# Flume 课堂笔记

<http://flume.apache.org/FlumeUserGuide.html> 官方用户向导

**Flume 主要用于实时采集文件源**

#### Source：Spooling Directory Source

不可能用Kafka做Channel，否则流程就是如下：

Flume（Kafka）——>Kafka

文件源——>Flume——> Kafka ——> 实时计算（Streaming、Storm）

——> 按小时或按天 offline消费写HDFS/Hive

——> hbase

flume-ng agent --conf conf --conf-file example.conf --name a1 -Dflume.root.logger=INFO,console

--conf 指定配置文件的目录

--conf-file 指定配置文件

--name a1 指定agent名称

学员提问：

罗剑平 20:35:52

这种架构的话，如果源端生产数据的速度远远超过消费端消费数据的速度，kafka有什么机制可以平衡生产和消费两端吗？

答：kafka是MQ产品，数据存储在硬盘上

李敏 21:17:26

接罗剑平的问题，消息产生端-》kafka-》消息消费端，如果产生端短时间有大量数据消息，kafka能否马上写入，消费端能否马上消费？

答：能。

这个是否和kafka集群和消息消费端的主机能力有关？是可以通过集群来扩展能力的？

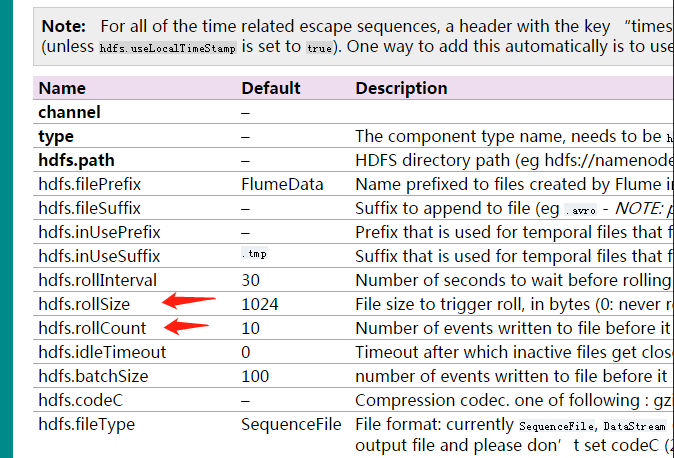
答：如果消费端出现消费延迟，建议1、增加消费端作业的资源；2、增加Topic分区数。

吴卓钧 21:03:02

spark-sql是调用hive的服务，hive又是基于MR，这个是不是和hadoop生态圈里面画的spark和MR之间关系有点不同？应该如何正确理解spark，hive，MR三者之间的关系？

答：Hive基于MR计算引擎，如果用 spark-sql是调用hive的话计算在Spark计算引擎上。

Hive是基于计算引擎上的一个Sql开发接口。



hdfs.rollSize：写到HDFS上的文件大小，1024即1M；

hdfs.rollCount：10 即 10个events写一次hdfs

HDFS小文件问题：造成Namenode 元数据的庞大