记录上一次的收集位置。

## 运行机制

Flume分布式系统中最核心的角色是agent（一个节点，一个进程），flume采集系统就是由一个个agent所连接起来形成。每一个agent相当于一个数据传递员（Source 到 Channel 到 Sink之间传递数据的形式是Event事件；Event事件是一个数据流单元。）

内部有三个组件：

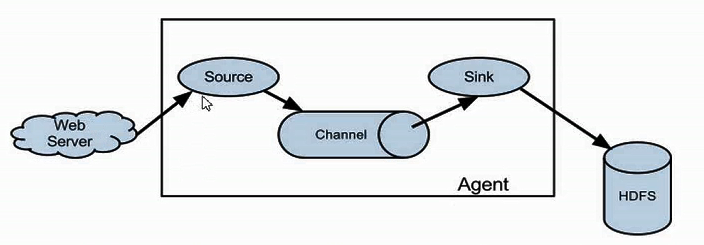
1）Source：采集源，用于跟数据源对接，以获取数据。

2）Sink：下沉地，采集数据的传送目的，用于往下一级agent传递数据或者往最终存储系统传递数据。

3）Channel：angent内部的数据传输通道，用于从source将数据传递到sink。

Flume传输过程：

source监控某个文件或数据流，数据源产生新的数据，拿到该数据后，将数据封装在一个Event中，并put到channel后commit提交，channel队列先进先出，sink去channel队列中拉取数据，然后写入到HDFS中。



Web server

Avro sink

客户端，发送数据到

请求服务端的监听端口 44444

服务端，接受数据

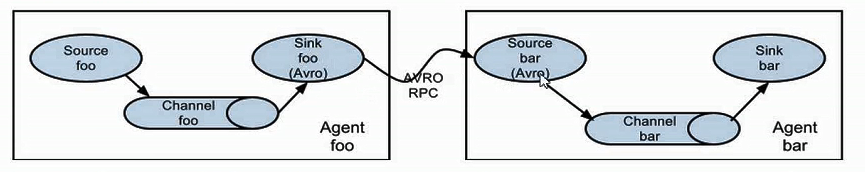
规划监听端口：44444

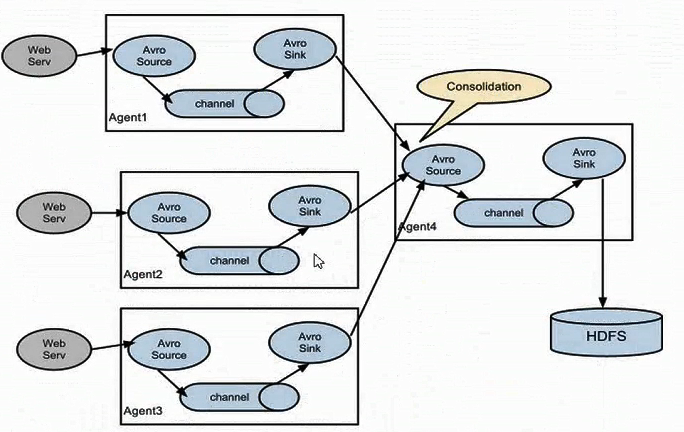
第一个f的数据写到第二个f，RPC协议

1. 主动source，由agent主动读取数据；

2.被动source，由外界把数据发送过来；

根据数据的位置，读取方式来选择使用哪种数据源





## flume 安装：

|  |
| --- |
| 安装（需要hadoop环境）  1、上传  2、解压：  *[root@node1 tools]# tar -zxvf apache-flume-1.8.0-bin.tar.gz -C /home/bigdata/*  3、修改conf/flume-env.sh 文件中的JDK目录  注意：JAVA\_OPTS 配置 如果我们传输文件过大 报内存溢出时 需要修改这个配置项  4、验证安装是否成功 ./flume-ng version  5、配置环境变量：  export FLUME\_HOME=/home/bigData/flume/flume-1.7.0  export PATH=$PATH:$FLUME\_HOME/bin |

## 案例：

### Flume 执行脚本命令说明：

|  |
| --- |
| *flume-ng 启动项*  *-n agent名称；*  *-c 配置文件后缀，可不写；*  *-f 配置文件*  *-D 日志输出级别和位置*  *Bin/flume-ng agent -n a1 -c conf -f 文件名.conf -D flume.root.logger=INFO,console* |

### 案例1、监听一个端口，接收数据，以log的形式发送

*编写配置文件 ： flumeTest01.conf*

|  |
| --- |
| *############################################################*  *# 1、命名：一个agent（进程）内各角色别名：a1*  *############################################################*  *#sources别名:r1*  *a1.sources = r1*  *#sink别名:k1*  *a1.sinks = k1*  *#channel别名：c1*  *a1.channels = c1*  *############################################################*  *# 2、source参数配置：*  *############################################################*  *#source类型：netcat 专门监听一个端口，当该端口有数据产生时，获取数据*  *a1.sources.r1.type = netcat*  *#监听的主机ip:*  *a1.sources.r1.bind = 192.168.100.101*  *#监听的端口：*  *a1.sources.r1.port = 8888*  *############################################################*  *# 3、channel参数配置：*  *############################################################*  *#channel类型：基于内存缓存*  *a1.channels.c1.type = memory*  *#默认该通道中最大的可以存储的event数量是1000*  *a1.channels.c1.capacity = 1000*  *#每次最大可以source中拿到或者送到sink中的event数量也是100*  *a1.channels.c1.transactionCapacity = 100*  *############################################################*  *# 4、channel参数配置：*  *############################################################*  *#sink类型：以日志的方式输出*  *a1.sinks.k1.type = logger*    *############################################################*  *# 5、集成一个agent当中的各个角色：*  *############################################################*  *a1.sources.r1.channels = c1*  *a1.sinks.k1.channel = c1* |

|  |
| --- |
| *1）运行flume脚本输出日志：*  *flume-ng agent -n a1 -f flumeTest01.conf -D flume.root.logger=INFO,console*  *2）安装telnet命令：yum install -y telnet*  *使用：telnet 192.168.9.xx 8888*  *退出 ctrl+] quit*   1. *测试：*   *[root@node1 flume-1.7.0]# telnet node1 8888*  *Trying 192.168.1.60...*  *Connected to node1.*  *Escape character is '^]'.*  *shi*  *OK*  *kkkkkk*  *OK*  *查看日志输出：*  *19/04/12 16:50:51 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChannelImpl[/192.168.*  *19/04/12 16:52:27 INFO sink.LoggerSink: Event: { headers:{} body: 73 68 69 79 6F 6E 67 78 69 6E 0D*  *19/04/12 16:52:44 INFO sink.LoggerSink: Event: { headers:{} body: 61 73 64 66 61 73 64 66 0D asdfasdf. }*  *19/04/12 16:52:44 INFO sink.LoggerSink: Event: { headers:{} body: 61 73 64 66 61 73 64 66 0D asdfasdf. }* |

### 案例2、监控一个文件内容的变化，当有数据追加时，实时打印到控制台

*编写flume脚本：flumeTest02.flume*

|  |
| --- |
| *############################################################*  *# 1、命名：一个agent（进程）内各角色别名：a1*  *############################################################*  *#sources别名:r1*  *a1.sources = r1*  *#sink别名:k1*  *a1.sinks = k1*  *#channel别名：c1*  *a1.channels = c1*  *############################################################*  *# 2、source参数配置：*  *############################################################*  *#source类型：exec ,监听一个文件内容的变化*  *a1.sources.r1.type = exec*  *#指定监听的文件：*  *a1.sources.r1.command = tail -F /home/case/test.log*  *############################################################*  *# 3、channel参数配置：*  *############################################################*  *#channel类型：*  *a1.channels.c1.type = memory*  *#默认该通道中最大的可以存储的event数量是100*  *a1.channels.c1.capacity = 1000*  *#每次最大可以source中拿到或者送到sink中的event数量也是100*  *a1.channels.c1.transactionCapacity = 100*  *############################################################*  *# 4、sink参数配置：*  *############################################################*  *#sink类型：以日志的方式输出*  *a1.sinks.k1.type = logger*  *############################################################*  *# 5、集成一个agent当中的各个角色：*  *############################################################*  *a1.sources.r1.channels = c1*  *a1.sinks.k1.channel = c1* |

执行脚本：

|  |
| --- |
| 1、创建空文件： test.log  2、执行脚本命令；  *flume-ng agent -n a1 -f flumeTest02.conf -D flume.root.logger=INFO,console*  3、循环添加数据 : 追加日志：*echo kkkkk >> /home/case/test.log* |

### 案例3、监控指定目录下新文件的生成，然后输出到控制台

编写flume脚本 : flumeTest03.flume

|  |
| --- |
| *############################################################*  *# 1、命名：一个agent（进程）内各角色别名：a1*  *############################################################*  *#sources别名:r1*  *a1.sources = r1*  *#sink别名:k1*  *a1.sinks = k1*  *#channel别名：c1*  *a1.channels = c1*  *############################################################*  *# 2、source参数配置：*  *############################################################*  *#source类型：*  *a1.sources.r1.type = spooldir*  *#指定监听目录：*  *a1.sources.r1.spoolDir = /home/logs*  *#文件头自动生成*  *a1.sources.r1.fileHeader = true*  *############################################################*  *# 3、channel参数配置：*  *############################################################*  *#channel类型：*  *a1.channels.c1.type = memory*  *#默认该通道中最大的可以存储的event数量是100*  *a1.channels.c1.capacity = 1000*  *#每次最大可以source中拿到或者送到sink中的event数量也是100*  *a1.channels.c1.transactionCapacity = 100*  *############################################################*  *# 4、sink参数配置：*  *############################################################*  *#sink类型：以日志的方式输出*  *a1.sinks.k1.type = logger*    *############################################################*  *# 5、集成一个agent当中的各个角色：*  *############################################################*  *a1.sources.r1.channels = c1*  *a1.sinks.k1.channel = c1* |

执行脚本：

|  |
| --- |
| 1. 执行脚本：   *flume-ng agent -n a1 -f flumeTest03.conf -D flume.root.logger=INFO,console*  2、拷贝文件：mkdir logs cp test.log logs/ |

### 案例4、实时读取本地文件到HDFS

编写flume脚本：flumeTest04.flume

|  |
| --- |
| *############################################################*  *# 1、命名：一个agent（进程）内各角色别名：a1*  *############################################################*  *a2.sources = r2*  *a2.sinks = k2*  *a2.channels = c2*  *############################################################*  *# 2、source参数配置：*  *############################################################*  *a2.sources.r2.type = exec*  *a2.sources.r2.command = tail -F /home/admin/modules/apache-hive-1.2.2-bin/hive.log*  *a2.sources.r2.shell = /bin/bash -c*  *############################################################*  *# 3、channel参数配置：*  *############################################################*  *a2.channels.c2.type = memory*  *a2.channels.c2.capacity = 1000*  *a2.channels.c2.transactionCapacity = 100*  *############################################################*  *# 4、sink参数配置：*  *############################################################*  *a2.sinks.k2.type = hdfs*  *a2.sinks.k2.hdfs.path = hdfs://linux01:8020/flume/%Y%m%d/%H*  *#上传文件的前缀*  *a2.sinks.k2.hdfs.filePrefix = logs-*  *#是否按照时间滚动文件夹*  *a2.sinks.k2.hdfs.round = true*  *#多少时间单位创建一个新的文件夹*  *a2.sinks.k2.hdfs.roundValue = 1*  *#重新定义时间单位*  *a2.sinks.k2.hdfs.roundUnit = hour*  *#是否使用本地时间戳*  *a2.sinks.k2.hdfs.useLocalTimeStamp = true*  *#积攒多少个Event才flush到HDFS一次*  *a2.sinks.k2.hdfs.batchSize = 1000*  *#设置文件类型，可支持压缩*  *a2.sinks.k2.hdfs.fileType = DataStream*  *#多久生成一个新的文件*  *a2.sinks.k2.hdfs.rollInterval = 8*  *#设置每个文件的滚动大小*  *a2.sinks.k2.hdfs.rollSize = 134217700*  *#文件的滚动与Event数量无关*  *a2.sinks.k2.hdfs.rollCount = 0*  *#最小冗余数*  *a2.sinks.k2.hdfs.minBlockReplicas = 1*  *############################################################*  *# 5、集成一个agent当中的各个角色：*  *############################################################*  *a2.sources.r2.channels = c2*  *a2.sinks.k2.channel = c2* |

执行脚本：

|  |
| --- |
| *[root@node1 case]# flume-ng agent -n a2 -f flumeTest04.conf -D flume.root.logger=INFO,console* |