Broadview
www.broadview.com.cn

PEARSON

寻路大数据

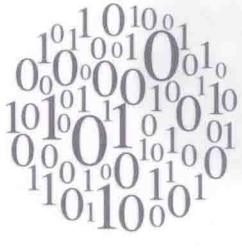
海量数据与大规模分析

Data Just Right

Introduction to Large-Scale Data & Analytics

[美] **Michael Manoochehri** 著

戴志伟 许杨毅 鄢博 陈冠诚 译



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

寻路大数据

海量数据与大规模分析

Data Just Right

Introduction to Large-Scale Data & Analytics

[美] **Michael Manoochehri** 著

戴志伟 许杨毅 鄢博 陈冠诚 译

電子工業出版社

Publishing House of Electronics Industry

北京•BEIJING

内容简介

这是一个数据爆发的时代，更是一个数据技术爆发的时代，各行各业都在因此进行深刻的变革。如何从众多的数据技术中选择正确的工具、如何使用这些工具从海量数据中挖掘出有价值的东西，无疑是非常具有挑战性的问题。

本书作者结合自己在 Google 大数据平台工作的丰富经验，阐述了数据技术的方方面面。从数据收集、共享到数据存储，从分布式数据平台、分析型数据库到数据可视化，从数据工作流构建到大规模数据分析，作者不仅进行了全面而深入的介绍，更覆盖了目前流行的各种数据技术与工具，同时对技术选型提出了指导性的建议。最后，作者对数据挑战的非技术因素进行了深刻的分析，并对数据技术的发展趋势进行了展望，引人深思。

本书对企业管理者、技术经理、数据分析师、数据应用开发人员和相关从业者都有很好的参考价值。决策者可以从中看到技术趋势，把握时代发展脉搏；数据分析人员可以看到经验的总结和工具的应用；其他从业者可以从中了解数据技术所涉及的各个方面。

图书在版编目（CIP）数据

寻路大数据：海量数据与大规模分析 /（美）马诺切里（Manoochehri,M.）著；戴志伟等译. —北京：电子工业出版社，2014.11

书名原文：Data just right:introduction to large-scale data & analytics

ISBN 978-7-121-24472-8

Ⅰ. ①寻… Ⅱ. ①马… ②戴… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2014）第 231194 号

策划编辑：张春雨

责任编辑：李云静

印　　刷：北京丰源印刷厂

装　　订：三河市鹏成印业有限公司

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：15.25　字数：264 千字

版　　次：2014 年 11 月第 1 版

印　　次：2014 年 11 月第 1 次印刷

定　　价：59.00 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。

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。

服务热线：（010）88258888。

推荐序1

《寻路大数据：海量数据与大规模分析》是一本系统化解读大数据工程处理基础技术的书籍。技术热潮已然催生了形形色色的大数据处理技术及工具，但重要的并非鱼与熊掌的孰优孰劣，而是如何选择或组合这些技术并应用于实现不同的商业目标。

本书正是从这个角度评估了针对不同使用场景的大数据处理技术，从而帮助技术负责人及架构师做出不同的选型决策。我很高兴看到了这本书的出版，它必将有益于大数据技术的各种具体实践。

刘子正
微博常务副总经理

推荐序2

大数据这个概念，提出来已经有好几年了。在这几年中，作为一名数据工作者，我经常会参加各种大数据相关的讨论，也会有朋友和企业提出一些大数据相关的咨询需求。我感觉到，经过一波又一波的质疑和辩论，很多企业已经接受了大数据这个概念，认可大数据对于企业的重要性。眼前的问题，已经不是要不要开展大数据相关的工作，而是如何开展大数据工作。

企业需要将大数据的概念、技术、能力和企业自身的数据充分结合，来切实地提升企业的业务能力。实现这个目标要解决的一个问题是，大数据技术的发展太快、太新，能够对大数据整体技术有了解的人很少。各类技术适合处理什么样的数据？适合解决什么样的问题？如何进行技术选型？这些问题对于实操规划企业整体大数据发展的人非常重要，他们迫切需要介于宏观的概念和细节的技术之间，对于规划和选型立刻就能够产生帮助的资源。

几周之前，我有幸提前阅读到本书的部分章节，很高兴地发现，这本书所提供的，正好是这种立刻能够产生帮助的信息，可以更好地帮助大家规划和设计与企业业务密切结合的大数据应用，而作者丰富的经验和对于大数据的深刻理解，也可以提高我们对于大数据的认知，避免在大数据的应用中走弯路，非常值得仔细阅读。

廖若雪

高德公司 大数据与移动技术中心副总裁

前百度主任架构师，百度公司技术委员会主席

推荐序3

中国改革开放的总设计师曾经说：“不管黑猫白猫，能抓住老鼠就是好猫”，针对大数据技术我想也是如此。我们曾经花费了很多时间讨论什么是大数据，多大才是大数据，而忽略了如何利用大数据技术来创造真正的价值。

随着云计算、智能设备、物联网的快速发展，以后每一个公司都会有大量的数据。而现在最重要的，我认为是快速学习大数据的思维、技术和方法解决工作中的实际问题，并对公司的决策提供支撑，对业务的发展提出建议。这才是大数据的真正价值。

正如本书中提到的“Data processing as a service”，从国内外的发展趋势来看，大数据和云计算的结合越来越紧密，各大云计算运营商都陆续推出大数据处理平台的产品，让每个企业的数据人员能够方便地应用大数据技术，从而专注于业务流程和数据本身，不被大数据基础架构的建设和维护成本所约束，从而快速地发挥大数据的价值。

国庆节前杨毅邀请我为本书作序，杨毅是具有丰富的实战经验和对于大数据有深刻理解的业内人士，本书内容也一样稳重而务实。不纠结于大数据名词的定义，而关注于：如何利用各种技术进行大数据处理，如何在各种应用场景下利用大数据产生价值，如何根据企业自身的实际情况选择合适的架构方案和技术解决实际问题，最终提升企业竞争力。这对于企业的技术负责人、大数据技术人员是非常有价值的。我认为每一个相信、追求并使用大数据的朋友都会和我一样，通过本书加深对大数据的理解，提高对大数据的把握能力，从而为业务创造更大的价值。

大数据，大价值！

季昕华

UCloud 创始人 & CEO

前盛大云 CEO，盛大在线首席安全官

译者序

大数据概念方兴未艾，大数据技术正在蓬勃发展，市场上商用的、开源的数据处理系统和工具不断涌现，你方唱罢我登场，好不热闹。作为从业者的我们，是否也开始有种“乱花渐欲迷人眼”的感觉呢？

本书就是为了应对这种情况而写就的。虽然只有二百多页，但本书的覆盖面相当之广，从数据采集、存储、传输、处理、分析，一直到最后的展示，整个链条上的流行技术工具都有所涉及。而且难能可贵的是，本书不仅指出了常用的解决方案，更给出了各种方案的对比，让读者可以少走弯路，避开其中的陷阱，这无疑为读者节约了大量时间成本。

除了技术方案之外，本书还对数据决策的非技术因素进行了深入的探讨，并对未来趋势进行了展望。对于决策者而言，本书同样具有很好的参考价值。

严复在《天演论》中讲道：“译事三难：信、达、雅。”这句话不仅道出了翻译的难点，也指出了翻译的目标。在本书的翻译过程中，译者力求做到信达二字，在此基础上尽力求“雅”。经过数月的翻译和数次审校，终于完成了本书的翻译工作，在此呈献给读者。由于译者水平所限，译文中可能仍有疏漏，真诚地希望广大读者不吝赐教。

衷心感谢电子工业出版社的张春雨老师、李云静老师和其他编辑人员，由于他们的努力，这本好书才可能与国内读者见面。

非常感谢我的家人，谢谢你们一如既往地支持我的工作。爱你们。

——戴志伟

大数据是毋庸置疑的技术趋势和话题焦点，但是多数情况下，可能人们都太纠结大数据“是什么？不是什么？”，这让话题看起来永远都不会有结论。而本书作者另辟蹊径，在另外的角度上对这一趋势进行解读，即什么才是大数据的正确之路。

现在我们就好像身处大数据技术演进的寒武纪时代，在这个技术变革时期，各种技术粉墨登场，快速演化。而这恰恰是人们对现有大数据技术一切疑惑和模棱两可观点的源头，也正是作者在开篇第 1 章和结尾表述的观点。因为我们身处变革当中，所以很好理解为什么作者并不急于回答“大数据是什么”。只要翻翻这本书，你一定会发现它和市面上其他大数据书籍的不同。

在书中，作者观点如炬，常常金句不断。既有具体数据处理工具的内容（比如用 ff 扩展包来解决 R 语言的单机数据集容量，采用 Python 的 Pandas 包来进行时间序列数据的上 / 下行重采样），也有侧重在战略决策和产业生态的内容。比如选择自研还是外购的关键其实在于能否给企业带来比较竞争优势，又如在“数据科学家兴衰”这一流行问题上的深入探讨，还有对人们乐于处理海量全样本数据这一癖好的挑战，书中让人茅塞顿开的观点比比皆是。作者还强调了当下普遍被忽视的一点，即那些经典的统计学方法（如统计抽样、假设检验、显著性分析等）依然是指导人们在“数据汪洋”中正确航行的“明灯”。

作者在最后的第 13 章和第 14 章提炼了本书的大部分观点并升华为决策指南。相信不同阅读目的的读者，无论是一线数据处理工程师，还是数据流水线架构师，或者是大数据平台和技术选型的决策人，都能从本书获益匪浅。

最后，我想感谢我的家人们。她们是刚刚一岁半的乖女儿啵妞和贤惠的老婆，还有无私奉献的老妈。谢谢你们支持我的工作。

——许杨毅

伴随着网络的高速发展，人们利用计算机存储与处理的数据正呈现爆发性增长的态势。继搜索引擎与社交网络之后，大数据相关技术日益成为行业关注的又一个热点技术话题。也正是由于这些应用需求的推动，早期的数据库等相关领域的技术逐渐演化发展为时下大数据技术的多项内容。

本书作者从事大数据相关领域多年，写作此书的风格与市面上已有的许多“大数据”书籍不同：作者立足于全局的高度来论述大数据技术的方方面面，其综合考虑应用场景、技术细节与选型的思路体现在各章节内容之中。本书概括综合的特点突出，并不是针对特定数据处理技术深入研究的书籍。在内容上，本书兼顾了大数据技术相关技术的各个方面，以数据存储、处理、展示的流程为线索展开具体各部分的叙述。对技术背景较少的读者来说，本书在内容安排上能使他们很快了解各部分技术的相关要点，主旨突出、脉络分明。作者论述精当，要言不烦，而各个层次的内容都有所覆盖，是大数据相关技术入门与选型方面不可多得的指南型佳作。

首先在此向冠诚、志伟、杨毅三位合译本书的成员表示感谢，能有机会与优秀的同行共事是我的荣幸。在合作翻译本书的过程中，得到各位的鼎力支持与帮助，与你们的讨论交流使我获益良多。还要谢谢长久以来支持我学习工作的父母与亲友，人生路上的点滴收获都献给为我默默付出的你们。

——鄢博

年初时，张春雨老师找到我，问我是否愿意接手翻译一本大数据的新书，我当时着实犹豫了一阵子。因为做过一些技术书籍的翻译工作，我深知要把一本技术书籍翻译好不是一件容易的事情。但在粗略读完本书的英文版之后，我的第一感觉就是：这本书跟别的大数据的书不一样，作者在大数据领域经验丰富、“干货”很多，把这本书翻译成中文版一定能让更多的朋友从中受益。

本书涉及的大数据技术覆盖的范围非常广泛，从 Hadoop 到 Spark，从 R 语言到 Google 的 BigQuery，从大规模机器学习到 NoSQL 等都有相关的介绍。不仅如此，作者还针对“如何基于商业需求进行大数据技术选型”做了非常有针对性的介绍。虽然现在市场上已经有了非常多介绍大数据的书籍，但却缺一本对大数据的相关技术有一个全面介绍的“技术选型指南”。如果你希望有这样一本书，即它能帮你快速了解大数据技术的一个全貌，并对大数据技术的选型有一个较深入的理解，那么本书绝对是不二之选。

经过几个月的翻译与校对工作，本书的中文版终于要正式与读者朋友们见面了。回想与几位译者以及编辑老师们一起走过的路，真的是感慨良多。感谢张春雨老师、李云静老师的辛苦工作，没有你们的大力支持，这本书不可能这么快与广大读者朋友见面。非常幸运能找到志伟、杨毅和鄢博来共同完成这本书的翻译，与你们的合作使我受益良多，并让我真切体验了一次“中国合伙人”的感觉。感谢我的家人，特别是我的老婆和我们可爱的儿子，谢谢你们对我工作一如既往的支持，我爱你们。

因译者水平有限，翻译过程中难免存在一些疏漏，恳请广大读者朋友批评指正。如果大家有和本书相关的内容需要探讨，或有相关的宝贵意见，欢迎通过 chenguancheng@gmail.com 和我们联系。真诚希望大家能从这本书中有所收获！

——陈冠诚

谨以此书献给我的父母：

Andrew Manoochehri 和 Cecelia Manoochehri，

他们为我能够接受到优质的教育付出了一切。

序

对数据进行采集、存储和分析的工具种类非常繁多，而且新的工具还在不断涌现。对于刚进入这个领域的新人来说，这往往意味着需要浏览众多网站和相关书籍才能对大数据处理的基础知识有个基本的了解。正因如此，这本书成为 Addison-Wesley 数据分析（*Data & Analytics*）丛书的一个有力补充：本书对构建大数据分析系统的工具、技术和实用技巧进行了全面的介绍。

Michael 是介绍大数据分析的绝佳人选，他曾在 Google 的云平台开发者关系组工作，帮助开发者使用 BigQuery（Google 的 TB 级数据分析平台）进行大规模数据分析。他将自己在大数据领域广阔的知识面带到了这本书中，为刚接触大数据的人和寻求建议、技巧和工具的人提供了非常实用的实战指南。

本书从大数据系统的成功应用开始介绍，之后陆续对 NoSQL、分布式计算和 CAP 理论进行了讲解。在介绍使用 Hadoop 和 Hive 分析大数据之后，又覆盖了使用 BigQuery 进行实时分析的相关内容。之后还包括了 MapReduce 流水线、Pig 和 Cascading、使用 Mahout 进行机器学习等高级课题。在书的最后，读者会看到将 Python 和 R 整合到大数据工具链中的实际案例。本书大部分章节都包含了很多例子以帮助读者学习和使用相关的大数据工具。如果你想要一本对大数据分析有一个全面了解的书籍，本书绝对是不二之选。

——Paul Dix，丛书编辑

前言

注意到了吗？移动技术和社会化媒体产生的数据已经超过了人类能够理解的范围，大规模数据分析突然变得魅力四射。

分布式和云计算领域正在快速发展以分析和处理这些数据。技术变革那令人难以置信的速度已经彻底颠覆了人们应对数据挑战的旧有观念，强迫他们跟上时代的步伐去评估一系列技术，而这些技术有时甚至是互相有冲突的。

很久以来，关系型数据库一直是商业智能应用的推进器，如今一些激进的开源NoSQL 新贵也加入了进来。二者的结合构成了一种全新的混合数据库解决方案。基于 Web 的计算所存在的优点驱使着大规模数据存储从定制数据中心转向可伸缩的“基础设施即服务”上来。另一方面，基于开源的 Hadoop 生态系统的项目使得普通开发者也能够接触到数据处理技术，这在以前只有一些做云计算的大公司，如 Amazon 和 Google 才能做到。

这些技术创新的结果通常被称为*大数据*（Big Data）。关于这个词汇的含义有很多争论。大数据是一个新产生的趋势，抑或只是老调重弹？大数据是如其字面意思那样意味着很多的数据，还是指使用新的方式去挖掘数据价值的过程呢？科学历史学家 George Dyson 总结得很好：“当扔掉数据的代价大于所需机器代价时，大数据才有了存在的价值。”换句话说，当数据本身的价值超过了收集和处理这些数据所需的计算能力时，就有了大数据。

尽管一些支持大数据运动的公司和开源项目的令人惊奇的成功的确是事实，但同时很多人也已经发现，去了解大量新的数据解决方案和服务提供商非常具有挑战性。而我发现设计解决方案去面对数据挑战的过程往往可以归纳为一系列共同的用例，这些用例在这些解决方案中一再出现。

寻找高效的数据解决方案就意味着权衡。一些技术是为某类特殊的数据用例专门优化的，因此对于其他类型的数据来说并不是最好的选择。一些数据库软件为了达到更高的分析速度而牺牲了灵活性，而另一些数据库软件可能为了更高的性能会牺牲一致性。本书会通过介绍实际用例和真实的成功案例帮助你学习如何去做选择。

本书的适用范围

在这个世界上没有使用无限的金钱和资源解决不了的问题。不管怎样，拥有大量资源的组织总是可以建造他们自己的系统去收集和分析任何规模的数据。本书并不是写给这些拥有无限的时间、一大群努力的工程师和无穷预算的人们的。

本书写给除此以外的其他人。这些人在寻找数据解决方案，但同时拥有的资源有限。大数据时代的一个主题是任何人都能够获取到合适的工具，而这些工具在几年前还只有少数几个大公司拥有。然而另一个摆在面前的事实是，很多工具非常新颖，并在快速演变，并不总是能够无缝地互相衔接。本书的目标就是向读者演示如何高效地将这些部件组装在一起建造成一个个完整的系统。我们会讨论解决数据问题的策略，如何使它们变得经济、可行，当然还必须具有实用性。

开源软件已经在无数的方面降低了人们获取技术的难度，在大数据领域也是如此。但是，本书中涉及的技术和解决方案并不全部是开源的，有时候涉及商业公司提供的计算资源服务。

尽管如此，很多基于云的服务是使用开源工具建造的，事实上，若没有这些开源工具，很多云服务根本就不会存在。因为规模经济效应，公用计算平台越来越多，用户可以按需付费购买超级计算资源，就像人们购买自来水和电一样。

我们会讨论在保证系统可伸缩性的同时尽量降低开销的策略。

为什么现在写这本书

有一件事仍然让我感到非常神奇，那就是如果不考虑经济发展程度不一和语言障碍，写一个能够被整个星球的人使用的软件并非不可能。像 Facebook、Google 搜

索、Yahoo！邮箱和中国的 QQ 空间这样的 Web 应用拥有几亿甚至几十亿的活跃用户并非不可能。Web 和相关开发工具的规模仅仅是大数据领域发展速度如此令人瞩目的原因之一。让我们来看看对此也有贡献的其他趋势吧。

开源大数据的成熟

2004 年，Google 发布了一篇著名的论文，文中详细介绍了一个叫作 MapReduce 的分布式计算框架。MapReduce 框架是 Google 用以将海量数据处理问题分割成多个更小问题的关键技术。不久之后，Google 发布了另一篇论文，介绍了 Google 内部使用的分布式数据库技术：BigTable。

从此，很多开源技术出现了，它们要么是这些 Google 论文中技术的实现，要么受到了其启发。同时，由于关系型数据库在分布式系统中使用所暴露出来的先天不足，新的数据库范式越来越为人们所接受。某些范式彻底避开了关系型数据库的核心属性，抛弃了标准化的模式、确保的一致性，甚至 SQL 本身。

Web 应用的崛起

随着喜欢使用 Web 的人越来越多，数据产生的速度也越来越快。Web 用户的增加也带来了 Web 应用的增加。

基于 Web 的软件通常基于应用程序接口（API）构建。应用程序接口能够将网络中独立的服务连接起来。例如，很多应用允许用户使用其 Twitter 账户信息来做认证，或者通过 Google 地图来可视化地分享自己的地理位置。每一个接口都有可能会提供某种类型的日志信息用于做数据驱动的决策。

另一个对现在的“数据洪流”有所贡献的是持续增长的用户产生内容和社会化网络的普及。互联网使得人们能够以极小的代价发布内容。尽管会有大量的噪声数据，但是从营销和广告方面看，懂得如何收集和分析雪崩式的社会化网络数据仍然是非常有用的。

根据从这些 Web 服务中收集的信息来辅助进行商业决策是可行的。例如，想象一下如何根据地理信息洞察销售规律：是否购买了某种产品的独立用户有 30% 来自于法国并且在 Facebook 上分享了他们的购物信息？像这样的数据很可能会帮助你决

定动用资源瞄准社会化网络上的法国客户。

移动设备

可伸缩的数据技术比过去任何时候都热门的另一个原因是，全球移动通信设备数量的爆炸。与其说这种趋势是由于个人对功能手机和智能手机的使用造成的，不如把这种趋势看作是以用户身份为中心而与设备无关的。假如你同时在使用一台计算机和一部智能手机，那么或许你就能够通过其中任何一个设备来存取自己的个人数据。而这份数据或许是存储在某个基础设施即服务的提供商的数据中心中。类似地，我的智能电视能够在空闲的时候把我关注的 Twitter 用户的推文作为屏幕保护显示出来。这些都是普适计算——基于你的身份从联网的任何设备存取资源的例子。

随着移动设备使用的加速增长，消费性移动设备越来越多地被用于商业目的。我们正处于普适计算的早期阶段，人们使用的设备仅仅是作为通过网络存取个人数据的工具。业界和政府正开始认识到使用 100% 基于云的商业生产力软件的关键优点，即能够改进雇员的机动性和提高工作效率。

总之，每天有数百万的用户开始通过持续增长的设备使用基于网络的应用程序。只要能够收集、处理和分析这些数据，就能够从中发现可用于进行商业决策的巨大价值。

物联网

未来，任何使用电的设备都可能会连接到互联网，因此会有很多数据在用户、设备和服务器之间来回传输。这通常被叫作物联网（Internet of Things）。如果你认为现在的数十亿互联网用户产生的数据很多的话，想想等到我们所有的汽车、手表、灯泡和面包机都上网了之后会怎么样吧。

尽管还不清楚能连接 Wi-Fi 的面包机有没有市场，但是越来越多的商业公司和个人爱好者开始使用廉价硬件进行物联网方面的探索。我们可以想象一下这种联网设备：用户能够完全通过智能机或平板电脑来操控。这种技术已经出现在了电视机上，应该很快就能取代微波炉上那些令人无法容忍的控制面板了。

就像前面描述的移动应用和网络应用趋势，物联网对个人隐私和政策的影响需要进行详细的评估：谁可以获知你在哪里如何使用你新买的支持 Wi-Fi 的电动牙刷？另一方面，从这些设备中收集的信息也可以使市场更加高效、自动检测设备中的潜在故障，或者向用户发出警告信息，这些警告信息可能会节省他们的时间和金钱。

通向普适计算之路

前面提到的信息太多，可能反而转移了大家的注意力，但是有一点是非常重要的：随着互联网背后的分布式计算技术使得人与人之间的交流越来越容易，大数据技术趋势也让寻找问题答案的过程从过去的不可能变得可能。

更重要的是，用户体验的进步意味着我们正在进入这样一个世界——探究我们产生的令人无法想象的海量数据的技术正在变得越来越透明、经济和唾手可得。

本书组织结构

处理海量的数据需要使用一系列专业技术，而每种技术本身都有取舍和挑战。本书分成几个部分，分别描述在一些常见用例下的数据挑战和成功的解决方案。第 1 部分“大数据时代指引”包含了第 1 章：“数据成功四原则”，本章描述了为什么大数据如此重要，以及为什么新技术的前景不仅意味着机遇也意味着挑战。本章介绍了贯穿全书的共同主题，如构建可伸缩的应用；构建解决数据孤岛问题的协同工具；在考虑使用什么技术之前先考虑应用场景；除非绝对必要，否则避免建造基础设施。

第 2 部分“收集和共享海量数据”描述了关于收集和共享大规模数据的应用场景。第 2 章“托管和共享 TB 级原始数据”描述了如何应对托管和共享大量文件中存在的看起来很简单的挑战。选择合适的数据格式是非常重要的，本章覆盖了共享数据时必须要考虑的问题，以及经济地托管大量数据时所需的基础设施。本章通过讨论传输数据时使用的序列化格式给出了相应的结论。

第 3 章“构建基于 NoSQL 的 Web 应用采集众包数据”介绍了可伸缩数据库技术领域。本章讨论了关系型数据库和非关系型数据库的历史，以及如何在二者之间