> Flume 是一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；flume具有高可用，分布式，配置工具，其设计的原理也是基于将数据流，如日志数据从各种网站服务器上汇集起来存储到HDFS,HBase等集中存储器中。

Flume安装

安装Flume

将 Flume 上传至新建的目录 `/opt/software/` 目录下，并解压至 `/usr/local/src/` 目录下。

配置环境变量

```
export FLUME_HOME=usr/local/src/flink
export PATH=$PATH:$FLUME_HOME/bin
```

Flume 配置

> Flume 包含Source、Channel、Sink三个组件，Source用来接收数据，Channel用来缓存数据，Sink用来发送数据，而且是主动push给下游，这就导致下游接收方只能是一个，因为如果下游有多个接收者，接收的速率不同就会导致接收速度低的接收者接收不到数据的情况(Channel会在Sink确认后删除数据)。如果想把数据发送给多个接收者，那就只能让Source把数据写到多个Channel，再由Channel经各自的Sink发送给不同的接收者。

新增配置文件

Flume 配置目录：

```
$FLUME_HOME/conf
```

新建配置文件 `config.conf` （此处文件名可自定义）并进行并进行编写：

```
# 当前定义当前agent的名字叫a1，可以随意设置

# 定义sources的名字为r1
a1.sources=r1
# 定义channel的名字为c1
a1.channels=c1
# 定义sinks的名字为k1
a1.sinks=k1

# -------- (1)使用端口接收数据 --------
# 定义sources的类型为网络字节流（从端口接收数据）
a1.sources.r1.type=netcat
# 设置监听主机
a1.sources.r1.bind=master
```

```
# 设置监听端口号
a1.sources.r1.port=26001

# -------- (2)使用文件接收数据 --------
# 定义sources的类型为命令输出流
a1.sources.r1.type=exec
# 设置消息来源为读取文件末尾
a1.sources.command.tail -F /opt/flume.txt
# ----------------------------------

# 设置使用内存作为缓存
a1.channels.c1.type=memory

# -------- (1)内容输出到日志（控制台） --------
# 设置Sink输出到控制台
a1.sinks.k1.type=logger

# -------- (2)内容输出到Kafka ---------------
# 设置Sink输出到Kafka
a1.sinks.k1.type=org.apache.flume.sink.kafka.KafkaSink
a1.sinks.k1.kafka.bootstrap.servers=master:9092,slave1:9092,slave2:9092
a1.sinks.k1.kafka.topic=[主题名称]

# 关联三个组件
# 设置r1的内容输出至c1
a1.sources.r1.channels=c1
# 设置k1的内容来源为c1
a1.sinks.k1.channel=c1
```

启动Flume

使用 `flume-ng agent` 命令启动Flume。

（-c：使用配置文件，-f：配置文件目录，--name：配置文件中agent的命名）

```
flume-ng agent -c conf -f /usr/local/src/flume/conf/config.conf --name a1 -
Dflume.root.logger=INFO,console
```

Flume使用

测试netcat：

使用其他机器终端，用nc命令连接。

```
nc 主机名 端口号
```

测试exec：

使用echo命令向指定文件追加内容。

```
echo "内容" >> 文件名
```

检测端口是否被占用：

```
netstat -tunlp | grep 44444
```