HZ BOOKS
华章IT

MANNING

基于Lambda架构的系统构建

Big Data

Principles and Best Practices of Scalable Realtime Data Systems

大数据系统构建

可扩展实时数据系统构建原理与最佳实践

[美] 南森·马茨（Nathan Marz）
詹姆斯·沃伦（James Warren） 著

马延辉 向磊 魏东琦 译

械工业出版社
a Machine Press

大数据技术丛书

Big Data

Principles and Best Practices of Scalable Realtime Data Systems

大数据系统构建

可扩展实时数据系统构建原理与最佳实践

[美] 南森·马茨（Nathan Marz） 詹姆斯·沃伦（James Warren） 著

马延辉 向磊 魏东琦 译

机械工业出版社 China Machine Press

图书在版编目（CIP）数据

大数据系统构建：可扩展实时数据系统构建原理与最佳实践 /（美）南森·马茨（Nathan Marz），（美）詹姆斯·沃伦（James Warren）著；马延辉，向磊，魏东琦译．—北京：机械工业出版社，2016.12
（大数据技术丛书）

书名原文：Big Data：Principles and Best Practices of Scalable Realtime Data Systems

ISBN 978-7-111-55294-9

I. 大… II. ① 南… ② 詹… ③ 马… ④ 向… ⑤ 魏… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2016）第 262539 号

本书版权登记号：图字：01-2015-7585

大数据系统构建
可扩展实时数据系统构建原理与最佳实践

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：吴晋瑜　　责任校对：殷　虹
印　　刷：北京诚信伟业印刷有限公司　　版　　次：2017 年 1 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：18.75
书　　号：ISBN 978-7-111-55294-9　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

The Translator's Words 译者序

首先，请允许我们对 Nathan Marz 致以崇高的敬意。

Nathan Marz 是分布式实时计算系统 Storm 的创始人，在 Twitter 收购社交媒体数据分析公司 BackType 前担任 BackType 的首席工程师，之后选择离开 Twitter，创立自己的公司。在实时大数据处理系统中，Storm 作为 Apache 顶级开源项目已经成为大数据界不可或缺的一部分。因此，对于能够翻译 Nathan Marz 的书籍，我们深感荣幸。

与大多数程序员一样，Nathan Marz 也是通过游戏进入开发者的世界的，在这一点上，似乎我们大多数人与 Nathan Marz 相差无几。但不同的是，Nathan Marz 开创性地设计并使用 Clojure 语言编写了 Storm，为我们揭开了大数据处理的新篇章，而我们未曾想过海量数据是可以实时分析并处理的，这也正是他与众不同的地方。Nathan Marz 对大数据概念的理解非常深刻，在编程技术上基础扎实，如同 Dean Jeffrey 和 Doug Cutting 那样，他用自己超凡的智慧，带领我们步入了一个全新的数据时代。

本书借一些虚构的社交媒体示例，来让读者深入理解以下几件事情：

1）什么是大数据，它们从哪里来？

2）社交媒体有哪些数据是有价值且需要我们去分析的？

3）在使用数据的过程中，我们需要用哪些思路、架构、工具来实现自己的目的？

4）对于不同的数据类型，我们如何选择正确的架构和模型去进行分析和挖掘？

在翻译的过程中，我们也了解到，Nathan Marz 不仅在数学与编程方面才华横溢，对各种开发工具与架构也是信手拈来，而且他所写的书籍也是字字珠玑，文不加点。他所写的内容深邃却并不晦涩，浅显易懂，贴近实战，原作行文流畅，文采炳焕。本书将大数据方方面面的工具以实例的形式引入内容中，令人读后有一种酣畅淋漓、耳目一新的感觉，在内容方面，从 Apache Thrift 的讲解到 Lambda 架构的实例、从 HDFS 和 MapReduce 的示范到架构和算法的实现以及针对不同类型数据模型的创建，一一涵盖其中。可以说，本书是

大数据技术的集大成者，是诸多大数据书籍中难得一见的实战参考书。

对于我们译者来说，之所以翻译本书，既是希望将国外实践大数据技术的重要经验引入国内，让国内的读者能够从中一窥究竟，同时也希望自己在翻译的过程中有所受益。站在巨人的肩膀上，才能让我们能够看得更远。

在本书的翻译过程中，我们得到了诸多朋友和家人的帮助、理解以及支持，在此对他们表示衷心的感谢。同时也对促成本书出版的机械工业出版社的王春华、杨福川编辑表示诚挚的谢意。

本书内容丰富，涵盖了大数据的诸多方面，如 Thrift、数据建模、HDFS、MapReduce、HBase、Lambda 等，这为本书的翻译增加了不少难度。尽管我们进行了多次校对和修改，甚至几位译者就某些专业词汇如何准确翻译进行了多次字斟句酌的讨论，但由于水平所限，恐难以将原作的内容全面还原，因此也难免出现纰漏和不足。在此，也恳请广大读者在阅读之余不吝赐教，给予批评指正。

向　磊

2016 年 10 月于北京

Preface 前　言

当第一次进入大数据的世界时，我仿佛置身于软件开发的美国西部荒原。许多人放弃了关系型数据库，转而选择带有高度受限模型的 NoSQL 数据库，主要是因为其使用体验良好、熟悉度较高且这种数据库可以扩展到成千上万台机器上。NoSQL 数据库的数量巨大，堪称铺天盖地，这些数据库中很多都只有细微的差别。一个名为“Hadoop”的新项目开始崭露头角，它宣称具备基于海量数据进行数据深度分析的能力。但弄清楚如何使用这些新工具很令人困惑。

当时，我正试图处理所在公司面临的扩展性问题。系统架构非常复杂——该 Web 系统包含共享关系型数据库、队列、工作节点、主节点和从节点。数据损坏渗透至数据库，为了处理这些损坏，我们使用了应用程序中的特殊代码，但从节点的操作总是落后于其他节点。我决定探索其他大数据技术，看看是否有比我们的数据架构更好的设计。

早期的软件工程职业生涯的经历，深刻影响了我对“系统该如何架构”的观点。我的一位同事花了几个星期将来自互联网的数据收集到一个共享文件系统。他在等待收集足够的数据，以便能在其上进行数据分析。有一天，在做一些日常维护时，我不小心删除了他的所有数据，导致他的项目延期了好几周。

我知道自己犯了一个大错，但作为一个软件工程师新手，我并不知道这会导致什么样的后果。我会不会因为粗心被解雇呢？我发了一封电子邮件向团队诚挚地道歉——让我惊喜的是，大家对此都表示非常同情。我永远不会忘记那个时刻——一个同事来到我的办公桌旁，拍着我的背说：“恭喜你！你现在是一个专业的软件工程师了！”

他玩笑式的表述道出了软件开发中不言而喻的“真理”——我们不知道如何创造完美的软件。软件可能有 bug 而且会被部署到生产中。如果应用程序可以写入数据库中，那么 bug 也可能写入数据库中。当着手重新设计我们的数据架构时，这样的经历深深地影响了我。我知道，新架构不但必须是可扩展的、对机器故障是可容错的，并且要易于推断故障

原因——但对人为错误也可容错。

重构那套系统的经验，促使我走上了一条“在数据库和数据管理方面怀疑一切我认为是正确的”道路。我想出了一个基于不可变数据和批量计算的架构，令我很惊讶的是，与仅仅基于增量计算的系统相比，新系统要简单得多。一切都变得更容易，包括操作、不断发展的系统以支持新的功能、从人为错误中恢复和性能优化等方面。该方法很通用，似乎可以用于任何数据系统。

但有些事情困扰着我。当观察其他行业时，我发现几乎没有人使用类似的技术。相反，在使用基于增量更新数据库的庞大集群架构中，令人生畏的复杂性是为人所接受的。这些架构的许多复杂性已经通过我所开发的方法完全避免或大大缓减了。

在接下来的几年中，我扩展了该方法，并使之正式成为我戏称的 Lambda 架构。在初创公司 BackType 工作时，我们的 5 人团队构建了一个社会化媒体分析产品，该产品支持在超过 100TB 的数据上进行多样化实时分析。我们的小团队还负责拥有数百台机器的集群的管理部署、运营和系统监控。当我们向别人展示自己的产品时，他们对这个团队只有 5 个人感到非常惊讶。他们经常会问“这么几个人做了这么多事情？怎么可能！？”我的回答很简单：“不是我们在做什么，而是我们没有做什么。”通过使用 Lambda 架构，我们避免了困扰传统架构的复杂性。通过避免这些复杂性，我们大大提高了工作效率。

大数据运动只是放大了已经存在了几十年的数据架构的复杂性。主要基于增量更新的大型数据库架构将遭受这些复杂性的折磨，从而导致错误、繁重的操作，并阻碍了生产力。尽管 SQL 和 NoSQL 数据库通常被描述成对立或相互对偶的关系，但从最基本的方面来说，它们实际上是一样的。它们都鼓励使用这种相同的架构——该架构具有不可避免的复杂性。复杂性是一个邪恶的野兽，无论你承认与否，它都会“咬”你。

为了传播 Lambda 架构以及它如何避免传统架构的复杂性等知识，我写了本书。它是我开始从事大数据工作时就希望有的。我希望你把这本书作为一个旅程——挑战你以为自己已经知道的关于数据系统的知识，并发现从事大数据工作也可以优雅、简单和有趣。

Nathan Marz

About This Book 关于本书

类似社交网络、网络分析和智能型电子商务这样的服务，通常需要在非常大规模的传统数据库上管理数据。复杂性随着规模与需求的增加而增加，而处理大数据并不是简单地将 RDBMS 扩大一倍或推出一些时髦的新技术。幸运的是，可扩展性和简单性并不是相互排斥的——你只需要采取不同的方法。大数据系统使用多台机器并行工作来存储和处理数据，它引入了大多数开发者并不熟悉的根本性的挑战。

本书将教你充分利用集群硬件优势的架构，以及专门用来捕获和分析网络规模数据的新工具，来创建这些系统。它将描述一个可扩展的、易于理解大数据系统的方法，可以由小团队构建并运行。本书利用一个实际示例，基于大数据系统的理论在实践中实现它们来指导读者。

本书不要求读者以前接触过大规模数据分析或 NoSQL 工具。熟悉传统数据库是有帮助的，但不是必需的。本书旨在教你如何思考数据系统，以及如何化繁为简。我们将从基本原理开始，从那些被认定为架构的各个组件所必需的属性开始。

路线图

本书包括 18 章，各章的主要内容如下。

第 1 章介绍了数据系统的原理，并给出了 Lambda 架构的概述：构建任何数据系统的广义方法。第 2～17 章以理论和示例交替讲解的方式深入介绍 Lambda 架构的所有内容。理论章节阐述了若干概念，这些概念对现有工具都是适用的；同时，示例章节使用现实世界中的工具来论证这些概念。不要被名字迷惑，虽然所有章节都是样例驱动。

第 2～9 章集中阐述 Lambda 架构的批处理层。在这里，你将了解如何为主数据集建模、如何使用批处理来创建数据的任意视图，以及如何进行增量和批处理之间的权衡。

第 10 章和第 11 章集中阐述服务层，它支持低延迟访问由批处理层中产生的视图。在这里，你将了解只批量写入的特定数据库。你将发现，这些数据库比传统数据库更简单，它们具有出色的性能，并具备可操作性、稳健性等特性。

第 12～17 章集中阐述速度层，该层弥补了批处理层的高延迟，为所有查询提供最新结果。在这里，你将了解 NoSQL 数据库、流处理和管理增量计算的复杂性。

第 18 章再次复习 Lambda 架构的相关知识，并进行查漏补缺。你将了解增量批处理、基本 Lambda 架构的变种，以及如何充分利用资源。

代码下载和约定

本书的源代码可以在 https://github.com/Big-Data-Manning 找到。我们提供了运行示例 SuperWebAnalytics.com 的源代码。

大量源代码以代码清单的形式给出。这些代码清单提供了完整的代码段。一些代码带有注释，以着重强调或解释某部分代码。在正文的其他地方，代码片段会在必要时使用。Courier 字体用来表示 Java 代码。在代码中，我们用粗体字来帮助你识别文本中的关键部分。

作者在线

购买本书的读者将可免费使用由 Manning 出版社运营的私人网络论坛。你可以在论坛上评论本书、提出技术问题，并从作者和其他用户处得到帮助。要访问和订阅该论坛，请在 Web 浏览器输入“www.manning.com/BigData”。在论坛上注册后，你可在作者在线（AO）页面查看如下信息：注册后如何登录论坛、哪些帮助可用以及论坛上的行为规则。

Manning 承诺为读者提供一个场所，以供个体读者之间、读者和作者之间进行有意义的对话。这并不是作者承诺进行任何特定数量的分享，作者对 AO 论坛的贡献是自愿的（无报酬的）。我们建议你尝试问作者一些有挑战性的问题，以免他们觉得了然无趣！

只要本书已出版，AO 论坛和以前讨论的归档文件即可以从发行商的网站访问。

关于封面插图

本书的封面插图是“Le Raccommodeur de Fiance”，意思是泥瓦匠。泥瓦匠擅长修补破

致　谢 *Acknowledgements*

如果没有周围很多人的帮助，这本书是无法完成的。我必须首先感谢我的父母，他们一直给我灌输热爱学习和探索世界的思想，并一直鼓励我在职业生涯中孜孜不倦地追求。

我的哥哥 Iorav 也一直在学术兴趣上给予我鼓励。我还记得，在我读小学时，他就教我学习代数。我第一次接触编程也是他介绍的——他教我 Visual Basic，因为他在高中学过该课程。这些课程引发了我对编程的热情，引导我走上了职业道路。

非常感谢 Michael Montano 和 Christopher Golda——BackType 的创始人。从他们把我作为他们的第一个员工开始，我就拥有可以自己做决定的极度自由。这种自由对我探索和最大限度地利用 Lambda 架构来说至关重要。他们从未质疑过开源的价值，并允许我自由地开源我们的技术。深入地参与开源已经成为我的特权之一。

特别感谢我在斯坦福学习时遇到的很多教授。Tim Roughgarden 是我见过的最好的老师——他从根本上提高了我严格分析、解构和解决困难问题的能力。尽可能多地上他的课是我一生中做出的最好决定之一。也感谢 Monica Lam 给我灌输了对 Datalog 优雅性的欣赏。许多年后我将 Datalog 与 MapReduce 结合，生成了平生第一个意义重大的开源项目——Cascalog。

Chris Wensel 是第一个给我展示大规模的数据处理也有可能优雅和高效的人。他的 Cascading 库改变了我看待大数据处理的方式。

如果没有大数据领域的先驱者，我的工作是不可能实现的。特别感谢最早的 MapReduce 论文的作者 Jeffrey Dean 和 Sanjay Ghemawat。感谢最初的 Dynamo 论文的作者 Giuseppe DeCandia、Deniz Hastorun、Madan Jampani、Gunavardhan Kakulapati、Avinash Lakshman、Alex Pilchin、Swaminathan Sivasubramanian、Peter Vosshall 和 Werner Vogels。感谢 Apache Hadoop 项目的创始人 Michael Cafarella 和 Doug Cutting。

在我的编程生涯中，Rich Hickey 一直是给我最多灵感的人之一。Clojure 是我至今用过的最好的语言，通过学习它，我成了一名更好的程序员。我很欣赏它的实用性和专注于简单性的特性。Rich 在编程的状态和复杂性理念方面已经深深地影响了我。

开始写这本书时，我几乎称不上是作者。Renae Gregoire，Manning 公司负责本书的策划编辑之一，特别感谢她帮助我提高写作技巧。她让我认识到使用例子来引入通用概念的重要性。在如何有效地组织技术写作方面，她给了我很多灵感。她教给我的技巧不仅适用于写作技术书籍，还适用于写博客、演讲和平时的沟通。因为掌握了一项重要的生活技能，所以我对她永远心存感激。

如果没有我的合著者 James Warren 的努力，本书将不会有现在的质量。为了让读者能够理解理论概念，并找到展示这些理论的更好方式，他做了大量的工作。本书之所以能够如此明晰，大部分是源于他强大的沟通技巧。

与我的出版商 Manning 合作是一种乐趣。他们的员工对我很耐心，并且理解寻找合适的方式写出这样一个大的话题是需要时间的。在整个过程中，他们都很支持我并帮助我，他们总是提供给我成功所需的资源。感谢 Marjan Bace 和 Michael Stephens 的支持，还有所有其他员工的帮助和全程指导。

我尝试尽可能多地学习其他作家的写作风格。Bradford Cross、Clayton Christensen、Paul Graham、Carl Sagan 和 Derek Sivers 对我的影响很大。

最后，十分感谢对本书提出意见、评论和反馈的数百余名读者。这些反馈促成了本书的多次修改、重写和重组架构，直到我们找到有效地展示材料的方法。在此特别感谢 Aaron Colcord、Aaron Crow、Alex Holmes、Arun Jacob、Asif Jan、Ayon Sinha、Bill Graham、Charles Brophy、David Beckwith、Derrick Burns、Douglas Duncan、Hugo Garza、Jason Courcoux、Jonathan Esterhazy、Karl Kuntz、Kevin Martin、Leo Polovets、Mark Fisher、Massimo Ilario、Michael Fogus、Michael G. Noll、Patrick Dennis、Pedro Ferrera Bertran、Philipp Janert、Rodrigo Abreu、Rudy Bonefas、Sam Ritchie、Siva Kalagarla、Soren Macbeth、Timothy Chklovski、Walid Farid 和 Zhenhua Guo。

Nathan Marz

在回想为本书做出贡献的人时，我很震惊这么多人帮助过我。虽然不能一一列举，但这并不能减轻我的谢意。尽管如此，我还是希望向一些人明确表达我的感激之情：

- 我的妻子 Wen-Ying Feng——感谢你的爱、鼓励和支持，不仅是这本书，还有我们

共同完成的一切。

- 我的父母 James 和 Gretta Warren——感谢你们带给我的无穷信仰，还有为我提供每个机会所做出的牺牲。
- 我的姐姐 Julia Warren-Ulanch——感谢你树立了一个光辉的榜样，使我可以跟随你的脚步。
- 我的两位导师 Ellen Toby 和 Sue Geller 教授——感谢你们悉心回答我的每个问题，并教导我学习的乐趣不仅源于获得知识，更在于分享知识。
- Chuck Lam——感谢你很多年前对我说："嘿，你听说过一个叫 Hadoop 的东西吗？"
- 我的朋友和 RockYou!、Storm8、Bina 的同事——感谢我们一起共享的经历和把理论运用到实践的机会。
- Marjan Bace、Michael Stephens、Jennifer Stout、Renae Gregoire 和整个 Manning 的编辑出版人员——感谢你们在本书出版过程中的指导和耐心。
- 本书的评论者和早期读者——感谢你们的评论和批评，推动我们更加清晰地阐述理论；感谢你们让这本书变得更好。

最后，我要对 Nathan 表达最真挚的谢意，感谢你邀请我一起完成本书。在加入这个项目之前，我已经非常仰慕你的工作，共事时因为你的想法和理念，使我更加尊重你。这是一项莫大的荣誉和特权。

James Warren

Contents 目　录

第一部分　批处理层