# Flume

## Flume介绍

Flume是Apache基金会组织的一个提供的高可用的，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；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。

当前Flume有两个版本，Flume 0.9x版本之前的统称为Flume-og，Flume1.X版本被统称为Flume-ng。

## Flume-og和Flume-ng的区别

主要区别如下：

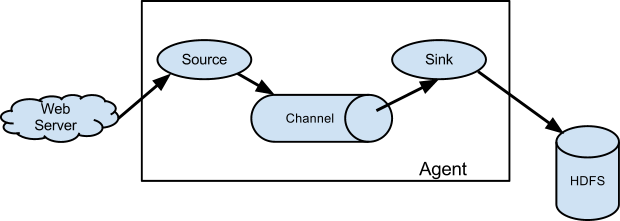
1. Flume-og中采用master结构，为了保证数据的一致性，引入zookeeper进行管理。Flume-ng中取消了集中master机制和zookeeper管理机制，变成了一个纯粹的传输工具。

2. Flume-ng中采用不同的线程进行数据的读写操作；在Flume-og中，读数据和写数据是由同一个线程操作的，如果写出比较慢的话，可能会阻塞flume的接收数据的能力。

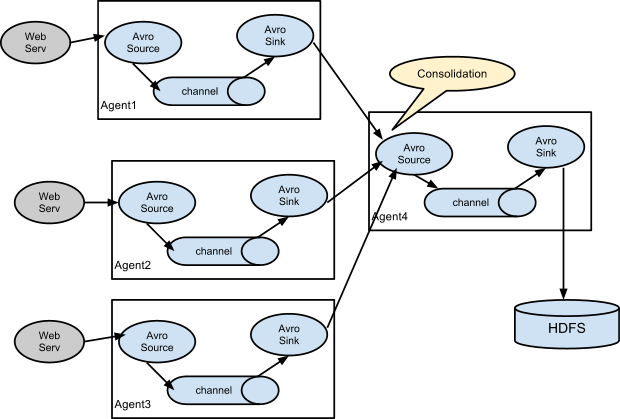
## Flume结构

Flume中以Agent为基本单位，一个agent可以包括source、channel、sink，三种组件都可以有多个。其中source组件主要功能是接收外部数据，并将数据传递到channel中；sink组件主要功能是发送flume接收到的数据到目的地；channel的主要作用就是数据传输和保存的一个作用。Flume主要分为三类结构：单agent结构、多agent链式结构和多路复用agent结构。

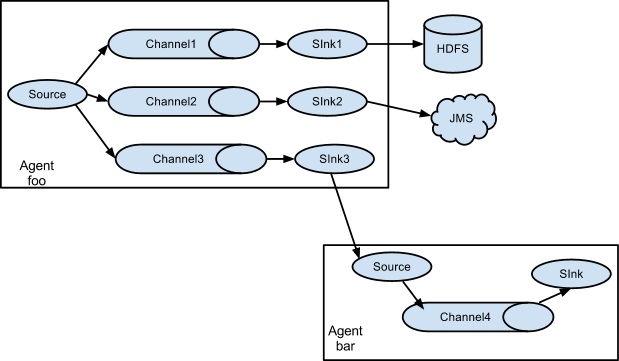
## 单agent结构



## 多agent链式结构



## 多路复用agent结构



## Source介绍

Source的主要作用是接收客户端发送的数据，并将数据发送到channel中，source和channel之间的关系是多对多关系，不过一般情况下使用一个source对应多个channel。通过名称区分不同的source。Flume常用source有：Avro Source、Thrift Source、Exec Source、Kafka Source、Netcat Source等。设置格式如下：

<agent-name>.sources=source\_names

<agent-name>.sources.<source\_name>.type=指定类型

<agent-name>.sources.<source\_name>.channels=channels

.... 其他对应source类型需要的参数

## Channel介绍

Channel的主要作用是提供一个数据传输通道，提供数据传输和数据存储(可选)等功能。source将数据放到channel中，sink从channel中拿数据。通过不同的名称来区分channel。Flume常用channel有：Memory Channel、JDBC Channel、Kafka Channel、File Channel等。设置格式如下：

<agent-name>.channels=channel\_names

<agent-name>.channels.<channel\_name>.type=指定类型

.... 其他对应channel类型需要的参数

## Sink介绍

Sink的主要作用是定义数据写出方式，一般情况下sink从channel中获取数据，然后将数据写出到file、hdfs或者网络上。channel和sink之间的关系是一对多的关系。通过不同的名称来区分sink。Flume常用sink有：Hdfs Sink、Hive Sink、File Sink、HBase Sink、Avro Sink、Thrift Sink、Logger Sink等。设置格式如下：

<agent-name>.sinks = sink\_names

<agent-name>.sinks.<sink\_name1>.type=指定类型

<agent-name>.sinks.<sink\_name1>.channel=<channe\_name>

.... 其他对应sink类型需要的参数

## Flume安装

安装步骤如下：

1. 下载flume：wget http://archive.cloudera.com/cdh5/cdh/5/flume-ng-1.5.0-cdh5.3.6.tar.gz

2. 解压flume。

3. 修改conf/flume-env.sh文件，如果没有就新建一个。

4. 添加flume的bin目录到环境变量中去。

5. 验证是否安装成功, flume-ng version