1. 知道 Flume 的 Channel 是啥吗

1. Channel 被设计为 Event 中转临时缓冲区，存储 Source 收集并且没有被 Sink 读取的 Event，为平衡 Source 收集和 Sink 读取的速度，可视为 Flume 内部的消息队列。

2. Channel 线程安全并且具有事务性，支持 Source 写失败写，和 Sink 读失败重复读的操作。常见的类型包括 Memory Channel，File Channel，Kafka Channel。

2. 介绍一下 Memory Channel

读写速度快，但是存储数据量小，Flume 进程挂掉、服务器停机或者重启都会导致数据丢失。资源充足、不关心数据丢失的场景下可以用。

3. 说说 File Channel

将 event 写入磁盘文件，与 Memory Channel 相比存储容量大，无数据丢失风险。File Channel 数据存储路径可以配置多磁盘文件路径，通过磁盘并行写入提高 File Channel 性能。Flume 将 Event 顺序写入到 File Channel 文件的末尾。可以在配置文件中通过设置 maxFileSize 参数配置数据文件大小，当被写入的文件大小达到上限的时候，Flume 会重新创建新的文件存储写入 Event。当一个已经关闭的只读数据文件的 Event 被读取完成，并且 Sink 已经提交读取完成的事务，则 Flume 把存储该数据的文件删除。

4. 说说 Kafka Channel

Memory Channel 有很大的丢数据风险，而且容量一般，File Channel 虽然能缓存更多的消息，但如果缓存下来的消息还没写入 Sink，此时 Agent 出现故障则 File Channel 中的消息一样不能被继续使用，直到该 Agent 恢复。而 Kafka Channel 容量大，容错能力强。

有了 Kafka Channel 可以在日志收集层只配置 Source 组件和 Kafka 组件，不需要再配置 Sink 组件，减少了日志收集层启动的进程数，有效降低服务器内存、磁盘等资源的使用率。而日志汇聚层，可以只配置 Kafka Channel 和 Sink，不需要再配置 Source。

kafka.consumer.auto.offset.reset，当 Kafka 中没有 Consumer 消费的初始偏移量或者当前偏移量在 Kafka 中不存在（比如数据已经被删除）情况下 Consumer 选择从 Kafka 拉取消息的方式，earliest 表示从最早的偏移量开始拉

取，latest 表示从最新的偏移量开始拉取，none 表示如果没有发现该 Consumer 组之前拉取的偏移量则抛出异常。

5. 介绍一下 Kafka 几种 Sink

1. HDFS Sink: 将 Event 写入 HDFS 文件存储，能够有效长期存储大量数据。
2. Kafka Sink: Flume 通过 Kafka Sink 将 Event 写入到 Kafka 中的主题，其他应用通过订阅主题消费数据。kafka.producer.acks 可以设置 Producer 端发送消息到 Broker 之后不需要等待 Broker 返回成功送达的信号。0表示 Producer 发送消息到 Broker 之后不需要等待 Broker 返回成功送达的信号，这种方式吞吐量高，但存在丢失数据的风险。1表示 Broker 接收到消息成功写入本地 log 文件后向 Producer 返回成功 接收的信号，不需要等待所有的 Follower 全部同步完消息后再做回应，这种方式在数据丢失风险和吞吐量之间做了平衡。-1表示 Broker 接收到 Producer 的消息成功写入本地 log 并且等待所有的 Follower 成功写入本地 log 后向 Producer 返回成功接收的信号，这种方式能够保证消息不丢失，但是性能最差（层层递进）。

6. 说说 Flume 的拦截器

Source 将 Event 写入到 Channel 之前可以使用拦截器对 Event 进行各种形式的处理，Source 和 Channel 之间可以有多个拦截器，不同拦截器使用不同的规则处理 Event，包括时间、主机、UUID、正则表达式等多种形式的拦截器。

7. 介绍一下什么是选择器

Source 发送的 Event 通过 Channel 选择器来选择以哪种方式写入到 Channel 中，Flume 提供三种类型 Channel 选择器，分别是复制、复用和自定义选择器。

1. 复制选择器: 一个 Source 以复制的方式将一个 Event 同时写入到多个 Channel 中，不同的 Sink 可以从不同的 Channel 中获取相同的 Event，比如一份日志数据同时写 Kafka 和 HDFS，一个 Event 同时写入两个 Channel，然后不同类型的 Sink 发送到不同的外部存储。
2. 复用选择器: 需要和拦截器配合使用，根据 Event 的头信息中不同键值数据来判断 Event 应该写入哪个 Channel 中。

8. 了解 Flume 的负载均衡和故障转移吗

目的是为了提高整个系统的容错能力和稳定性。简单配置就可以轻松实现，首先需要设置 Sink 组，同一个 Sink 组内有多个子 Sink，不同 Sink 之间可以配置成负载均衡或者故障转移。