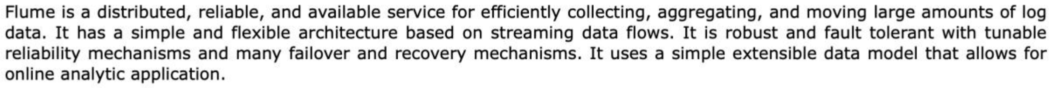
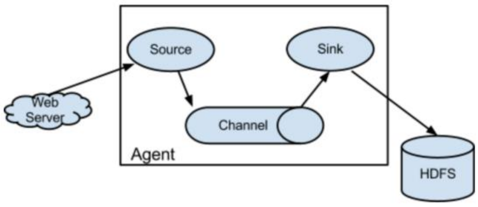
Flume

- ## 概述



Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.



- ## 环境搭建

tar zxf apache-flume-1.6.0-bin.tar.gz -C /opt

cp conf/ flume-env.sh.template flume-env.sh

配置 JAVA_HOME

bin/flume-ng

- ## 简单案例

telnet 安装

上传 telnet rpm 包

rpm -ivh *.rpm

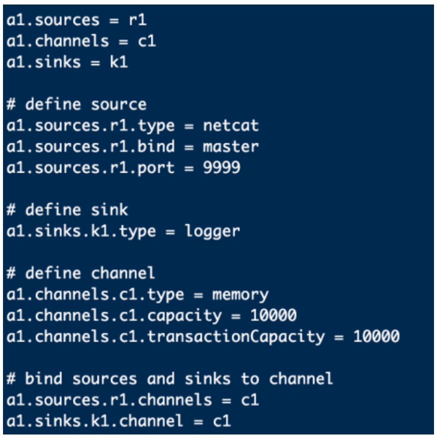
/etc/rc.d/init.d/xinetd start

telnet host port

配置 netcat agent

cp flume-conf.properties.template a1.conf

vi a1.conf



```
a1.sources = r1
a1.channels = c1
a1.sinks = k1

# define source
a1.sources.r1.type = netcat
a1.sources.r1.bind = master
a1.sources.r1.port = 9999

# define sink
a1.sinks.k1.type = logger

# define channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 10000
a1.channels.c1.transactionCapacity = 10000

# bind sources and sinks to channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

启动 agent

```
bin/flume-ng agent --name a1 --conf conf --conf-file conf/a1.conf -
Dflume.root.logger=DEBUG,console
```

启动 telnet

```
telnet master 9999
```

- ## 案例实战

hive sources(实时采集 hive 日志)

```
hive.sources = r2
hive.channels = c2
hive.sinks = k2

# define sources
hive.sources.r2.type = exec
hive.sources.r2.command = tail -f /opt/modules/apache-hive-1.2.1-bin/log/hive.log
hive.sources.r2.shell = /bin/bash -c
hive.sources.r2.logStdErr = false
hive.sources.r2.channels = c2

# define channels
hive.channels.c2.type = memory
hive.channels.c2.capacity = 1000
hive.channels.c2.transactionCapacity = 100

# define sinks
hive.sinks.k2.type = hdfs
hive.sinks.k2.hdfs.path = hdfs://master:9000/user/root/flume-hive
hive.sinks.k2.hdfs.fileType = DataStream
hive.sinks.k2.hdfs.writeFormat = Text
hive.sinks.k2.hdfs.round = true
hive.sinks.k2.hdfs.minBlockReplicas=1
hive.sinks.k2.hdfs.roundValue=5
hive.sinks.k2.hdfs.roundUnit=minute
hive.sinks.k2.hdfs.rollInterval=30
hive.sinks.k2.hdfs.rollSize = 0
hive.sinks.k2.hdfs.rollCount = 0
hive.sinks.k2.channel = c2
```

- ## HDFS Sink

hdfs.filePrefix 写入 hdfs 的文件名前缀，可以使用 flume 提供的日期及%{host}表达式

hdfs.fileSuffix 写入 hdfs 的文件名后缀，比如：.lzo .log 等

hdfs.rollInterval hdfs sink 间隔多长将临时文件滚动成最终目标文件，单位：秒

hdfs.rollSize 当临时文件达到该大小（单位：bytes）时，滚动成目标文件 如果设置成 0，则表示不根据临时文件大小来滚动文件

hdfs.rollCount 当 events 数据达到该数量时候，将临时文件滚动成目标文件

如果设置成 0，则表示不根据 events 数据来滚动文件

hdfs.idleTimeout 当目前被打开的临时文件在该参数指定的时间（秒）内，没有任何数据写入，则将该临时文件关闭并重命名成目标文件

hdfs.batchSize 每个批次刷新到 HDFS 上的 events 数量

hdfs.codeC 文件压缩格式，包括：gzip, bzip2, lzo, lzop, snappy

hdfs.fileType 文件格式，包括：SequenceFile, DataStream,CompressedStream

hdfs.maxOpenFiles 最大允许打开的 HDFS 文件数，当打开的文件数达到该值，最早打开的文件将会被关闭

hdfs.minBlockReplicas 写入 HDFS 文件块的最小副本数，该参数会影响文件的滚动配置，一般将该参数配置成 1，才可以按照配置正确滚动文件

hdfs.writeFormat 写 sequence 文件的格式。包含：Text, Writable（默认）

hdfs.callTimeout 执行 HDFS 操作的超时时间（单位：毫秒）

hdfs.threadsPoolSize hdfs sink 启动的操作 HDFS 的线程数。

hdfs.round　　是否启用时间上的”舍弃”，这里的”舍弃”，类似于”四舍五入”。如果启用，则会影响除了%t 的其他所有时间表达式

hdfs.roundValue 时间上进行“舍弃”的值

hdfs.roundUnit　时间上进行”舍弃”的单位，包含：second,minute,hour

hdfs.retryInterval hdfs sink 尝试关闭文件的时间间隔，如果设置为 0，表示不尝试，相当于于将 hdfs.closeTries 设置成 1.

数据分区

```
# define sinks
hive.sinks.k2.type = hdfs
hive.sinks.k2.hdfs.useLocalTimeStamp = true
hive.sinks.k2.hdfs.path = hdfs://master:9000/user/root/flume-hive/%Y/%m/%d
hive.sinks.k2.hdfs.fileType = DataStream
hive.sinks.k2.hdfs.writeFormat = Text
hive.sinks.k2.hdfs.rollInterval=30
hive.sinks.k2.hdfs.rollSize = 0
hive.sinks.k2.hdfs.rollCount = 0
hive.sinks.k2.channel = c2
```

- Spooling Directory Source

1. 监控目录
2. FileChannel
3. HDFS Store

```
spool.sources = r2
spool.channels = c2
spool.sinks = k2

# define sources
spool.sources.r2.type = spooldir
spool.sources.r2.spoolDir = /opt/modules/apache-flume-1.6.0-bin/spool_data
spool.sources.r2.fileHeader = true
spool.sources.r2.ignorePattern = ^(.)*\\.log$
spool.sources.r2.channels = c2

# define channels
spool.channels.c2.type = file
spool.channels.c2.checkpointDir = /opt/modules/apache-flume-1.6.0-bin/filechannel/checkpoint
spool.channels.c2.dataDirs = /opt/modules/apache-flume-1.6.0-bin/filechannel/data

# define sinks
spool.sinks.k2.type = hdfs
spool.sinks.k2.hdfs.useLocalTimeStamp = true
spool.sinks.k2.hdfs.path = hdfs://master:9000/user/root/flume-spool/%Y-%m-%d
spool.sinks.k2.hdfs.fileType = DataStream
spool.sinks.k2.hdfs.writeFormat = Text
spool.sinks.k2.hdfs.rollInterval=30
spool.sinks.k2.hdfs.rollSize = 0
spool.sinks.k2.hdfs.rollCount = 0
spool.sinks.k2.channel = c2
```