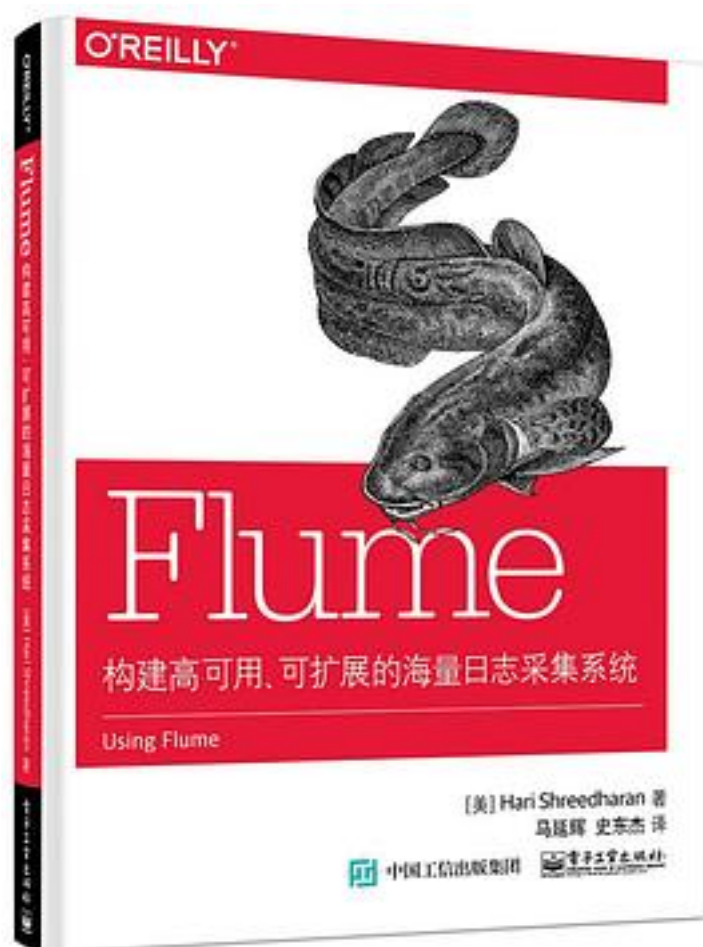


Flume



[Flume_下载链接1](#)

著者:【美】 Hari Shreedharan (哈里·史瑞德哈伦)

出版者:电子工业出版社

出版时间:2015-8-1

装帧:平装

isbn:9787121265587

《Flume：构建高可用、可扩展的海量日志采集系统》从Flume的基本概念和设计原理开始讲解，分别介绍了不同种类的组件、如何配置组件、如何运

行Flume Agent 等。同时，分别讨论Source、Channel 和Sink 三种核心组件，不仅仅阐述每个组件的基本概念，而且结合实际的编程案例，深入、全面地介绍每个组件的详细用法，并且这部分内容也是整个Flume 框架的重中之重。之后，讲解拦截器、Channel选择器、Sink 组和Sink 处理器等内容，它们为Flume 提供灵活的扩展支持。最后，介绍了Flume 的高级使用，如何使用Flume 软件开发工具集（SDK）和Embedded Agent API，如何设计、部署和监控Flume 生产集群。

总而言之，《Flume：构建高可用、可扩展的海量日志采集系统》是一本理论结合实战，深度、广度兼备的海量日志采集系统的著作。

作者介绍:

Hari Shreedharan是Cloudera的一名软件工程师，他工作于Apache Spark、Apache Flume和Apache Sqoop。他也是Flume项目的一个提交者和PMC成员，帮助项目的方向做决定。

目录: 译者序 v

序xiii

前言 x

第1 章 认识Apache Hadoop 和Apache HBase 1

 分布式文件系统HDFS1

 HDFS 的数据格式3

 处理HDFS 中的数据4

 Apache HBase4

 总结5

 参考文献6

第2 章 用Apache Flume 处理流数据 7

 我们需要Flume7

 Flume 是否适合呢?9

 Flume Agent 内部原理10

 配置Flume Agent13

 Flume Agent 之间的相互通信17

 复杂的流17

 复制数据到不同目的地20

 动态路由21

 Flume 的无数据丢失保证，Channel 和事务22

 Flume Channel 中的事务23

 Agent 失败和数据丢失25

 批量的重要性26

 重复怎么样?27

 运行Flume Agent27

 总结29

 参考文献30

第3 章 源（Source） 31

 Source 的生命周期31

 Sink-to-Source 通信33

 Avro Source34

 Thrift Source37

 RPC Sources 的失败处理39

 HTTP Source40

 针对HTTP Source 写处理程序*42

Spooling Directory Source	47
使用Deserializers 读取自定义格式*	50
Spooling Directory Source 性能.....	55
Syslog Source	55
Exec Source	59
JMS Source	61
转换JMS 消息为Flume 事件*	63
编写自定义Source*	65
Event-Driven Source 和Pollable Source	66
总结	73
参考文献	73
第4 章 Channel	75
事务工作流	76
Flume 自带的Channel	78
Memory Channel	78
File Channel	80
总结	86
参考文献	86
第5 章 Sink	87
Sink 的生命周期	88
优化Sink 的性能	89
写入到HDFS： HDFS Sink	89
理解Bucket	90
配置HDFS Sink	93
使用序列化器控制数据格式*	100
HBase Sink	106
用序列化器将Flume 事件转换成HBase Put 和Increment*	108
RPC Sink	113
Avro Sink	113
Thrift Sink	115
Morphline Solr Sink	116
Elastic Search Sink	119
自定义数据格式*	121
其他Sink： Null Sink、Rolling File Sink 和Logger Sink	124
编写自定义Sink*	125
总结	129
参考文献	129
第6 章 拦截器、Channel 选择器、Sink 组和Sink 处理器	131
拦截器	131
时间戳拦截器	132
主机拦截器	133
静态拦截器	133
正则过滤拦截器	134
Morphline 拦截器	135
UUID 拦截器	136
编写拦截器*	137
Channel 选择器	140
复制Channel 选择器	140
多路复用Channel 选择器	141
自定义Channel 选择器*	144
Sink 组和Sink 处理器	146
Load-Balancing Sink 处理器	148
Failover Sink 处理器	151

总结	153
参考文献	154
第7章 发送数据到Flume*	155
构建Flume 事件	155
Flume 客户端SDK	156
创建Flume RPC 客户端	157
RPC 客户端接口	157
所有RPC 客户端的公共配置参数	158
默认RPC 客户端.....	165
Load-Balancing RPC 客户端	168
Failover RPC 客户端	171
Thrift RPC 客户端	172
嵌入式Agent	173
配置嵌入式Agent	175
log4j Appender	180
Load-Balancing log4j Appender	181
总结	182
参考文献	183
第8章 规划、部署和监控Flume	185
规划一个Flume 部署	185
修复时间	185
我的Flume Channel 需要多少容量?	186
多少层?	186
通过跨数据中心链接发送数据	188
层分片	190
部署Flume	191
部署自定义代码	191
监控Flume	193
从自定义组件报告度量	196
总结	196
参考文献	196
索引	197
• • • • • (收起)	

[Flume 下载链接1](#)

标签

大数据

flume

日志

计算机

翻译质量差

数据挖掘

编程

程序设计

评论

还好吧。

简单粗读，待实战仔细体会

基本把相关组件关系原理理顺了。翻译的很一般，读起来不流畅。

说真的，我完全看不懂作者在讲什么，嘚啵嘚的好几章，还不如直接在百度上搜一个简单介绍看的明白。

建议看原版书 这个翻译的有点儿…… 至少影响阅读流畅度 (之所以四星主要是因为翻译的原因 如果不考虑翻译的因素 一定是五星)

flume功能，使用场景，原理配置（source，channel，sink，拦截器，选择器，sink组与处理器），规划部署相当全面深入

flume就是一个采集日记的工具，然后把日志存放到hbase或者hdfs上面。由source，c

hannel和sink组成。

配合官方文档阅读更佳

比看文档快 适合快速掌握后用于生产

中规中矩，将官方文档做了整理，加上了部分经验之谈。

有段时间公司用到，但鉴于该项目的活跃程度，很少会有公司再选型，更多应该是ELK Stack，除非特别要用到HDFS

就这翻译也好意思出书，垃圾

[Flume_下载链接1](#)

书评

[Flume_下载链接1](#)