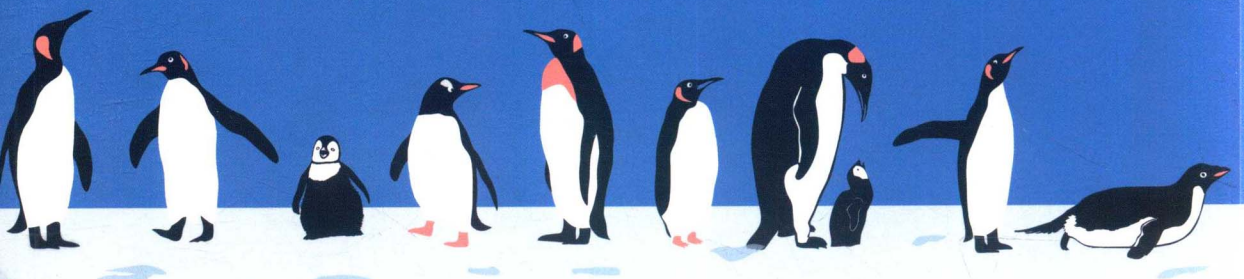
Broadview®
www.broadview.com.cn

Apache Kafka 实战

胡夕 著



- 基于Apache Kafka 1.0.0版本进行介绍，Kafka Contributor执笔。
- 从Kafka基本概念与特性开始，详细介绍了Kafka的部署、开发、运营、监控、调试、优化以及重要组件的设计原理，并给出了翔实的案例。
- 本书既适合作为Kafka的入门书籍，也适合系统架构师和一线开发工程师参考阅读。

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Apache Kafka 实战

Apache Kafka实战

胡夕 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内容简介

本书是涵盖 Apache Kafka 各方面的具有实践指导意义的工具书和参考书。作者结合典型的使用场景，对 Kafka 整个技术体系进行了较为全面的讲解，以便读者能够举一反三，直接应用于实践。同时，本书还对 Kafka 的设计原理及其流式处理组件进行了较深入的探讨，并给出了翔实的案例。

本书共分为 10 章：第 1 章全面介绍消息引擎系统以及 Kafka 的基本概念与特性，快速带领读者走进 Kafka 的世界；第 2 章简要回顾了 Apache Kafka 的发展历史；第 3 章详细介绍了 Kafka 集群环境的搭建；第 4、5 章深入探讨了 Kafka 客户端的使用方法；第 6 章带领读者一览 Kafka 内部设计原理；第 7~9 章以实例的方式讲解了 Kafka 集群的管理、监控与调优；第 10 章介绍了 Kafka 新引入的流式处理组件。

本书适合所有对云计算、大数据处理感兴趣的技术人员阅读，尤其适合对消息引擎、流式处理技术及框架感兴趣的技术人员参考阅读。

图书在版编目（CIP）数据

Apache Kafka 实战 / 胡夕著. —北京：电子工业出版社，2018.5
ISBN 978-7-121-33776-5

Ⅰ. ①A… Ⅱ. ①胡… Ⅲ. ①分布式操作系统 Ⅳ. ①TP316.4

中国版本图书馆 CIP 数据核字（2018）第 037942 号

责任编辑：付　睿
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：25　　字数：557 千字
版　　次：2018 年 5 月第 1 版
印　　次：2018 年 5 月第 1 次印刷
定　　价：89.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

前言

2011 年年初，美国领英公司（LinkedIn）开源了一款基础架构软件，以奥地利作家弗兰兹·卡夫卡（Franz Kafka）的名字命名，之后 LinkedIn 将其贡献给 Apache 基金会，随后该软件于 2012 年 10 月成功完成孵化并顺利晋升为 Apache 顶级项目——这便是大名鼎鼎的 Apache Kafka。历经 7 年发展，2017 年 11 月，Apache Kafka 正式演进到 1.0 时代，本书就是基于 1.0.0 版本来展开介绍 Kafka 的设计原理与实战的。

背景

这是一个最好的大数据时代，这是一个最坏的大数据时代！

很抱歉，我使用了这句改编后的狄更斯名言作为开头，我想没有谁会质疑“当今是大数据时代”这个论点。今年（2018 年）两会上李克强总理所做的政府工作报告中多次提及大数据等关键词，这已然是“大数据”第 5 次被写入政府工作报告了。具体到大数据行业内，各种各样的大数据产业方兴未艾，其中在实时流式处理领域涌现出大量的技术与框架，令技术人员们应接不暇。实时流式处理系统在克服了传统批处理系统延时方面的固有缺陷的同时，还摆脱了设计上的桎梏，实现了“梦寐以求”的正确性。可以说，对于流式处理从业人员来说，这正是摩拳擦掌、大展宏图的最好时代。

与此同时，我们也清醒地意识到当今大数据领域内的细分越来越精细化。不必说日渐火爆的人工智能和机器学习潮流引诱着我们改弦易辙，也不必说那些纷繁复杂的技术框架令人眼花缭乱，单是静下心来沉淀所学、思考方向的片刻时光于我们这些从业者来说都已显得弥足珍贵。我们仿佛在黑暗密林中徘徊，试图找出那条通往光明的“康庄大道”。每当发现了一条羊肠小路都好似救命稻草一般紧紧抓住。多年后我们回望那只不过是不断追逐热点罢了，在技术的海洋中我们迷失了前进的方向。从这个意义上说，这实在是一个糟糕的时代。

时光切回到 4 年前的某个下午，那时我正在做着 Kafka 的大数据项目。我突然发现与其盲目跟风各种技术趋势，何不精进手头的工作，把当前工作中用到的技术搞明白，于是我萌发了

研究 Kafka 的想法。直到今天，我都无比庆幸那个午后做出的冲动决定，正如 Adam Grant 在《离经叛道》一书中所说：最正确的决定都是在冲动之下做出的。诚不欺我！

想要深入学习 Kafka，不掌握 Scala 语言是不行的，毕竟 Kafka 就是使用 Scala 语言编写的。苦于当时没有合适的 Scala 中文书籍，我依稀记得找到了一本 600 多页的 Scala 原版书（*Programming Scala Edition 2*）进行学习。那段时间实在是难熬！不得不说，英文版书籍虽然内容翔实，但在表述上实在晦涩难懂，比如 partially applied function 和 partial function 两者的区别直至今天我都不是特别清晰，还是要不断地翻阅资料才能隐约记得它们之间的不同。庆幸的是，我没有半途而废，600 多页的英文文档硬是啃了下来。对于 Scala 的初步掌握也让我觉得研究 Kafka 的时机到了。有意思的是，在之后通读 Kafka 的源码时我不禁大呼上当，Kafka 的源码中只使用了最简单的函数式编程，我有些后悔自己花了那么多时间去学习 Scala 的函数式编程，当然这是后话。

既然是研究 Kafka，那么研读源码是必不可少的步骤。如果不分析源码，我们就无法定位问题发生的根本原因。实话实说，阅读别人源码的过程是痛苦的，因而在理解的过程中我走了不少弯路。为了记录阅读 Kafka 源码的心得，我努力为每个 Kafka 源码包撰写博客。现在翻看我之前的博客，大家还能看到那好似流水账一般的 Kafka 源码分析系列文章。

随着对源码的不断熟悉，我加入了 Apache Kafka 社区，希望贡献自己的微薄之力。时至今日，我依然记得当初发送邮件要求加入开发组时的惶恐，也记得第一次贡献代码时的惴惴不安；我记得为了研究某个 Kafka bug，自己曾忘记吃中饭的执着，也记得自己被标记为“Kafka contributor”时的喜悦。在混迹社区的日子里，我逐渐认识了一些 Kafka 的 committer 们，比如 Kafka PMC 成员王国璋，国璋兄对于网上 Kafka 问题的权威解答令我受教良多，同时我也很感激他于百忙之中为本书写推荐语。还有 Kafka 的三位原作者之一的饶军（Rao Jun），几次问题交流让我看到了他霸气的决断能力以及对于疑难问题原因的毒辣分析。当然还有非常敬业的 Ijuma，他是我见过的最勤劳的 Kafka committer，没有之一。在编写本书的过程中，我都或多或少地得到过他们的帮助，再次表示衷心感谢。

由于对 Kafka 研究的日益深入，我终于有了写书的冲动。我希望通过把学到的知识和原理集中整理并书写成文字来帮助那些尚未接触 Kafka 的广大读者快速上手，降低他们学习使用 Kafka 的成本，于是有了今天这本《Apache Kafka 实战》。借着写作本书的契机，我本人对 Kafka 的方方面面做了梳理，自觉收获良多。每当搞懂了一个以前未了解的机制时，心中的那种满足感和兴奋感至今都令人神往。在此，我深深地希望读者在阅读完本书后也能有这样的体会。

面向的读者

我衷心希望本书可以成为各行各业的大数据从业者使用消息队列甚至是进入流式处理领域

内的“敲门砖”，也希望各大公司能够充分利用 Kafka 来实现自己的业务目标。

在编写本书的过程中，我阅读了大量的英文资料和源代码，试图通过自己的理解将 Kafka 的使用实战技巧深入浅出地呈现给广大读者。没错，我希望这本书给人的感觉是通俗易懂、深入浅出，从而方便引领读者快速进入 Kafka 学习的大门。

我本人维护了一个微信公众号（名为“大数据 Kafka 技术分享”），希望在该公众号中我能和读者朋友们一起深入交流和探讨 Kafka 学习过程中碰到的各种问题，同时我也会及时分享和推送各种最新的 Kafka 使用心得。

致谢

非常感谢 Kafka PMC 成员、Kafka Committer 王国璋对本书的大力支持。自开始编写本书之日起，国璋兄就给予我很大的鼓励与帮助，这也让我坚定了传播 Kafka 实战心得的决心。

感谢腾讯 AI 平台助理总经理王迪先生和我的好友贾兴华，你们对本书的评价之高实在是过誉了，但也令本人倍感振奋。

感谢我的前同事、新浪微博技术专家付稳。付总对本书整体结构和具体知识点的建议发人深省，其独到的行业见解令人佩服。

非常感谢电子工业出版社的编辑付睿女士。她细致、专业、严谨的工作作风深深地感染了我，在本书编写过程中她总是能及时地就书中的内容给出合理的建议和指导。

另外，我还想感谢一下我的家人，特别是我的妻子刘丹女士。过去一年中正是你坚定的支持和默默的付出才成就我撰写本书。对于你偶尔在学术上给予的提点我既感到惊讶，同时也欣慰不已。这为我漫长枯燥的写书过程平添了很多温暖。

最后，非常感谢本书的每一位读者。本人已经在写作过程中收获良多，我衷心希望你们在阅读本书时也有大呼过瘾的感觉。另外，我在“知乎”（ID：huxihx）的 Kafka 专栏以及 StackOverflow 网站上也会尽力回答关于 Kafka 的各类问题，希望通过这些途径可以和读者进行更加深入的交流。

由于本人水平有限，书中难免有遗漏和疏忽，也恳请各位读者多多指正。

胡夕

2018 年 3 月 15 日于北京

个人博客：https://www.cnblogs.com/huxi2b/

微信公众号：大数据 Kafka 技术分享

电子邮箱：huxi_2b@hotmail.com

目录

第 1 章
认识 Apache Kafka

随着大数据时代的到来，数据中蕴含的价值日益得到展现，仿佛一座待人挖掘的金矿，引来无数的掘金者。但随着数据量越来越大，如何实时准确地收集并分析数据成为摆在所有从业人员面前的难题。

本章作为本书的第 1 章，将带领读者对 Apache Kafka 系统及其生态圈建立一个宏观的概念和认识。同时，本章将结合消息引擎系统的相关知识与设计理念，循序渐进地对 Kafka 系统的设计架构和相关概念进行展开，并给出简单示例以快速上手 Kafka。

学习本章，你将了解到以下内容。

- 消息引擎系统的定义与特点。
- Apache Kafka 系统概要设计及术语。
- Apache Kafka 快速入门。

1.1 Kafka 快速入门

Kafka 的核心功能是什么？一言以蔽之，高性能的消息发送与高性能的消息消费。接下来，本书打算先给出 Kafka 的快速入门教程，即 Kafka 的“Hello, world”示例。这么做与其他书不同，其目的就是可以让各位读者快速体验 Kafka 的核心功能——消息的发送与消费。这对于很多不太熟悉 Kafka 的初学者来说，可以让他们瞬间提升学习的快感，从而增强对于掌握与理解 Kafka 系统的信心。

作为消息引擎系统中的佼佼者，Kafka 诸多独特的设计理念及强大的功能特性非常值得学习，但现在我们首先跑通一个 Kafka 的简单示例，切身感受一下 Kafka 这个目前消息引擎领域