HDFS 或 NFS 将数据移动到 HDFS 也是有意义的，尤其是如果写出的数据量相对小——每几小时有几个 GB 的若干文件不会损害 HDFS。在这种情况下，设计、配置和部署 Flume 可能不值得做。Flume 真正做的是实时推送事件，数据流是持续且量级很大的情况。

如前面所指出的，Flume 最简单的部署单元叫作 *Flume Agent*。Agent 是一个 Java 应用程序，接收或者生成数据并缓冲数据，直到最终写入到 Agent 或者存储或索引系统。我们将在下一节论述 Flume Agent 的三个主要组件（Source、Channel、Sink）。

Flume Agent 内部原理

如前面所说，每个 Flume Agent 包含三个主要组件：Source、Channel、Sink。在本节，我们将描述这些组建以及其他组件，以及它们是如何协同工作的。

Source 是从一些其他产生数据的应用中接收数据的活跃组件。有自己产生数据的 Source，不过这些 Source 通常用于测试目的。Source 可以监听一个或者多个网络端口，用于接收数据或者可以从本地文件系统读取数据。每个 Source 必须至少连接一个 Channel。基于一些标准，一个 Source 可以写入几个 Channel，复制事件到所有或某些 Channel。

11 一般来说，Channel 是被动组件（虽然它们可以为了清理或者垃圾回收运行自己的线程），缓冲 Agent 已经接收，但尚未写出到另一个 Agent 或者存储系统的数据。Channel 的行为像队列，Source 写入到它们，Sink 从它们中读取。多个 Source 可以安全地写入到相同的 Channel，并且多个 Sink 可以从相同的 Channel 进行读取。可是一个 Sink 只能从一个 Channel 读取。如果多个 Sink 从相同的 Channel 读取，它可以保证只有一个 Sink 将会从 Channel 读取（和提交——更多参见第 4 章）一个指定特定的事件。

Sink 连续轮询各自的 Channel 来读取和删除事件。Sink 将事件推送到下一阶段（RPC Sink 的情况下），或到最终目的地。一旦在下一阶段或其目的地中数据是安全的，Sink 通过事务提交通知 Channel，可以从 Channel 中删除这些事件。

图 2-1 展示了拥有一个 Source、Channel 和 Sink 的简单 Flume Agent。

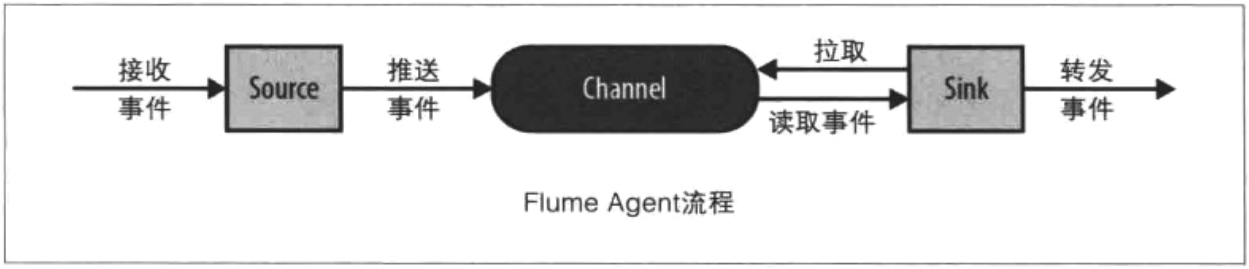

图2-1 带有一个Souce、Channel和Sink的简单Flume Agent