Filebeat 调研报告

指导老师：杨嵘

郑延钦

2021.09

目录

1. 工具简介

Filebeat 是 Beats 中的一员。Beats 是一个轻量级日志采集器，其包含六种工具：Packetbeat 收集网络流量数据，Metricbeat 收集系统、进程和文件系统级别的 CPU 和内存使用情况等数据, Filebeat 收集文件数据, Winlogbeat 收集 Windows 事件日志数据，Auditbeat 收集审计日志，Heartbeat 收集系统运行时的数据。Filebeat 是用于转发和集中日志数据的轻量级传送工具，监视您指定的日志文件或位置，收集日志事件，并将它们转发到 Elasticsearch 或 Logstash 进行索引。

Filebeat 由两个主要组成部分组成：prospector 和 harvesters。这些组件一起工作来读取文件并将事件数据发送到指定的 output。当你开启 Filebeat 程序的时候，它会启动一个或多个探测器（prospectors）去检测你指定的日志目录或文件，对于探测器找出的每一个日志文件，Filebeat 启动收割进程（harvester），每一个收割进程读取一个日志文件的新内容，并发送这些新的日志数据到处理程序（spooler），处理程序会集合这些事件，最后 filebeat 会发送集合的数据到你指定的地点。

2. 主流采集工具对比

大数据的数据采集工作是大数据技术中非常重要、基础的部分，具体场景使用合适的采集工具，可以大大提高效率和可靠性，并降低资源成本。Flume、Logstash 和 Filebeat 都是可以作为日志采集的工具，本章节将针对这三者进行对比分析。

2.1. Flume 简介

Flume 是一种分布式、高可靠和高可用的服务，用于高效地收集、聚合和移动大量日志数据。它有一个简单而灵活的基于流数据流的体系结构。它具有可调的可靠性机制、故障转移和恢复机制，具有强大的容错能力。它使用一个简单的可扩展数据模型，允许在线分析应用程序。

Flume 的设计宗旨是向 Hadoop 集群批量导入基于事件的海量数据。

2.2. Logstach 简介

Logstash 是开源的服务器端数据处理管道，能够同时从多个来源采集数据，转换数据，然后将数据发送到存储库中。数据从源传输到存储库的过程中，Logstash 过滤器能够解析各个事件，识别已命名的字段以构建结构，并将它们转换成通用格式，以便更轻松、更快速地分析和实现商业价值。

2.3. 三款工具对比

	Flume	Logstash	Filebeat
内存	大	大	小
cpu		大	小
背压敏感协议	否	否	是
插件	需要 API	多	多
功能	多输入端采集，多输出端输出	多输入端采集，实时分解并转换数据，多输出端输出	多输入端采集，多输出端输出
轻重	相对较重	相对较重	轻量二进制文件
过滤能力	自带分区和拦截器功能	强大的过滤能力	有过滤能力但是较弱
进程	一台服务器可以拉起多个进程，挂掉后去手动拉起	一台服务器只允许一个 Logstash 进程，挂掉后需手动拉起	十分稳定
编写语言	java	jruby	go
集群	分布式	单节点	单节点
输出至多个接收方	支持	支持	6.0 前支持
扩展开发	一般	难	易

3. Filebeat 日志采集

日志数据通常以文本文件的形式存储于磁盘中，使用 Filebeat 收集日志数据的方法有以下两种：

（1）数据采集模块：常见日志格式的采集、解析和可视化

（2）手动 Filebeat 配置

3.1. 预置模块支持

Filebeat 预置了多种服务的日志采集模块，包括 Kafka module、Microsoft module、MySQL module、Nginx module、Office 365 module、Oracle module、Redis module、System module、Tomcat module 等 71 个模块，通过快速启动模块可以对多个服务定向采集日志。如果日志不在默认位置，需在 modules.d/*.yml 配置文件中设置 path 变量，或通过命令行进行设置。

若需启动预置模块应用的日志采集，需在 filebeat.yml 中配置以下内容（该例子为采集 nginx、mysql 与 system 日志）：

filebeat.modules:

```
- module: nginx
- module: mysql
- module: system
```

3.2. 手动配置

对于未提供服务的应用程序，您可以自定义配置实现日志采集，最少需在 filebeat.yml 中配置以下内容：

```
filebeat.inputs:
- type: log
  enabled: true
  paths:
- /var/log/*.log
```

此示例中 Filebeat 将收集/var/log/ 以.log.结尾的目录中的所有文件。

3.3. ECS 记录器

虽然 Filebeat 可用于摄取原始的纯文本应用程序日志，但还是建议在摄取时重新构建日志。这使您可以提取字段，例如日志级别和异常堆栈跟踪。Elastic 通过提供各种流行编程语言的应用程序日志格式化程序来简化此过程。这些插件将您的日志格式化为与 ECS 兼容的 JSON，从而无需手动解析日志。

若不采用 ECS logger，则采集的日志为原始格式。

若采用了 ECS logger，则会将日志解析为 json 格式，例如：{"@timestamp":"2019-08-06T12:09:12.375Z", "log.level": "INFO", "message":"Tomcat started on port(s): 8080 (http) with context path "", "service.name":"spring-petclinic","process.thread.name":"restartedMain","log.logger": "org.springframework.boot.web.embedded.tomcat.TomcatWebServer"}

每一条日志会解析为一个 json 格式的内容，其通常包含但不限于：

- @timestamp：日志事件的时间戳。
- log.level：日志事件的级别或严重性。
- log.logger：应用程序中记录器的名称。
- log.origin.file.name：包含产生日志事件的源代码的文件的名称。
- log.origin.file.line：包含产生日志事件的源代码的文件的行号。
- log.origin.function：产生日志事件的函数或方法的名称。
- message：日志消息。
- error.type：仅适用于包含异常或错误的日志。如果此日志事件包含异常，则为错误的类型或类别。
- error.message：仅适用于包含异常或错误的日志。异常或错误的消息。
- error.stack_trace：仅适用于包含异常或错误的日志。作为原始字符串的异常或错误的完整堆栈跟踪。
- process.thread.name：记录事件的线程的名称。
- service.name：帮助按服务对日志进行归档。
- event.dataset：启用日志率异常检测。

对于不同模块的日志，还会单独解析出不同的字段。

ECS logger 解析服务时，并非切割日志原文，若日志本身为 json 格式，则将和 ECS logger 配置的其他关键字合并为一个 json 格式，若日志解析 json 格式失败，则整体存入 message 字段。

4. Filebeat 数据导出

Filebeat 将日志收集后，可以将事件流式传输到 Elasticsearch（需额外安装服务）。已将日志流式传输到 Elasticsearch 后，可以通过 kibana 可视化您的数据统计结果。建议安装其他 Beats 程序，可以获得更多的系统数据并对相关数据进行监控和分析。

除此之外，Filebeat 也可以将采集的日志数据配置传递给 Logstash、kafka、redis，或是生成文本（json 格式）进行本地存储。

4.1. 导出至文本配置

配置选项编辑，您可以将采集的日志按照文件的形式存储在指定路径。在 filebeat.yml 配置文件的部分中指定以下选项：

- enabled 编辑：启用的配置是一个布尔设置，用于启用或禁用输出。如果设置为 false，则禁用输出。默认值是 true。
- path 编辑：将保存生成的文件的目录的路径。此选项是强制性的。
- filename 编辑：生成的文件的名称。默认设置为节拍名称。例如，默认为 Filebeat 生成的文件将是“filebeat”、“filebeat.1”、“filebeat.2”等。
- rotate_every_kb 编辑：每个文件的最大大小（以千字节为单位）。当达到此大小时，文件将被旋转。默认值为 10240 KB。
- number_of_files 编辑：下保存的最大文件数 path。当达到这个文件数时，最旧的文件被删除，其余的文件从最后移到第一个。文件数必须在 2 到 1024 之间。默认值为 7。
- permissions 编辑：用于文件创建的权限。默认值为 0600。

4.2. 导出至 kafka 配置

此输出适用于 0.11 和 2.2.2 之间的所有 Kafka 版本。旧版本也可能工作，但不受支持。配置选项编辑，您可以将采集的日志输出到 kafka 中。在 filebeat.yml 配置文件的部分中指定以下选项：

- enabled 编辑：的 enabled 配置是一个布尔设置可启用或禁止输出。如果设置为 false，则禁用输出。默认值是 true。
- hosts 编辑：从哪里获取集群元数据的 Kafka 代理地址列表。集群元数据包含实际发布到的 Kafka 代理事件。
- version 编辑：假设运行 Kafka 版本的 filebeat。默认为 1.0.0。如果启用了 0.10.0.0+ 版本，将添加事件时间戳。

- username 编辑：用于连接到 Kafka 的用户名。如果配置了用户名，则还必须配置密码。
- password 编辑：连接 Kafka 的密码。
- topic 编辑：用于生产事件的 Kafka 主题。
- topics 编辑：一组主题选择器规则。每个规则指定 topic 用于匹配规则的事件。在发布期间，Filebeattopic 根据数组中的第一个匹配规则为每个事件设置。规则可以包含条件、基于字符串的格式字段和名称映射。如果 topics 缺少设置或没有规则匹配，topic 则使用该字段。
- key 编辑：指定 Kafka 事件键的可选格式化字符串。如果已配置，则可以使用格式字符串从事件中提取事件密钥。
- partition 编辑：Kafka 输出代理事件分区策略。必须是一个 random，round_robin 或 hash。默认情况下使用 hash 分区程序。
- ssl 编辑：SSL 参数的配置选项，例如 Kafka 连接的根 CA。应该使用-keyalg RSA 参数创建 Kafka 主机密钥库，以确保它使用 Filebeat 的 Kafka 库支持的密码。Filebeat 是一个日志文件托运工具，在服务器上安装客户端后，Filebeat 会监控日志目录或者指定的日志文件，追踪读取这些文件（追踪文件的变化，不停的读），并且转发这些信息到 ElasticSearch 或者 Logstarsh 中存放。

5. 总结

Filebeat 是一个日志文件托运工具，在服务器上安装客户端后，Filebeat 会监控日志目录或者指定的日志文件，追踪读取这些文件（追踪文件的变化，不停的读），并且转发这些信息到 ElasticSearch 或者 Logstarsh 中存放。其优势在于轻量级，占用资源小，便于开发且十分稳定，是一款非常优秀的日志数据采集工具。

参考资料

https://www.elastic.co/guide/en/beats/filebeat/current/index.html