一个合理的批量大小是多少？

这取决于部署、硬件和其他几个因素。没有大量的试错测试，批量大小不应该最终确定，因为批量太大也会导致自己的问题，如网络上太多的碎片等。同时，批量太大会增加事件重复的风险，因为每一批失败最终可能会导致大量的事件再次写入，如果有些事件已经成功写入 HDFS，这些事件最终会重新写入。

要为 RPC 和终端 Sink 选择正确的批量大小，开始是相当于几百 KB 到 1MB 的批量大小，然后基于你看到的超时和重复率，从那里向上或向下调整。如果有太多的重复或超时太多，你必须减小你的批量大小；在相反的情况下，增加批量大小直到开始出现超时。一旦你看到超时，说明已经达到阈值，应该减少几个百分点。

重复怎么样？

Flume 提供至少一次保证，这基本上意味着通过 Flume 到存储系统的发送事件至少将存储一次。但是最终 Flume 可能会不止一次地存储数据。有很多场景会导致重复，有些是由于错误，另一些是由于配置。

因为每个 agent-to-agent RPC 调用有一个可配置的超时，如果没有在超时时间内得到响应，即使 RPC 没有失败，发送事件的 Agent 也有可能认为 RPC 失败，从而引发重试。如果 RPC 没有失败，重试将导致相同事件再次发送，造成重复。这种情况可能发生在终端 Sink，如 HDFS 或 HBase Sink。

而且，由于 Flume Source 可以写入到多个 Channel，如果相同的 Source 配置了多个 Channel，同样的事件基本上可以重复。如果 Sink 从 Channel 读取事件并最终推送事件到相同的存储系统，这也可能会导致重复。

如果用例是重复敏感型的，在事件中插入唯一标识符通常是一个好主意。后续的处理工作可以使用这些标识符删除重复的数据，如使用 Spark、MapReduce 等。

运行 Flume Agent

本节假设 Flume 目录结构没有改变，当前工作目录是 Flume 目录结构的顶层。每个 Flume Agent 使用 flume-ng 命令从命令行启动。这个命令需要几个参数：要启动的 Flume Agent 名称、使用的配置文件、使用的配置目录。

Flume 配置文件可以包含多个 Flume Agent 的配置，每个由一个唯一名称确定。当启动