

Kafka Streams实战



[Kafka Streams实战 下载链接1](#)

著者:[美]小威廉·P. 贝杰克 (William P. Bejeck Jr.)

出版者:人民邮电出版社

出版时间:2018-7

装帧:平装

isbn:9787115507396

Kafka

Streams是Kafka提供的一个用于构建流式处理程序的Java库，它与Storm、Spark等流式处理框架不同，是一个仅依赖于Kafka的Java库，而不是一个流式处理框架。除Kafka之外，Kafka

Streams不需要额外的流式处理集群，提供了轻量级、易用的流式处理API。

本书包括4部分，共9章，从基础API到复杂拓扑的应用，通过具体示例由浅入深地详细介绍了Kafka

Streams基础知识及使用方法。本书的主要内容包含流式处理发展历程和Kafka Streams工作原理的介绍，Kafka基础知识的介绍，使用Kafka

Streams实现一个具体流式处理应用程序，讨论状态存储及其使用方法，讨论表和流的二元性及使用场景，介绍Kafka Streams应用程序的监控及测试方法，介绍使用Kafka Connect将现有数据源集成到Kafka Streams中，使用KSQL进行交互式查询等。本书适合使用Kafka Streams实现流式处理应用的开发人员阅读。

作者介绍:

小威廉·P. 贝杰克 (William P. Bejeck Jr., 本名Bill Bejeck) 是Kafka Streams源码贡献者，在Confluent公司的Kafka Streams团队工作，有15年以上的软件开发经验，其中8年专注于后端开发，特别是处理大量数据，在数据提炼团队中，使用Kafka来改善下游客户的数据流。

目录: 第一部分 开启Kafka Streams之旅

第1章 欢迎来到Kafka Streams 3

1.1 大数据的发展以及它是如何改变程序设计方式的 3

1.1.1 大数据起源 4

1.1.2 MapReduce中的重要概念 5

1.1.3 批处理还不够 7

1.2 流式处理简介 8

1.3 处理购买交易 9

1.3.1 权衡流式处理的选择 9

1.3.2 将需求解构为图表 10

1.4 改变看待购买交易的视角 10

1.4.1 源节点 11

1.4.2 信用卡屏蔽节点 11

1.4.3 模式节点 11

1.4.4 奖励节点 12

1.4.5 存储节点 13

1.5 Kafka Streams在购买处理节点图中的应用 13

1.6 Kafka Streams在购买交易流中的应用 14

1.6.1 定义源 15

1.6.2 第一个处理器：屏蔽信用卡号码 15

1.6.3 第二个处理器：购买模式 16

1.6.4 第三个处理器：客户奖励 17

1.6.5 第四个处理器：写入购买记录 18

1.7 小结 18

第2章 Kafka快速指南 20

2.1 数据问题 20

2.2 使用Kafka处理数据 21

2.2.1 ZMart原始的数据平台 21

2.2.2 一个Kafka销售交易数据中心 22

2.3 Kafka架构 23

2.3.1 Kafka是一个消息代理 23

2.3.2 Kafka是一个日志 24

2.3.3 Kafka日志工作原理 25

2.3.4 Kafka和分区 25

2.3.5 分区按键对数据进行分组 26

2.3.6 编写自定义分区器 27

2.3.7 指定一个自定义分区器 28

2.3.8 确定恰当的分区数 29

2.3.9 分布式日志 29

2.3.10 ZooKeeper：领导者、追随者和副本 30

- 2.3.11 Apache ZooKeeper 31
- 2.3.12 选择一个控制器 31
- 2.3.13 副本 31
- 2.3.14 控制器的职责 32
- 2.3.15 日志管理 33
- 2.3.16 日志删除 33
- 2.3.17 日志压缩 35
- 2.4 生产者发送消息 36
 - 2.4.1 生产者属性 38
 - 2.4.2 指定分区和时间戳 39
 - 2.4.3 指定分区 39
 - 2.4.4 Kafka中的时间戳 40
- 2.5 消费者读取消息 40
 - 2.5.1 管理偏移量 41
 - 2.5.2 自动提交偏移量 42
 - 2.5.3 手动提交偏移量 42
 - 2.5.4 创建消费者 43
 - 2.5.5 消费者和分区 43
 - 2.5.6 再平衡 43
 - 2.5.7 更细粒度的消费者分配 44
 - 2.5.8 消费者示例 44
- 2.6 安装和运行Kafka 45
 - 2.6.1 Kafka本地配置 45
 - 2.6.2 运行Kafka 46
 - 2.6.3 发送第一条消息 47
- 2.7 小结 49
- 第二部分 Kafka Streams开发篇
- 第3章 开发Kafka Streams 53
 - 3.1 流式处理器API 53
 - 3.2 Kafka Streams的Hello World 54
 - 3.2.1 构建“Yelling App”的拓扑 55
 - 3.2.2 Kafka Streams配置 58
 - 3.2.3 Serde的创建 59
 - 3.3 处理客户数据 60
 - 3.3.1 构建一个拓扑 61
 - 3.3.2 创建一个自定义的Serde 67
 - 3.4 交互式开发 69
 - 3.5 下一步 71
 - 3.5.1 新需求 71
 - 3.5.2 将记录写入Kafka之外 76
 - 3.6 小结 78
- 第4章 流和状态 79
 - 4.1 事件的思考 79
 - 4.2 将状态操作应用到Kafka Stream 81
 - 4.2.1 值转换处理器 82
 - 4.2.2 有状态的客户端奖励 82
 - 4.2.3 初始化值转换器 84
 - 4.2.4 使用状态将Purchase对象映射为Reward Accumulator 84
 - 4.2.5 更新奖励处理器 88
 - 4.3 使用状态存储查找和记录以前看到的数据 89
 - 4.3.1 数据本地化 90
 - 4.3.2 故障恢复和容错 91
 - 4.3.3 Kafka Streams使用状态存储 91
 - 4.3.4 其他键/值存储供应者 92

- 4.3.5 状态存储容错 93
- 4.3.6 配置变更日志主题 93
- 4.4 连接流以增加洞察力 94
 - 4.4.1 设置数据 95
 - 4.4.2 生成包含客户ID的键来执行连接 96
 - 4.4.3 构建连接 98
 - 4.4.4 其他连接选项 102
- 4.5 Kafka Streams中的时间戳 104
 - 4.5.1 自带的时间戳提取器实现类 105
 - 4.5.2 WallclockTimestampExtractor 106
 - 4.5.3 自定义时间戳提取器 106
 - 4.5.4 指定一个时间戳提取器 107
- 4.6 小结 108
- 第5章 KTable API 109
 - 5.1 流和表之间的关系 110
 - 5.1.1 记录流 110
 - 5.1.2 更新记录或变更日志 111
 - 5.1.3 事件流与更新流对比 113
 - 5.2 记录更新和KTable配置 115
 - 5.2.1 设置缓存缓冲大小 115
 - 5.2.2 设置提交间隔 116
 - 5.3 聚合和开窗操作 117
 - 5.3.1 按行业汇总股票成交量 118
 - 5.3.2 开窗操作 122
 - 5.3.3 连接KStream和KTable 128
 - 5.3.4 GlobalKTable 130
 - 5.3.5 可查询的状态 133
 - 5.4 小结 133
- 第6章 处理器API 135
 - 6.1 更高阶抽象与更多控制的权衡 135
 - 6.2 使用源、处理器和接收器创建一个拓扑 136
 - 6.2.1 添加一个源节点 136
 - 6.2.2 添加一个处理器节点 137
 - 6.2.3 增加一个接收器节点 140
 - 6.3 通过股票分析处理器深入研究处理器API 141
 - 6.3.1 股票表现处理器应用程序 142
 - 6.3.2 process()方法 145
 - 6.3.3 punctuator执行 147
 - 6.4 组合处理器 148
 - 6.5 集成处理器API和Kafka Streams API 158
 - 6.6 小结 159
- 第三部分 管理Kafka Streams
- 第7章 监控和性能 163
 - 7.1 Kafka基本监控 163
 - 7.1.1 测评消费者和生产者性能 164
 - 7.1.2 检查消费滞后 165
 - 7.1.3 拦截生产者和消费者 166
 - 7.2 应用程序指标 169
 - 7.2.1 指标配置 171
 - 7.2.2 如何连接到收集到的指标 172
 - 7.2.3 使用JMX 172
 - 7.2.4 查看指标 176
 - 7.3 更多Kafka Streams调试技术 177
 - 7.3.1 查看应用程序的表现形式 177

- 7.3.2 获取应用程序各种状态的通知 178
- 7.3.3 使用状态监听器 179
- 7.3.4 状态恢复监听器 181
- 7.3.5 未捕获的异常处理器 184
- 7.4 小结 184
- 第8章 测试Kafka Streams应用程序 185
- 8.1 测试拓扑 186
 - 8.1.1 构建测试用例 188
 - 8.1.2 测试拓扑中的状态存储 190
 - 8.1.3 测试处理器和转换器 191
- 8.2 集成测试 193
- 8.3 小结 199
- 第四部分 Kafka Streams进阶
- 第9章 Kafka Streams的高级应用 203
- 9.1 将Kafka与其他数据源集成 204
 - 9.1.1 使用Kafka Connect集成数据 205
 - 9.1.2 配置Kafka Connect 205
 - 9.1.3 转换数据 207
- 9.2 替代数据库 211
 - 9.2.1 交互式查询的工作原理 213
 - 9.2.2 分配状态存储 213
 - 9.2.3 创建和查找分布式状态存储 215
 - 9.2.4 编写交互式查询 216
 - 9.2.5 查询服务器内部 218
- 9.3 KSQL 221
 - 9.3.1 KSQL流和表 222
 - 9.3.2 KSQL架构 222
 - 9.3.3 安装和运行KSQL 224
 - 9.3.4 创建一个KSQL流 224
 - 9.3.5 编写KSQL查询 226
 - 9.3.6 创建一张KSQL表 227
 - 9.3.7 配置KSQL 227
- 9.4 小结 228
- 附录A 额外的配置信息 229
- 附录B 精确一次处理语义 234
- • • • • [\(收起\)](#)

[Kafka Streams实战 下载链接1](#)

标签

大数据

流计算

流式计算

消息引擎

数据湖

工作学习

中间件

Kafka

评论

分布式系统不是从应用程序中学习，而是在大数据底层架构HDFS构造和中间件的使用和配置过程中慢慢理解。单点和压力是所有程序最重要的困难点。

代码贴的有点多了，略显啰嗦

基础的概念和一些KStream
join、map、reduce之类的接口说明，完整的一个例子演示。
原理上设计展开不多，类似最后一章中“交互式查询基于state
store的设计”部分解读很少有。

[Kafka Streams实战_下载链接1](#)

书评

结论 本书非常适合新手，循规蹈矩按部就班即可入门，但排版非常粗糙！ 首先说明：
我是在亚马逊中国购买的Kindle版，主要在Mac以及Kindle上进行阅读
因没接触到纸质书籍和PDF格式，因此所论述仅限于Kindle版本 排版太粗糙了！
Kindle版本整体给人感觉，排版比较粗糙，编辑或作...

[Kafka Streams实战_下载链接1](#)