第 5 章 93

Sink

Flume 被设计为带有插入几乎每个组件的能力，包括那些写数据到最终目的地的组件——大多数情况下，目的地是一些数据存储。

从 Flume Agent 移除数据并写入到另一个 Agent 或数据存储或一些其他系统的组件被称为 *Sink*。为了推动这个过程，Flume 允许用户配置 Sink，可以是 Flume 内置的 Sink 之一或是由用户自己写的 Sink（对于非 Flume 内置的自定义 Sink，JAR 包应该放入 Flume 的 *plugins.d* 目录）。

Sink 是 Flume Agent 中的组件，能移除 Channel 中的数据，所以 Source 可以持续接收事件并写入到 Channel。Sink 不断地轮询 Channel 中的事件且批量地移除它们。这些事件批量写入到存储或索引系统，或者被发送到另一个 Flume Agent。

Sink 是完全事务性的。在从 Channel 批量移除数据之前，每个 Sink 用 Channel 启动一个事务。批量事件一旦成功写出到存储系统或下一个 Flume Agent，Sink 就利用 Channel 提交事务。事务一旦被提交，该 Channel 从自己的内部缓冲区删除事件。

Flume 封装了许多 Sink，可以写到诸如 HDFS、HBase、Solr、Elastic Search 等存储和索引系统。这些 Sink 通常被称为*终端* Sink，因为它们通常出现在 Flume 管道的终点。Flume 管道通过发送数据到另一个 Flume Agent 的 Agent 创建。通过 RPC sink-source 对发生这种通信。Flume 封装的 Avro- 和 Thrift-based RPC Sink，可以用于将数据发送给远程 Flume Agent 上各自的 RPC Source。

在本章中，我们将讨论不同种类的 Sink，它们的配置和管理，如何针对每个 Sink 序列化数据，以使得数据可以被写为用户选择的一种格式，以及该如何编写自定义 Sink。 94