实验一：多层 Flume 收集数据

一、实验介绍

1、about

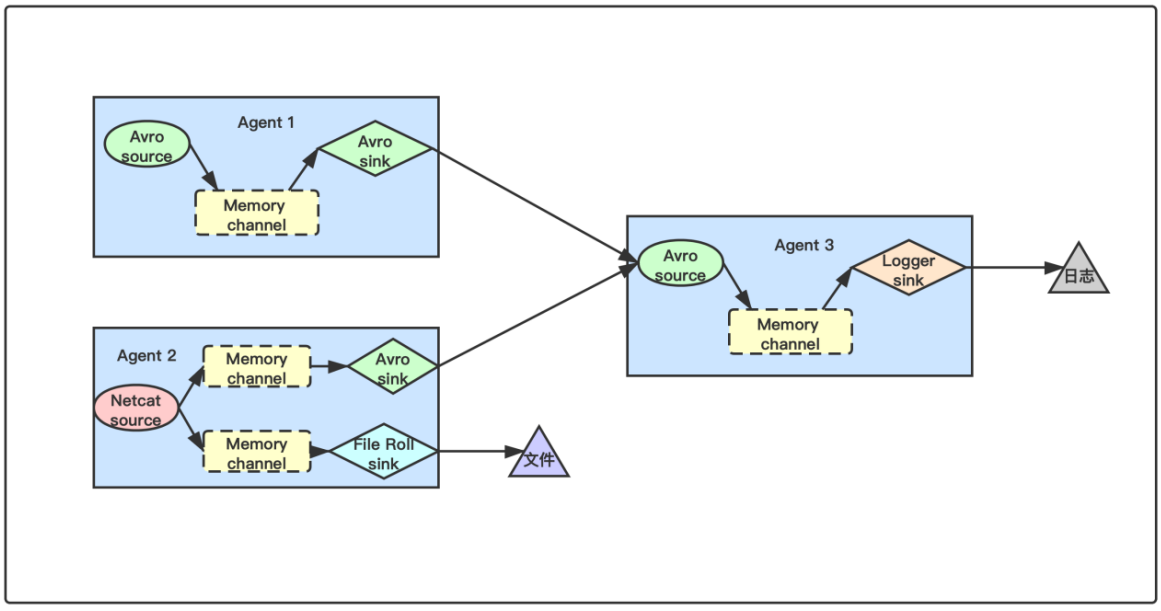
> 本实验在 Flume 配置实验 的基础之上，通过搭建多层 Flume 网络进行数据收集任务，帮助同学们掌握 Flume Agent 的组成结构、各组件的功能及灵活组合 Flume Agent 实现多层数据采集。

2、实验目的

- 了解 Flume agent 的结构组成及各组件功能。
- 了解 Flume agent 组件 source 、 channel 和 source 的不同类型及其功能和应用场景。
- 掌握多层 Flume 拓扑网络的配置和使用，实现数据收集任务。

二、实验内容

通过编写配置文件，实现如下图所示的 Flume 拓扑网络结构



其中，各 Flume Agent 的组件构成为：

Agent1

组件	类型	备注
Source	Avro Source	收集 Avro 客户端发来的数据转换为 Event 保存到 Channel
Channel	Memory Channel	将 Source 传来的 Event 保存于内存中，等待 Sink 取走
Sink	Avro Sink	将 Event 发送到指定的主机和端口

Agent2

组件	类型	备注
Source	Netcat Source	监听指定端口，将从该端口收集到的 TCP 协议文本数据 转换为 Event 保存到 Channel 中
Channel	Memory Channel	将 Source 传来的 Event 保存于内存中，等待 Sink 取走
Sink	Avro Sink	将 Event 发送到指定的主机和端口
Sink	File Roll Sink	将 Event 以文件形式保存到指定目录

Agent3

组件	类型	备注
Source	Avro Source	收集 Avro 客户端发来的数据转换为 Event 保存到 Channel
Channel	Memory Channel	将 Source 传来的 Event 保存于内存中，等待 Sink 取走
Sink	Logger Sink	将 Event 内容输出到日志中（可在客户端看到）

其中， Agent 1 和 Agent 2 可以收集不同类型的数据并将收集到的数据发送给Agent 3， Agent 3 可将收集到的数据输出到屏幕日志中以显示，此外，Agent 2 还会将自己收集到的数据保存一份文件副本。

三、实验步骤

1、实验环境

- CentOS Linux release 7.8.2003 (Core)
- oracle jdk version "1.8.0_341"
- apache-flume-1.9.0-bin.tar.gz

2、实验准备

2.1 修改主机名：主机名格式：学号+姓名简称

2.2 在用户目录下新建文件夹，用于保存本次实验文件

```
mkdir -p ~/experiment/ex_log
```

2.3 安装 flume 环境（略）

已在之前的实验中提及，本次实验可以直接在上次的实验环境中接着做

2.4 安装 netcat

```
yum install -y nc
```

3、Flume 配置

Agent 1

在 `${FLUME_HOME}/conf` 目录下编写配置文件 `a1.conf` ，内容如下：

```
在 ${FLUME_HOME}/conf 目录下编写配置文件 a1.conf ，内容如下：a1.sources = r1

a1.sinks = k1

a1.channels = c1


a1.sources.r1.channels = c1

a1.sinks.k1.channel = c1


a1.sources.r1.type = avro

a1.sources.r1.bind = 0.0.0.0

a1.sources.r1.port = 10000

a1.sources.r1.threads = 5


a1.sinks.k1.type = avro

a1.sinks.k1.hostname = 0.0.0.0

a1.sinks.k1.port = 10002


a1.channels.c1.type = memory

a1.channels.c1.capacity = 1000
```

```
a1.channels.c1.transactionCapacity = 100
```

Agent 2

在 `${FLUME_HOME}/conf` 目录下编写配置文件 `a2.conf` ，内容如下：（文件目录名注意配置）

```
a2.sources = r1

a2.sinks = k1 k2

a2.channels = c1 c2


a2.sources.r1.channels = c1 c2

a2.sources.r1.selector.type = replicating


a2.sinks.k1.channel = c1

a2.sinks.k2.channel = c2


# 组件配置

a2.sources.r1.type = netcat

a2.sources.r1.bind = 0.0.0.0

a2.sources.r1.port = 10001


a2.sinks.k1.type = avro

a2.sinks.k1.hostname = 0.0.0.0

a2.sinks.k1.port = 10002


a2.sinks.k2.type = file_roll

# 这里的文件名注意配置

a2.sinks.k2.sink.directory = /root/experiment/ex_log

a2.sinks.k2.sink.rollInterval = 0


a2.channels.c1.type = memory

a2.channels.c1.capacity = 1000

a2.channels.c1.transactionCapacity = 100


a2.channels.c2.type = memory

a2.channels.c2.capacity = 1000
```

```
a2.channels.c2.transactionCapacity = 100
```

Agent 3

在 `${FLUME_HOME}/conf` 目录下编写配置文件 `a3.conf` ，内容如下：

```
# Agent a3内部的数据流
a3.sources = r1
a3.sinks = k1
a3.channels = c1

a3.sources.r1.channels = c1
a3.sinks.k1.channel = c1

# 组件配置
a3.sources.r1.type = avro
a3.sources.r1.bind = 0.0.0.0
a3.sources.r1.port = 10002
a3.sources.r1.threads = 5

a3.sinks.k1.type = logger

a3.channels.c1.type = memory
a3.channels.c1.capacity = 1000
a3.channels.c1.transactionCapacity = 100
```

4、开始数据收集

4.1 在新的终端 ${FLUME_HOME} 目录下启动3个 Flume Agent

> **注意启动次序要求： a3 早于 a1 和 a2 ，建议启动次序 a3 -> a2 -> a1 【重要】**

```
flume-ng agent --conf conf --conf-file conf/a3.conf --name a3 -Dflume.root.logger=INFO,console
flume-ng agent --conf conf --conf-file conf/a2.conf --name a2 -Dflume.root.logger=INFO,console
flume-ng agent --conf conf --conf-file conf/a1.conf --name a1 -Dflume.root.logger=INFO,console
```

4.2 新的终端中用户目录下新建文件 a1_input.txt 并写入信息，并使用 avro-client 向 a1 监听的 avro 服务发送文件

```
flume-ng avro-client -c ${FLUME_HOME}/conf/ -H 0.0.0.0 -p 10000 -F ~/a1-input.txt
```

`a1-input.txt` 的格式为

```
[学号]

[姓名]

[ip]
```

4.3 新的终端中启动 netcat 服务，向 a2 监听的 netcat 服务发送消息。

```
curl telnet://localhost:10001

# 或者用 nc 命令

nc localhost 10001
```

消息内容必须包含学号

4.4 查看数据收集结果

```
cat /home/zkpk/experiment/ex_log/*
```

四、附加实验

设计并实现更复杂的 Flume 拓扑网络，比如三层或更多的网络，并画出你的数据是如何在 Flume 组件间流动的。