计数器组成。考虑下面的配置：

```
agent.sinks.hdfsSink.hdfs.filePrefix = UsingFlume
agent.sinks.hdfsSink.hdfs.fileSuffix = .oreilly
```

配置将最终生成例如以 *UsingFlume.33434321.oreilly* 和 *UsingFlume.33434322.oreilly* 等命名的文件。

当文件仍然被写入时，建议类似Hive或MapReduce的系统忽略该文件直到Flume关闭它。这能确保当 MapReduce 或 Hive 读取文件的时候，文件内容不会更新。不幸的是，确定文件是否正被写入和内容是否更新是不容易的。为了解决这个问题，一旦 Flume 关闭了文件，HDFS Sink 允许用户添加后缀和前缀到将要移除的文件名。

通过使用这些参数，MapReduce 作业或 Hive 查询可以过滤一些文件。当文件正被写入，文件名将会以 in-use 前缀开始，以 in-use 后缀结尾，中间是最终的文件名（基于前面解释过的文件前缀和文件后缀参数）。我们再次看一下配置例子：

```
agent.sinks.hdfsSink.hdfs.filePrefix = UsingFlume
agent.sinks.hdfsSink.hdfs.fileSuffix = .oreilly
agent.sinks.hdfsSink.hdfs.inUsePrefix = .
agent.sinks.hdfsSink.hdfs.inUseSuffix = .temp
```

这将创建以 *.UsingFlume.33434321.oreilly.temp*、*.UsingFlume.33434322.oreilly.temp* 等命名的文件，一旦关闭将会重命名为 *UsingFlume.33434321.oreilly*、*UsingFlume.33434322.oreilly* 等。前面的配置确保了文件是隐藏的直到它们最终被关闭，重命名了文件名，文件名是前面描述过的 `hdfs.filePrefix` 和 `hdfs.fileSuffix` 参数指定的。

HDFS Sink 可以基于时间分桶，在第 4 章“理解 bucket”一节中描述过。Sink 基于 `hdfs.timeZone` 参数转换 Epoch 时间戳为分桶的日期和时间。如果没有指定的时区，那么就使用运行 Agent 机器的当地时区。时区按照 Internet Assigned Numbers Authority (IANA) 标准格式 [tz-list]。

104

为了确保文件是关闭，且对于处理数据的系统数据是可用的，HDFS Sink 可以被配置用基于时间、事件计数或写入到文件的预压缩大小来滚动文件。`hdfs.rollInterval` 参数控制基于时间的文件滚动。每个文件在该参数指定的时间（以秒为单位）之后被刷新、关闭和重命名。设置为 0 禁用基于时间的滚动。

HDFS Sink 也可以基于写入事件的数量来滚动文件。`hdfs.rollCount` 参数控制了这一点。设置为 0 表示禁用基于计数的滚动文件。最终，有可能使用 `hdfs.rollSize` 参数（即使写入的数据是压缩的格式，但这是未压缩的大小），基于文件中事件主体的总大小来滚动文件。该参数的值指定为以字节为单位。只要滚动文件的参数之一达到了阈值，文件