本章代码链接：https://github.com/gitdreams/hadoop-book/tree/master/ch14-flume

1. 要想使用Flume,就需要运行Flume代理，Flume代理有持续运行的**source**（数据来源），**sink**（数据目标）以及**channel**（用于连接source和sink）构成的java进程。

源，水池，渠道

1. 事务和可靠性，Flume使用**两个独立的事务**负责**source🡪channel**和**channel🡪sink**的事件传递。
2. channel 可以分为file channel 和 memory channel ，其中file channel 具有持久性，重启代理事件也不会丢失；memory channel 不具有持久性，重启代理事件会丢失，但是吞吐量较高。
3. at-least-once 至少一次， exactly-once 恰好一次
4. HDFS Sink，使用spooling diretoty source 和HDFS sink的Flume配置

agent1.sources = source1

agent1.sinks=sink1

agent1.channels=channel1

agent1.sources.source1.channels=channel1

agent1.sinks.sink1.channel=channel1

agent1.sources.source1.type =spooldir

agent1.sinks.sink1.channel=/tmp/spooldir

agent1.sinks.sink1.type=hdfs

agent1.sinks.sink1.hdfs.path/tmp/flume

agent1.sinks.sink1.hdfs.filePrefix=events

agent1.sinks.sink1.hdfs.fileSuffix=.log

agent1.sinks.sink1.hdfs.inUsePrefix=\_

agent1.sinks.sink2.hdfs.fileType=DataStream

agent1.channels.channel1.type=file

配置文件大体内容

1. 分区和拦截器

大型数据集常常被组织为分区（partition），Flume事件通常按照时间来分区，一个Flume时间被写入哪个分区是由事件的header中的时间戳决定的，可以使用时间拦截器，它将为每一个source产生的每个事件添加一个时间戳header，

Agent1.sources.source1.interceptors=interceptor1

Agent1.sources.source1.interceptors.interceptor1.type=timestamp

1. 文件格式

默认的Hdfs.fileType文件格式是SequenceFile序列文件。键的数据类型为LongWritable（header中的时间戳，如果没有就用当前时间），数据的类型为ByteWritable。

1. 扇出

扇出（fan out）是指一个source向多个channel，亦即向多个sink传递事件

1. 交付保证
2. 复制和复用选择器

Source想使一些事件流向一个channel，另一个事件流向另一个channel，可以通过在source上设置一个**复用选择器**来实现

1. 通过代理层分发