目录

[1. Flume 采集数据会丢失吗?](#_Toc2415_WPSOffice_Level1) [1](#_Toc2415_WPSOffice_Level1)

[2. Flume 与 Kafka 的选取？](#_Toc24056_WPSOffice_Level1) [1](#_Toc24056_WPSOffice_Level1)

[3. 数据怎么采集到 Kafka，实现方式？](#_Toc5148_WPSOffice_Level1) [2](#_Toc5148_WPSOffice_Level1)

[4. flume 管道内存，flume 宕机了数据丢失怎么解决？](#_Toc24037_WPSOffice_Level1) [2](#_Toc24037_WPSOffice_Level1)

[5. flume 和 kafka 采集日志区别，采集日志时中间停了，怎么记录之前的日志？](#_Toc4383_WPSOffice_Level1) [2](#_Toc4383_WPSOffice_Level1)

[6. flume 有哪些组件，flume 的 source、channel、sink 具体是做什么的？](#_Toc1568_WPSOffice_Level1) [3](#_Toc1568_WPSOffice_Level1)

[7. 为什么使用Flume？](#_Toc31165_WPSOffice_Level1) [3](#_Toc31165_WPSOffice_Level1)

[8. Flume组成架构？](#_Toc27383_WPSOffice_Level1) [4](#_Toc27383_WPSOffice_Level1)

[9. FlumeAgent内部原理？](#_Toc10058_WPSOffice_Level1) [4](#_Toc10058_WPSOffice_Level1)

[10.Flume Event 是数据流的基本单元](#_Toc9660_WPSOffice_Level1) [5](#_Toc9660_WPSOffice_Level1)

[11. Flume agent](#_Toc9689_WPSOffice_Level1) [5](#_Toc9689_WPSOffice_Level1)

[12. 你是如何实现Flume数据传输的监控的](#_Toc12824_WPSOffice_Level1) [8](#_Toc12824_WPSOffice_Level1)

[13. flume 调优 ：](#_Toc10949_WPSOffice_Level1) [8](#_Toc10949_WPSOffice_Level1)

### Flume 采集数据会丢失吗?

不会，Channel 存储可以存储在 File 中，数据传输自身有事务。

### **Flume 与 Kafka 的选取？**

采集层主要可以使用 Flume、Kafka 两种技术。

Flume：Flume 是管道流方式，提供了很多的默认实现，让用户通过参数部署，及扩展 API。

Kafka：Kafka 是一个可持久化的分布式的消息队列。

Kafka 是一个非常通用的系统。你可以有许多生产者和很多的消费者共享多个主题Topics。相比之下，Flume 是一个专用工具被设计为旨在往 HDFS，HBase 发送数据。它对HDFS 有特殊的优化，并且集成了 Hadoop 的安全特性。所以，Cloudera 建议如果数据被多个系统消费的话，使用 kafka；如果数据被设计给 Hadoop 使用，使用 Flume。正如你们所知 Flume 内置很多的 source 和 sink 组件。然而，Kafka 明显有一个更小的生产消费者生态系统，并且 Kafka 的社区支持不好。希望将来这种情况会得到改善，但是目前：使用 Kafka 意味着你准备好了编写你自己的生产者和消费者代码。如果已经存在的 Flume Sources 和 Sinks 满足你的需求，并且你更喜欢不需要任何开发的系统，请使用 Flume。Flume 可以使用拦截器实时处理数据。这些对数据屏蔽或者过量是很有用的。Kafka 需要外部的流处理系统才能做到。

Kafka 和 Flume 都是可靠的系统,通过适当的配置能保证零数据丢失。然而，Flume 不支持副本事件。于是，如果 Flume 代理的一个节点奔溃了，即使使用了可靠的文件管道方式，你也将丢失这些事件直到你恢复这些磁盘。如果你需要一个高可靠行的管道，那么使用Kafka 是个更好的选择。

Flume 和 Kafka 可以很好地结合起来使用。如果你的设计需要从 Kafka 到 Hadoop 的流数据，使用 Flume 代理并配置 Kafka 的 Source 读取数据也是可行的：你没有必要实现自己的消费者。你可以直接利用Flume 与HDFS 及HBase 的结合的所有好处。你可以使用ClouderaManager 对消费者的监控，并且你甚至可以添加拦截器进行一些流处理。

### **数据怎么采集到 Kafka，实现方式？**

使用官方提供的 flumeKafka 插件，插件的实现方式是自定义了 flume 的 sink，将数据从channle 中取出，通过 kafka 的producer 写入到 kafka 中，可以自定义分区等。

### **flume 管道内存，flume 宕机了数据丢失怎么解决？**

1）Flume 的 channel 分为很多种，可以将数据写入到文件。

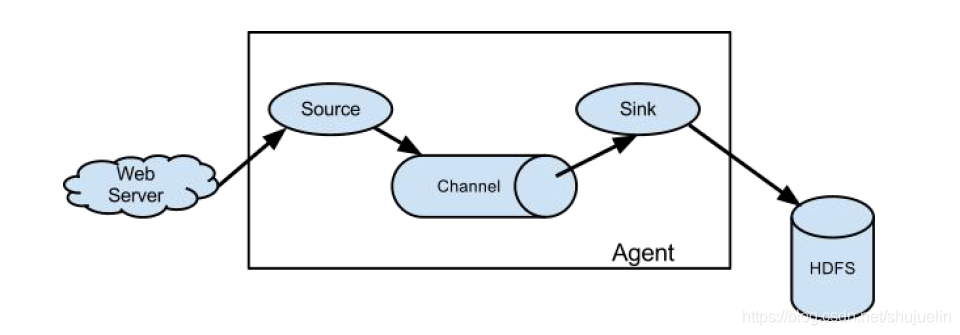
2）防止非首个 agent 宕机的方法数可以做集群或者主备

### **flume 和 kafka 采集日志区别，采集日志时中间停了，怎么记录之前的日志？**

Flume 采集日志是通过流的方式直接将日志收集到存储层，而 kafka 是将缓存在 kafka集群，待后期可以采集到存储层。

Flume 采集中间停了，可以采用文件的方式记录之前的日志，而 kafka 是采用 offset 的方式记录之前的日志。

### **flume 有哪些组件，flume 的 source、channel、sink 具体是做什么的？**



1）source：用于采集数据，Source 是产生数据流的地方，同时 Source 会将产生的数据

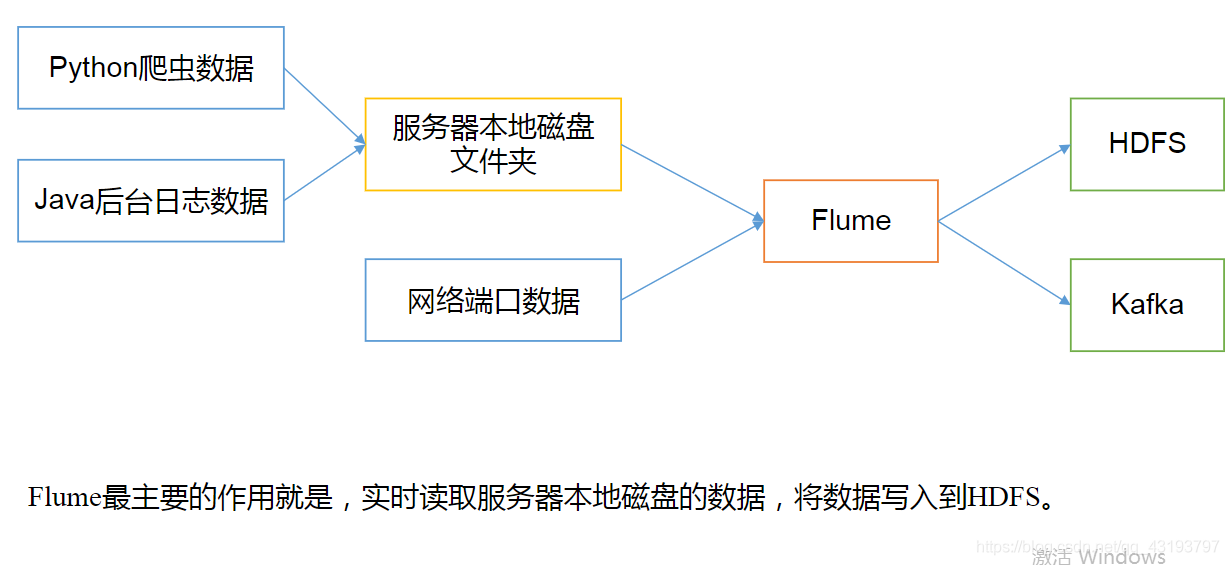
流传输到 Channel，这个有点类似于 Java IO 部分的 Channel。

2）channel：用于桥接 Sources 和 Sinks，类似于一个队列。

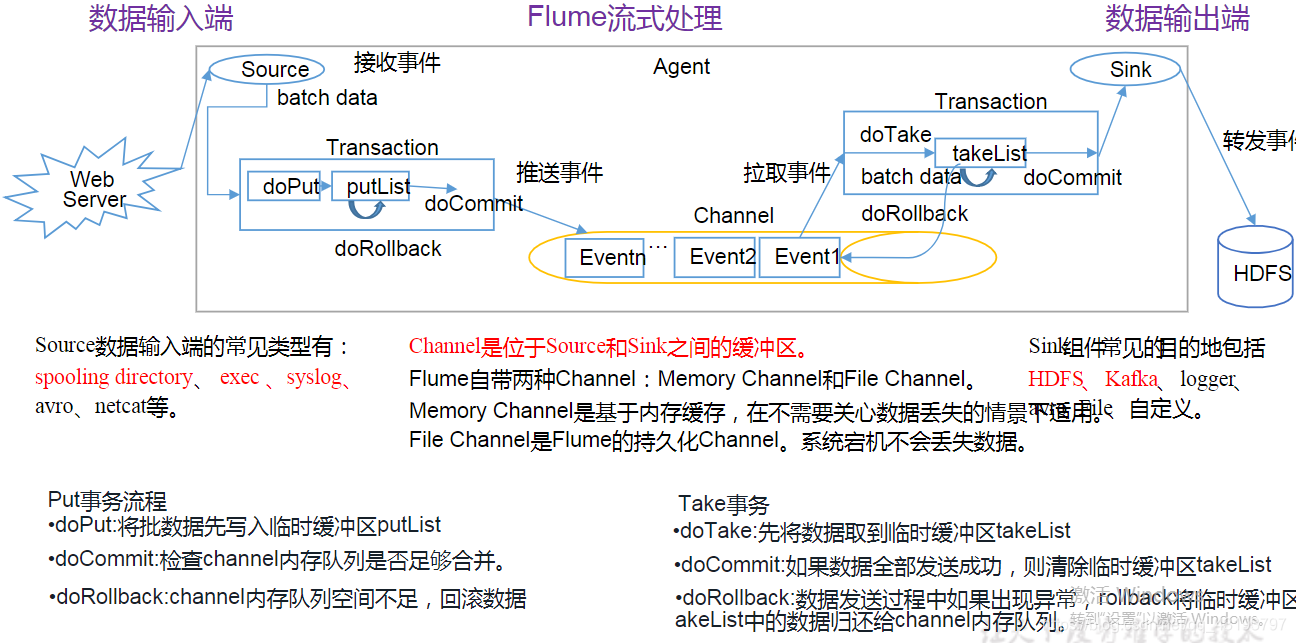
3）sink：从 Channel 收集数据，将数据写到目标源(可以是下一个 Source，也可以是 HDFS

或者 HBase)。

### **为什么使用Flume？**



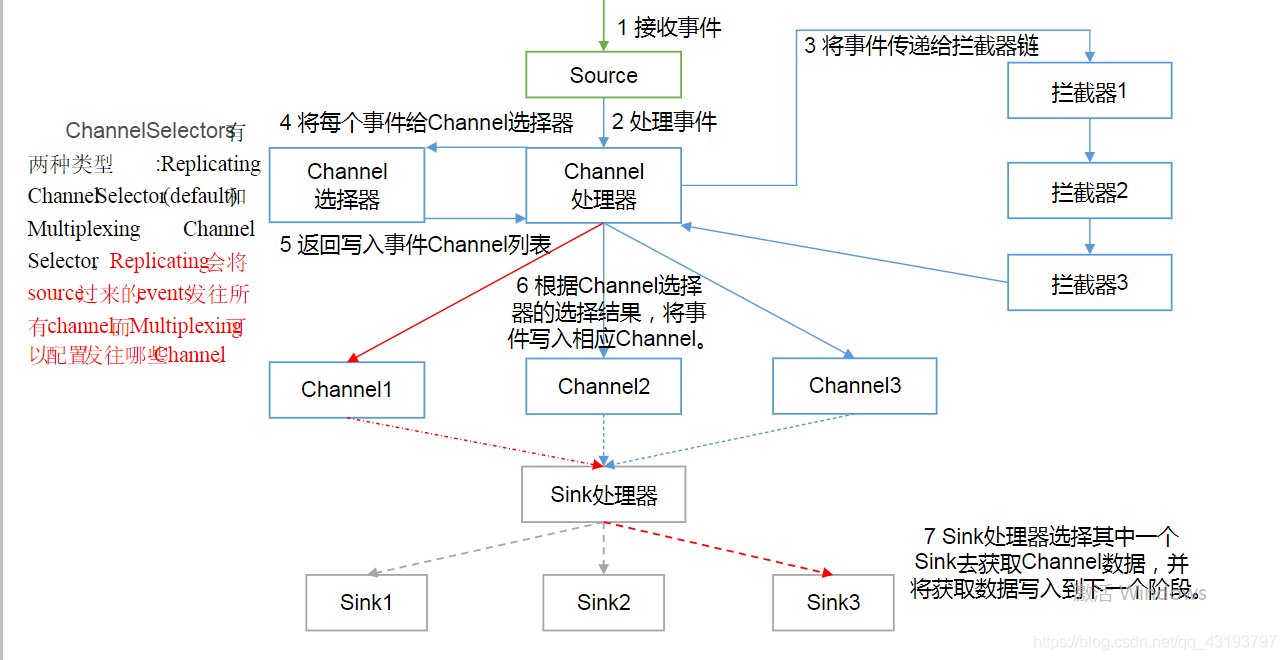
### **Flume组成架构？**



关于flume事务

flume要尽可能的保证数据的安全性，其在source推送数据到channel以及sink从channel拉取数据时都是以事务方式进行的。因为在agent内的两次数据传递间都会涉及到数据的传送、从数据上游删除数据的问题；就比如sink从channel拉取数据并提交到数据下游之后需要从channel中删除已获取到的批次数据，其中跨越了多个原子事件，故而需要以事务的方式将这些原子事件进一步绑定在一起，以便在其中某个环节出错时进行回滚防止数据丢失。所以在选用file channel时一般来说是不会丢失数据的。

### **FlumeAgent内部原理？**



### **Flume Event 是数据流的基本单元**

它由一个装载数据的字节数组(byte payload)和一系列可选的字符串属性来组成(可选头部).



### **Flume agent**

Agent 是一个 JVM 进程，它以事件的形式将数据从源头送至目的。  
Agent 主要有 3 个部分组成， Source、 Channel、 Sink

Flume source 消耗从类似于 web 服务器这样的外部源传来的 events.

外部数据源以一种 Flume source 能够认识的格式发送 event 给 Flume source.

Flume source 组件可以处理各种类型、各种格式的日志数据，包括 avro、thrift、exec、jms、spooling directory、netcat、sequence generator、syslog、http、legacy.

flume source 是负责接收数据到 Flume Agent 的组件



Flume channel

当 Flume source 接受到一个 event 的时, Flume source 会把这个 event 存储在一个或多个 channel 中.

Channel 是连接Source和Sink的组件, 是位于 Source 和 Sink 之间的数据缓冲区。

Flume channel 使用被动存储机制. 它存储的数据的写入是靠 Flume source 来完成的, 数据的读取是靠后面的组件 Flume sink 来完成的.

Channel 是线程安全的，可以同时处理几个 Source 的写入操作和几个 Sink 的读取操作。

Flume 自带两种 Channel：

Memory Channel

Memory Channel是内存中的队列。

Memory Channel在不需要关心数据丢失的情景下适用。

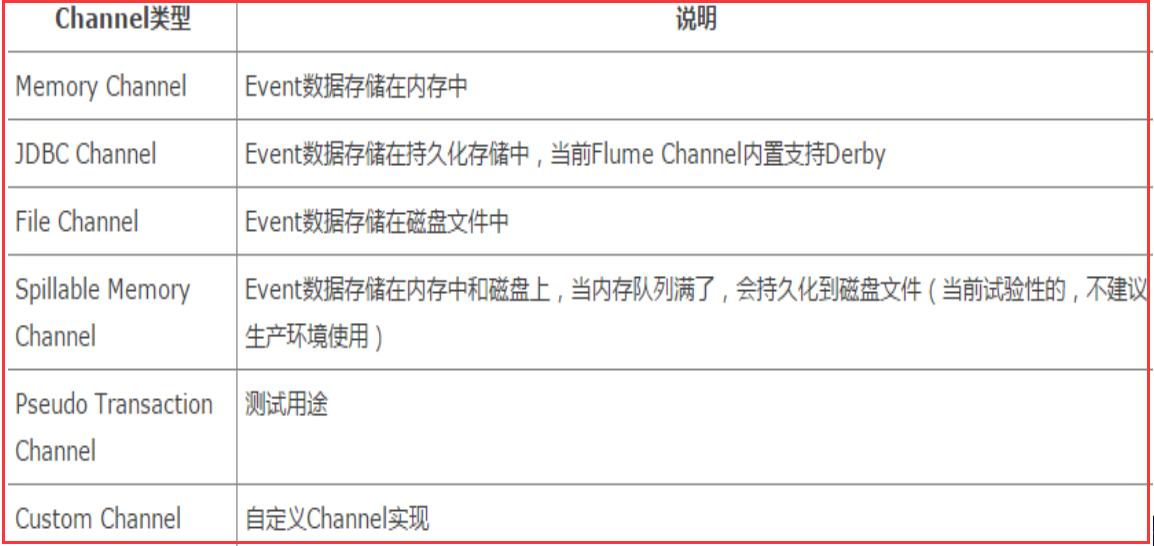
如果需要关心数据丢失，那么Memory Channel就不应该使用，因为程序死亡、机器宕机或者重启都会导致数据丢失。

File Channel。

File Channel将所有事件写到磁盘。

因此在程序关闭或机器宕机的情况下不会丢失数据。

还可以有其他的 channel: 比如 JDBC channel.



Flume sink

Sink 不断地轮询 Channel 中的事件且批量地移除它们，并将这些事件批量写入到存储或索引系统、或者发送到另一个Flume Agent。

Sink 是完全事务性的。

~~在从 Channel 批量删除数据之前，每个 Sink 用 Channel 启动一个事务。批量事件一旦成功写出到存储系统或下一个Flume Agent，Sink 就利用 Channel 提交事务。事务一旦被提交，该 Channel 从自己的内部缓冲区删除事件。如果写入失败，将缓冲区takeList中的数据归还给Channel。~~

Sink组件目的地包括hdfs、logger、avro、thrift、ipc、file、null、HBase、solr、自定义。



### **你是如何实现Flume数据传输的监控的**

使用第三方框架Ganglia实时监控Flume。

### **flume 调优**

source ：

1 ，增加 source 个数，可以增大 source 读取能力。

具体做法 ： 如果一个目录下生成的文件过多，可以将它拆分成多个目录。每个目录都配置一个 source 。

2 ，增大 batchSize ： 可以增大一次性批处理的 event 条数，适当调大这个参数，可以调高 source 搬运数据到 channel 的性能。

channel ：

4 ，memory ：性能好，但是，如果发生意外，可能丢失数据。

5 ，使用 file channel 时，dataDirs 配置多个不同盘下的目录可以提高性能。

a2.channels.c20.type = file

a2.channels.c20.checkpointDir = /export/flume/realtime/file\_channel20/checkpoint

#存放检查点目录，checkpointDir是一个目录

a2.channels.c20.dataDirs = /export/flume/realtime/file\_channel20/data

#存放数据的目录，dataDirs可以是多个目录，以逗号隔开。用独立的多个磁盘上的多个目录可以提高file channel的性能

————————————————

6 ，transactionCapacity 需要大于 source 和 sink 的 batchSize 参数

sink ：

增加 sink 个数可以增加消费 event 能力