异步图书
www.epubit.com

MANNING

Kafka Streams 实战

Kafka Streams IN ACTION

[美] 小威廉·P. 贝杰克（William P. Bejeck Jr.） 著
牟大恩 译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

MANNING

Kafka Streams 实战

Kafka Streams IN ACTION

[美] 小威廉·P. 贝杰克 (William P. Bejeck Jr.) 著
牟大恩 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Kafka Streams实战 / (美) 小威廉 • P. 贝杰克 (William P. Bejeck) 著 ; 牟大恩译. -- 北京 : 人民邮电出版社, 2019.5
书名原文: Kafka Streams in Action
ISBN 978-7-115-50739-6

Ⅰ. ①K… Ⅱ. ①小… ②牟… Ⅲ. ①分布式操作系统 Ⅳ. ①TP316.4

中国版本图书馆CIP数据核字(2019)第022454号

版权声明

◆ 著　[美] 小威廉 • P. 贝杰克（William P. Bejeck Jr.）
译　牟大恩
责任编辑　杨海玲
责任印制　焦志炜

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：800×1000　1/16
印张：16
字数：344 千字　2019 年 5 月第 1 版
印数：1 – 3 000 册　2019 年 5 月河北第 1 次印刷

著作权合同登记号　图字：01-2018-7736 号

定价：69.00 元

读者服务热线：(010) 81055410　印装质量热线：(010) 81055316
反盗版热线：(010) 81055315
广告经营许可证：京东工商广登字 20170147 号

内容提要

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 到复杂拓扑的高级应用，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（包括高级特性），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。

中文版序

当我在 2015 年从领英的流数据架构组离职加入 Confluent 的时候，与 Jay 和 Neha 两人有过一次长时间的交流。当时公司刚刚成立，一切都还是从零起步。Jay 问我，接下来想要开展哪些工作，我回答说，我已经在流式存储层面，也就是 Kafka Core 做了两年多的时间，接下来我的兴趣是在存储上，也就是计算层面寻求一些新的挑战。大数据这个提法叫了这么多年了，可是一直以来我们都致力在数据的大规模（volume）上，比如数据系统的可延展性等；我觉得接下来大数据的趋势会向第二个“V”，也就是快速率（velocity）发展，因为越来越多的人已经不满意批处理带来的时间延迟，他们需要的是就在下一秒，从收集的数据中获得信息，产生效益。

所以，接下来我想做流式数据处理。这个想法和他们一拍即合，从那时候开始我投入到 Kafka Streams 的开发中来。

从写下第一行 Kafka Streams 的代码到今天已经快 4 年的时间了，在这期间我有幸目睹了流式数据处理和流事件驱动架构在硅谷的互联网行业，进而在全世界的各个商业领域中突飞猛进的发展。越来越多的人开始从请求/响应以及批处理的应用编程模式向流式处理转移，越来越多的企业开始思考实时计算如何能够给他们的产品或者服务带来信息收益，而 Apache Kafka 作为当今流数据平台的事实标准，正在被越来越多的人注意和使用。而 Kafka Streams 作为 Apache Kafka 项目下原生的流式处理库，也越来越多地被投入到生产环境中，并且得到了大量社区贡献者的帮助。这对我本人而言，是莫大的喜悦和欣慰。

在今年上半年，我的同事 Bill Bejeck 完成了这本《Kafka Streams 实战》，本书是 Bill 通过总结自身开发并维护真实生产环境下的 Kafka Streams 的经验完成的，对于想要学习并掌握 Kafka Streams 以及流事件驱动架构的读者来说是最好的方式之一。本书的译者牟大恩对 Kafka 源代码了解颇深，此前已著有《Kafka 入门与实践》一书，我相信一定能够准确还原 Bill 在书中想要带给大家的关于流式数据处理应用实践的思维模式。

祝各位读者在探索 Kafka Streams 的路上不断有惊喜的发现！

——王国璋（Guozhang Wang）

Confluent 流数据处理系统架构师

Apache Kafka PMC，Kafka Streams 作者之一

译者序

Kafka 在 0.10 版本中引入了 Kafka Streams，它是一个轻量级、简单易用的基于 Kafka 实现的构建流式处理应用程序的 Java 库。虽然它只是一个 Java 库，但具备了流式处理的基本功能，同时它利用 Kafka 的分区特性很容易实现透明的负载均衡以及水平扩展，从而达到高吞吐量。

一年前我在写《Kafka 入门与实践》一书时，用了专门一章讲解 Kafka Streams，由于那是一本关于 Kafka 的书，因此对 Kafka Streams 的讲解并没有面面俱到。巧合的是，本书作为一本关于 Kafka Streams 的书，也是用专门一章来介绍 Kafka。就我个人而言，我觉得这两本书中的内容在某种程度上可以互为补充，大家可以根据自己的偏好选择适合自己的 Kafka 书籍。

我很荣幸有机会翻译本书。通过翻译本书，无论是 Kafka Streams 知识本身还是本书作者的写作编排方式，都使我收获颇多。Kafka Streams 的诸多设计优点在本书中都有详细介绍，并结合具体示例对相关 API 进行讲解。本书通过模拟近乎真实的场景，从场景描述开始，逐步对问题进行剖析，然后利用 Kafka Streams 解决问题。阅读本书，读者不仅能够全面掌握 Kafka Streams 相关的 API，而且能够轻松学会如何使用 Kafka Streams 解决具体问题。

在翻译本书的过程当中，我理解最深的是，国外的技术书籍不是直接给出解决问题的完整代码，而是在场景描述、问题分析、技术选型等方面给予更多的篇幅，这种方式更能够帮助读者真正深入地掌握相关技术的要领，正所谓“授人以鱼，不如授人以渔”。

在此特别感谢人民邮电出版社的杨海玲编辑及其团队，正是他们一丝不苟、认真专业的工作态度，才使本书得以圆满完成。借此机会，我还要感谢我公司信息技术部副总经理、开发中心总经理王洪涛和部门经理熊友根对我的培养，以及同事给予我的帮助。同时还要感谢我的妻子吴小华，姐姐屈海林、尚立霞，妹妹石俊豪，感谢她们在我翻译本书时对我和我儿子的照顾，正是有了她们的帮助，才使我下班回到家时可以全身心投入到翻译工作中。同时，将本书送给我的宝贝儿子牟经纬，作为宝宝周岁的生日礼物，祝他健康、茁壮成长！

虽然在翻译过程中我力争做到“信、达、雅”，但本书许多概念和术语目前尚无公认的中文翻译，加之译者水平有限，译文中难免有不妥或错误之处，恳请读者批评指正。

牟大恩

2018 年 10 月

译者简介

牟大恩，武汉大学硕士研究生毕业，曾先后在网易杭州研究院、掌门科技、优酷土豆集团担任高级开发工程师和资深开发工程师职务，目前就职于海通证券总部。有多年的 Java 开发及系统设计经验，专注于互联网金融及大数据应用相关领域。著有《Kafka 入门与实践》，已提交技术发明专利两项，发表论文一篇。

序

我相信以实时事件流和流式处理为中心的架构将在未来几年变得无处不在。像 Netflix、Uber、Goldman Sachs、Bloomberg 等技术先进的公司已经建立了这种大规模运行的大型事件流平台。虽然这是一个大胆的断言，但我认为流式处理和事件驱动架构的出现将会对公司如何使用数据产生与关系数据库同样大的影响。

如果你还处在请求/响应风格的应用程序以及使用关系型数据库的思维模式，那么围绕流式处理的事件思维和构建面向事件驱动的应用程序需要你改变这种思维模式，这就是本书的作用所在。

流式处理需要从命令式思维向事件思维的根本性转变——这种转变使响应式的、事件驱动的、可扩展的、灵活的、实时的应用程序成为可能。在业务中，事件思维为组织提供了实时、上下文敏感的决策和操作。在技术上，事件思维可以产生更多自主的和解耦的软件应用，从而产生伸缩自如和可扩展的系统。

在这两种情况下，最终的好处是更大的敏捷性——在业务以及促进业务的技术方面。将事件思维应用于整个组织是事件驱动架构的基础，而流式处理是实现这种转换的技术。

Kafka Streams 是原生的 Apache Kafka 流式处理库，它用 Java 语言实现，用于构建事件驱动的应用程序。使用 Kafka Streams 的应用程序可以对数据流进行复杂转换，这些数据流能够自动容错，透明且弹性地分布在应用程序的实例上。自 2016 年在 Apache Kafka 的 0.10 版本中首次发布以来，许多公司已经将 Kafka Streams 投入生产环境，这些公司包括 P 站（Pinterest）、纽约时报（The New York Times）、拉博银行（Rabobank）、连我（LINE）等。

我们使用 Kafka Streams 和 KSQL 的目标是使流式处理足够简单，并使流式处理成为构建响应事件的事件驱动应用程序的自然方式，而不仅是处理大数据的一个重量级框架。在我们的模型中，主要实体不是用于数据处理的代码，而是 Kafka 中的数据流。

这是了解 Kafka Streams 以及 Kafka Streams 如何成为事件驱动应用程序的关键推动者的极好方式。我希望你和我一样喜欢本书！

——Neha Narkhede
Confluent 联合创始人兼首席技术官
Apache Kafka 联合创作者

前言

在我作为软件开发人员期间，我有幸在一些令人兴奋的项目上使用了当前软件。起初我客户端和后端都做，但我发现我更喜欢后端开发，因此我扎根于后端开发。随着时间的推移，我开始从事分布式系统相关的工作，从 Hadoop 开始（那时还是在 1.0 版本之前）。快进到一个新项目，我有机会使用了 Kafka。我最初的印象是使用 Kafka 工作起来非常简单，也带来很多的强大功能和灵活性。我发现越来越多的方法将 Kafka 集成到交付项目数据中。编写生产者和消费者的代码很简单，并且 Kafka 提升了系统的性能。

然后我学习 Kafka Streams 相关的内容，我立刻意识到："我为什么需要另一个从 Kafka 读取数据的处理集群，难道只是为了回写？"当我查看 API 时，我找到了我所需的流式处理的一切——连接、映射值、归约以及分组。更重要的是，添加状态的方法比我在此之前使用过的任何方法都要好。

我一直热衷于用一种简单易懂的方式向别人解释概念。当我有机会写关于 Kafka Streams 的书时，我知道这是一项艰苦的工作，但是很值得。我希望为本书付出的辛勤工作能证明一个事实，那就是 Kafka Streams 是一个简单但优雅且功能强大的执行流式处理的方法。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

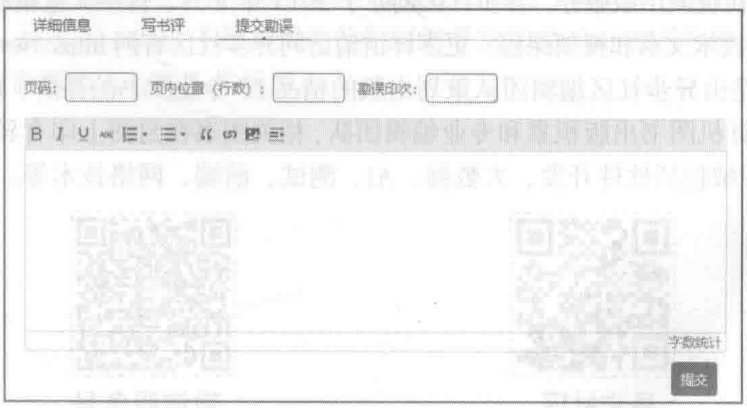
配套资源

本书提供源代码下载，要获得以上配套资源，请在异步社区本书页面中点击 配套资源 ，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，点击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

致谢

首先，我要感谢我的妻子 Beth，感谢她在这一过程中给予我的支持。写一本书是一项耗时的任务，没有她的鼓励，这本书就不会完成。Beth，你太棒了，我很感激你能成为我的妻子。我也要感谢我的孩子们，他们在大多数周末都忍受整天坐在办公室里的爸爸，当他们问我什么时候能写完的时候，我总模糊地回答“很快”。

接下来，我要感谢 Kafka Streams 的核心开发者 Guozhang Wang、Matthias Sax、Damian Guy 和 Eno Thereska。如果没有他们卓越的洞察力和辛勤的工作，就不会有 Kafka Streams，我也就没机会写这个颠覆性的工具。

感谢本书的编辑，Manning 出版社的 Frances Lefkowitz，她的专业指导和无限的耐心让写书变得很有趣。我还要感谢 John Hyaduck 提供的准确的技术反馈，以及技术校对者 Valentin Crettaz 对代码的出色审查。此外，我还要感谢审稿人的辛勤工作和宝贵的反馈，正是他们使本书更高质量地服务于所有读者，这些审稿人是 Alexander Koutmos、Bojan Djurkovic、Dylan Scott、Hamish Dickson、James Frohnhofer、Jim Manthely、Jose San Leandro、Kerry Koitzsch、László Hegedüs、Matt Belanger、Michele Adduci、Nicholas Whitehead、Ricardo Jorge Pereira Mano、Robin Coe、Sumant Tambe 和 Venkata Marrapu。

最后，我要感谢 Kafka 的所有开发人员，因为他们构建了如此高质量的软件，特别是 Jay Kreps、Neha Narkhede 和 Jun Rao，不仅是因为他们当初开发了 Kafka，也因为他们创办了 Confluent 公司——一个优秀而鼓舞人心的工作场所。

关于作者

William P. Bejeck Jr.（本名 Bill Bejeck），是 Kafka 的贡献者，在 Confluent 公司的 Kafka Streams 团队工作。他已从事软件开发近 15 年，其中有 6 年专注于后端开发，特别是处理大量数据，并在数据提炼团队中，使用 Kafka 来改善下游客户的数据流。他是 *Getting Started with Google Guava*（Packt，2013）的作者和“编码随想”（Random Thoughts on Coding）的博主。

关于本书

我写本书的目的是教大家如何开始使用 Kafka Streams，更确切地说，是教大家总体了解如何进行流式处理。我写这本书的方式是以结对编程的视角，我假想当你在编码和学习 API 时，我就坐在你旁边。你将从构建一个简单的应用程序开始，在深入研究 Kafka Streams 时将添加更多的特性。你将会了解到如何对 Kafka Streams 应用程序进行测试和监控，最后通过开发一个高级 Kafka Streams 应用程序来整合这些功能。

读者对象

本书适合任何想要进入流式处理的开发人员。虽然没有严格要求，但是具有分布式编程的知识对理解 Kafka 和 Kafka Streams 很有帮助。Kafka 本身的知识是有用的，但不是必需的，我将会教你需要知道的内容。经验丰富的 Kafka 开发人员以及 Kafka 新手将会学习如何使用 Kafka Streams 开发引人注目的流式处理应用程序。熟悉序列化之类的 Java 中、高级开发人员将学习如何使用这些技能来构建 Kafka Streams 应用程序。本书源代码是用 Java 8 编写的，大量使用 Java 8 的 lambda 语法，因此具有 lambda（即使是另一种开发语言）程序的开发经验会很有帮助。

本书组织结构：路线图

本书有 4 部分，共 9 章。第一部分介绍了一个 Kafka Streams 的心智模型，从宏观上向你展示它是如何工作的。以下章节也为那些想学习或想回顾的人提供了 Kafka 的基础知识。

- 第 1 章介绍流式处理如何以及为何成为处理大规模实时数据的必需方式的历史，并提出 Kafka Streams 的心智模型，没有详细介绍任何代码，而是描述 Kafka Streams 是如何工作的。
- 第 2 章为 Kafka 新手介绍一些 Kafka 入门知识。Kafka 经验丰富的读者可以跳过这一章，直接进入 Kafka Streams。

第二部分继续讨论 Kafka Streams，从基础 API 开始，一直到更复杂的特性，第二部分各章介绍如下。

- 第 3 章介绍一个 Hello World 应用程序，然后介绍一个更实际的应用程序示例——为虚构的零售商开发应用程序，包括高级特性。
- 第 4 章讨论状态，并解释流式应用程序有时是如何需要状态的。同时读者还将了解如何实现状态存储以及如何在 Kafka Streams 中执行连接。
- 第 5 章讨论表和流的二元性，并引入一个新概念——`KTable`。KStream 是事件流，而 `KTable` 是相关事件的流或者更新流。
- 第 6 章介绍低阶处理器 API。到此时，一直使用的是高阶 DSL，但是在这里，读者将学习如何在编写应用程序的自定义部分时使用处理器 API。

第三部分将从开发 Kafka Streams 应用程序转到对 Kafka Streams 的管理知识的讨论。

- 第 7 章介绍如何监控 Kafka Streams 应用程序，以查看处理记录所需要的时间以及定位潜在的处理瓶颈。
- 第 8 章介绍如何测试 Kafka Streams 应用程序。读者将学习如何对整个拓扑进行测试，对单个处理器进行单元测试，以及使用嵌入式 Kafka 代理进行集成测试。

第四部分是本书的压轴部分，在这里你将深入研究使用 Kafka Streams 开发高级应用程序。

- 第 9 章介绍使用 Kafka Connect 将现有的数据源集成到 Kafka Streams 中。你将会学习如何在流式应用程序中包括数据库表。然后你将看到数据在 Kafka Streams 中流动时如何使用交互式查询来提供可视化和仪表板应用程序，而无需关系型数据库。这一章还会介绍 KSQL，可以使用它在 Kafka 运行连续的查询，除了使用 SQL 之外并不需要编写任何代码。

关于代码

本书包含了很多源代码的例子，包括书中编号的代码清单所标明的代码，以及内联在普通文本中的代码。在这两种情况下，源代码都采用固定宽度字体的格式，以便与普通文本区分开。

在很多情况下，原始源代码已经被重新格式化了。我们增加了断行以及重新缩进，以适应书中可用的页面空间。在极少数情况下，甚至空间还不够，代码清单中包括续行标识（➥）。此外，当在文本中描述代码时，源代码中的注释常常从代码清单中删除。代码清单中附带的许多代码注释，突出显示重要的概念。

最后，需要注意的是：许多代码示例并不是独立存在的，它们只是包含当前讨论的最相关部分代码的节选。你在本书附带的源代码中将会找到所有示例的完整代码。

本书的源代码是使用 Gradle 工具构建的一个包括所有代码的项目。你可以使用合适的命令将项目导入 IntelliJ 或 Eclipse 中。在附带的 README.md 文件中可以找到使用和导航源代码的完整说明。

图书论坛

购买本书可以免费访问一个由 Manning 出版社运营的私人网络论坛，可以在论坛上对本书进行评论、咨询技术问题、接受本书作者或者其他用户的帮助。要访问该论坛，请访问 Manning 出版社官方网站本书页面。你还可以从 Manning 出版社官方网站了解更多关于 Manning 论坛及其行为规则。

Manning 的论坛承诺为我们的读者提供一个可以在读者之间，以及读者与作者之间进行有意义对话的地方，但并不承诺作者的参与程度，作者对论坛的贡献是自愿的（并没有报酬）。建议你试着问他一些有挑战性的问题，以免他对你的问题没有兴趣！只要本书在印刷中，论坛和之前所讨论的问题归档就会从出版社的网站上获得。

其他在线资源

- Apache Kafka 文档：见 Apache Kafka 官方网站。
- Confluent 文档：见 Confluent 官方网站。
- Kafka Streams 文档：见 Confluent 官方网站。
- KSQL 文档：见 Confluent 官方网站。

关于封面插图

本书封面上的图片描述的是“18 世纪一位土耳其绅士的习惯”，这幅插图来自 Thomas Jefferys 的 *A Collection of the Dresses of Different Nations, Ancient and Modern*（共 4 卷），于 1757 年和 1772 年之间出版于伦敦。扉页上写着：这些是手工着色的铜版雕刻品，用阿拉伯胶加深了颜色。Thomas Jefferys（1719—1771）被称为“乔治三世的地理学家”。他是一位英国制图师，是当时主要的地图供应商。他为政府和其他官方机构雕刻和印刷地图，制作了各种商业地图和地图集，尤其是北美地区的。作为一名地图制作者，他在所调查和绘制的地区激起了人们对当地服饰习俗的兴趣，这些都在这本图集中得到了很好的展示。向往远方、为快乐而旅行，在 18 世纪后期还是相对较新的现象，类似于这套服饰集的书非常受欢迎，把旅行者和神游的旅行者介绍给其他国家的居民。Jefferys 卷宗中绘画的多样性生动地说明了 200 多年前世界各国的独特性和个性。从那时起，着装样式已经发生了变化，各个国家和地区当时非常丰富的着装多样性也逐渐消失。现在仅依靠衣着很难把一个大陆的居民和另一个大陆的居民区分开来。或许我们已经用文化和视觉上的多样性换取了个人生活的多样化——当然是更为丰富和有趣的文化和艺术生活。

在一个很难将计算机书籍区分开的时代，Manning 以两个世纪以前丰富多样的地区生活为基础，通过以 Jefferys 的图片作为书籍封面来庆祝计算机行业的创造性和首创精神。