大数据科学丛书

数据质量征途

Journey to Data Quality

黄伟　王嘉寅　苏秦　冯耕中　编译

[美] Yang W. Lee
Leo L .Pipino
James D.Funk
Richard Y.Wang
著

高等教育出版社

大数据科学丛书

Journey to Data Quality

数据质量征途

Shuju Zhiliang Zhengtu

[美]Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang 著

黄伟　王嘉寅　苏秦　冯耕中　编译

高等教育出版社·北京

图书在版编目(CIP)数据

数据质量征途 / (美)李(Lee, Y. W.)等著;黄伟等编译. -- 北京 :高等教育出版社, 2015.7
(大数据科学丛书)
书名原文:Journey to Data Quality
ISBN 978-7-04-042675-5

Ⅰ. ①数… Ⅱ. ①李… ②黄… Ⅲ. ①数据处理
Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 107660 号

策划编辑 冯 英　　责任编辑 冯 英　　封面设计 王 鹏　　版式设计 童 丹
责任校对 胡美萍　　责任印制 尤 静

出版发行 高等教育出版社
社　　址 北京市西城区德外大街 4 号
邮政编码 100120
印　　刷 北京宏信印刷厂
开　　本 787mm ×1092mm 1/16
印　　张 13
字　　数 230 千字
购书热线 010-58581118

咨询电话 400-810-0598
网　　址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.landraco.com
http://www.landraco.com.cn
版　　次 2015年7月第1版
印　　次 2015年7月第1次印刷
定　　价 39.00 元

物 料 号 42675-00

序一

大数据是数字时代的新型战略资源，也是服务创新、驱动发展的重要抓手。大数据是数据科学的一个应用，也是数据科学重要的发展方向。近年来，大数据的热潮与数据科学的发展互为促进，正改变着人们的生产、生活方式。而对于学者和业界来说，是否能够抓住机遇、深入研究、形成解决方案，就显得非常必要和紧迫。

由于大数据具有分散存储、整合使用，分析处理的时间、空间复杂度高，以及数据整体及其关系协同呈现高价值的三大特征，数据质量往往难以保障。但是数据质量对于使用、用好大数据起到决定性的作用。数据质量低不仅会降低决策质量，更可能带来难以估量的灾难性损失。保障和提高大数据的质量迫在眉睫。

《Journey to Data Quality》一书堪称数据质量领域的经典之作。该书从数据质量的概念入手，结合案例和分析工具，深入浅出地总结了美国学术界和产业界十余年的成果和经验，具有很强的指导性和实用性。由黄伟教授统筹，联合了国内外几位学者翻译该书，并融入最近几年的研究成果，很有意义。对于国内致力于数据质量的学者和业界来说，本书可以提供基础性的介绍和指导，为解决大数据环境下的数据质量问题指出方向。

徐宗本教授

中国科学院院士

2015 年 5 月

序二

大数据时代的到来，不仅不断改变着人们认知世界的方式，而且正以其独特的影响力，推动着各行各业发展进步的新潮流。伴随着各种大数据技术的不断涌现，无论是在生命科学、医学，还是在金融、管理等诸多领域，大数据都显现出越来越重要的作用。大数据产业和管理正是将信息技术广泛而深入地应用于不同的行业，使后者拥有快速获取、高效存储、精准分析、正确判断各类数据和信息的能力，从而实现组织的科学决策。

积极抓住大数据带来的技术创新和产业变革的契机，成为国家间竞争和取得领先优势的关键。美国的国家大数据计划法案提出加强大数据相关领域的研究，将大数据及其产业作为国家推动的一个战略方向。党的十八大明确提出了创新驱动发展战略，大数据产业的发展以其内生的创新优势也必将成为重中之重。如何利用大数据提高国家的竞争力，如何发挥大数据的作用，成为政府、各行业都十分关注的问题——这其中，数据质量是一切的保障。

由黄伟教授牵头，联合西安交通大学和美国圣路易斯华盛顿大学的几位学者共同翻译、修编了《Journey to Data Quality》一书。该书不仅在数据质量研究领域享有盛誉，而且得到产业界的高度赞许和推崇，这是非常难能可贵的。书中不仅总结了前沿的研究成果和产业界多年的实践经验，而且通过对经典案例的深入剖析实现了理论与实践应用的统一。本书对于研究者、实践者，特别是管理者来说，将会带来重要的启迪。

汪应洛教授

中国工程院院士

西安交通大学管理学院名誉院长

2015 年 5 月

译者前言

信息技术的快速发展和普及推动着“大数据时代”骤然而至。短短的几年中，以复杂网络系列理论为基础的社会－社交网络大数据支撑着众多的社交平台，近乎颠覆性地改变着人们日常的社交活动；以自组织体系结构为基础的智能电网技术，以及陆续涌现的智慧城市、智慧田园、智慧海洋大数据带动着诸多关键技术的跨领域交叉融合和跃进，不知不觉中改善着人们的社会生活；以高通量测序技术为基础的生物大数据不仅在农作物精准育种方面大显身手，而且与医学大数据结合形成的基因尺度特征－复杂疾病－药物功效关联网络使得个体化医疗变为现实。此外，在金融工程、先进制造、纳米材料等前沿领域和产业，大数据也正发挥着巨大的、乃至决定性的作用。

新的科技革命和新的产业变革已经初现端倪，抢抓机遇成为大国的必然选择。2012 年美国政府公布了“大数据研发计划”（Big Data Research and Development Initiative），以期加强大数据相关领域的研究并带动产业发展，这是世界上首个大数据发展的国家战略。我国做出实施创新驱动发展战略的重大部署以来，大数据产业和大数据研究也迅速起步。可以预见在未来的十年，大数据产业将成为助推产业经济转型的强劲引擎。为满足我国大数据产业发展的需求，培养市场、社会急需的人才，推动大数据相关研究，西安交通大学管理学院与美国麻省理工学院（Massachusetts Institute of Technology）合作组建了数据科学与数据质量研究中心。作为一项基础性的工作，中心将引进、编著一系列大数据理论、技术、产业和管理相关的教材，希望借此帮助国内的学者特别是产业界，了解世界范围内大数据的前沿研究和应用，同时作为教学之用。

基础不牢，地动山摇。大数据产业的根基是数据，倘若数据的质量出了问题，大数据产业就难以获得高质量的产品，也很难得到长足的发展。大数据领域有句俗话“进去的是垃圾，出去的就是垃圾”（garbage in，garbage out），说的就是这种现象。不少研究指出，产业发展越迅速，低质量数据的危害就越严重。正因为如此，作为“大数据科学”丛书的首册，我们编译引进了 Yang W Lee、Leo L Pipino、James D Funk 和 Richard Y Wang 合作编著的《Journey to Data Quality》一书。四位作者都是国际数据质量研究的先驱，作者在书中总结了十余年的研究成果和实践经验，包括对数据质量概念的翔实介绍，对数据质量项目案例的深入

剖析，以及信息产品地图等数据质量分析工具的应用指南。无论是对初窥门径的学生、还是对拥有多年工作经验的实践者，本书都有极高的参考价值。

本书的编译工作由西安交通大学黄伟教授、苏秦教授和冯耕中教授统筹。原书共12章，翻译初稿的第1～4章由张坦完成，第5、6章分别由张宏云、张舟完成，第7、8章由陈静完成，第9、10章由李雅慧完成，第11、12章分别由徐丰、杨彤完成。翻译初稿完成后，在一定范围内征求了意见。根据反馈，由刘跃文、张宏云、吴悦和马续补对翻译初稿的部分内容作了修改。全书由圣路易斯华盛顿大学王嘉寅博士修编并统稿，并在原著的基础上编入了两个附录，分别介绍了信息产品地图的配套绘制软件——迅图，由韩博编写，以及近年来数据质量研究的新进展，由刘跃文、张宏云、徐丰、黄伟编写。

在本书的编译过程中，时任西安交通大学副校长蒋庄德院士、徐宗本院士和宋晓平教授给予我们很多鼓励和帮助，谨在此深表感谢。西安交通大学管理学院名誉院长汪应洛院士、中国科学院大学石勇教授、美国麻省理工学院 Harry Zhu 博士、Xitong Li 博士、西安交通大学郭菊娥教授在书稿翻译、修订和编写中给予我们很多指导、建议和意见，特在此表示感谢。西安交通大学人文社会科学处贾毅华处长等在立项和配套等很多方面给予我们大量支持，在此致谢。我们也非常感恩自己的家人，感谢徐忠锋教授的指导和帮助，以及大家对我们这项工作的理解和支持。

正如本书的书名，编译工作也是自我学习的旅途。无论是大数据还是数据质量都是全新的领域，书中难免有各种各样的不足，我们诚恳地希望读者向我们反馈，相互学习提高。

黄伟、苏秦、冯耕中于西安交通大学
王嘉寅于圣路易斯华盛顿大学
2015年6月

前言

本书汇集了作者和本领域众多学者的研究成果，也包括他们在政府和相关行业实践工作中累积的经验。本书尝试将分别来自于学术期刊和学术实践性会议中的诸多观点、概念予以总结和升华，向读者展现这些观点和概念是如何被许多组织采纳并用作数据质量管理和实践的原则、政策和技术工具的。进一步地，作者将通过具体的现实例子和来自行业的案例对本书中的理论观点和方法加以讨论。

本书的读者群主要是企业的管理层、从事数据质量方面的工作人员、数据质量领域的研究者和学生。对于业界人员，本书有助于深入理解他们所从事工作的理论基础，为将来更好地解决问题并付诸实践做好准备。研究人员能够通过本书了解数据质量理论是怎样被应用到实践中的，进而有助于更加专注于未来的研究领域。而对于学生来说，本书能够提供对于这个领域的宏观认知，为今后在这一领域的学习和研究奠定坚实基础。管理层人员则可能会对本书的前几章和第 11 章（数据质量政策）更感兴趣并得以裨益。

致谢

很多人都为本书的出版做出了贡献。我们感谢匿名评审，是你们看到本书第一稿中的知识和实际潜力，并用心帮助我们形成连贯的书。我们感谢同事们审阅本书，并提出宝贵的意见。在之前的研究和咨询工作中，许多研究和从业人员与我们一起合作，从而奠定了这本书的基础。他们是 Stuart Madnick、Don Ballou、Harry Pazer、Giri Tayi、Diane Strong、Beverly Kahn、Elizabeth Pierce、Stanley Dobbs 和 Shobha Chengular-Smith。Raïssa Katz-Haas 和 Bruce Davidson 为本书中的数据质量实践案例提供了很多素材。

此外，我们感谢产业界的实践者，开放他们的实践环节和机构，并允许我们将此作为我们的实验场。

本书中的一些成果借鉴了其他作者之前的出版物。感谢下列期刊允许我们使用相关内容：Communications of the ACM、Sloan Management Review、Prentice Hall、Journal of Management Information Systems、Journal of Database Management、Information and Management、International Journal of Healthcare Technology and Management 和 IEEE Computer。

麻省理工学院（Massachusetts Institute of Technology，MIT）出版社的 Doug Sery 在整个出版过程中不遗余力地为本书提供独特的见解和毫无保留的支持。

如果没有卓有成效的工作环境，这本书也不可能完成，感谢麻省理工学院信息质量项目、剑桥（Cambridge）研究小组和麻省理工学院全面数据质量管理（TDQM）项目，以及睿智和富于奉献精神的同事们：Tony Nguyen、Karen Tran、Rith Peou、Andrey Sutanto、John Maglio 和 Jeff Richardson，感谢东北大学（Northeastern University）工商管理学院的大力支持，感谢马萨诸塞大学洛厄尔分校（University of Massachusetts Lowell）管理学院的支持。

最后，我们感谢自己的家人，谢谢他们在编写这本书的漫长过程中付出的爱、支持和理解。Albina Bertolotti、Marcella Croci、Karen Funk、Karalyn Smith、Marc Smith、Jacob Smith、Cameron Smith、Kirsten Ertl、Phil Ertl、Austin Ertl、Logan Ertl、Lorenzina Gustafson、Laura Gustafson、Juliana Gustafson 和 Fori Wang，给我们的生活带来这么多快乐和幸福。特别地，我们要感谢我们的父母，谢谢他们培养我们对学习的热爱。

目录

第1章 引言

Jane Fine是一家全球制造公司的信息系统主管。她正坐在自己的位子上思考着下一步应该怎么办:对于销售副总裁提出的"对某个大客户的销售总量是多少?"的问题,她很难作出回答。因为她知道公司的数据库中存在重复的客户标识和重复的产品代码,这样不恰当的数据库结构及其带来的不准确内容,令她难以回答销售总量的问题。在一千英里①之外的一家大型教学医院,负责医疗事务的常务副总裁Jim Brace也正面临着不同的窘境。他刚刚参加了一个由医院负责人召集的会议,州政府的监管机构对医院的数据报告中部分数据的真实性提出了质疑,这导致委员会拒绝批准医院的报告,而医院将因此失去州政府的财政补贴,这一情况必须尽快得到处理。与此同时,政府人力资源部门的高级信息专员Dean Gary在审核人事文件时,发现一些文件中的数据与该部门最新的人事报告中的一些汇总数字存在出入,数据存在不一致的现象。尽管这个问题目前还没有被人事工作报告的使用者发现,也暂时没有从其他方面体现出来,但是不可能永远不被发现。

在以上案例中,前两个是信息系统的上层管理者发现的数据质量问题。如果问题继续存在,组织的高层管理者很可能基于低质量的数据作出糟糕的决策。例如,在大型教学医院的案例中,医院负责人对数据质量问题的关注是由外部监管机构的质疑引起的。

在大多数组织中,如何让管理层相信数据质量存在问题,并制定一个正式的数据质量计划,是一项挑战。比如,在全球制造公司中,高层通常要在一个可接受的时间内获取他们要求的数据。对他们而言,这些数据不存在问题。他们也看不到下面的人为了满足他们的数据需求,清理不一致数据所做的额外工作。尽管在修正数据质量问题时花费的时间并不多,但是这些时间本可以用于更加高效地完成其他任务。

在以上三个案例中,管理者们不得不面对他们组织中的数据质量问题。可以猜测大多数组织应该也面临着相似的问题。

① 译者注:1 mile = 1 609.344 m。

1.1 信息可以被共享吗

一些组织在执行跨业务的流程，或者尝试跨系统、跨组织交互时，常常难以充分地利用信息。当这些组织相信他们拥有完成业务功能的数据但却不能顺利开展业务时，组织内部就容易产生挫败情绪。例如，某公司希望做一些趋势分析，以便与客户和其他合作者构建更紧密的关系，但是该公司的信息技术部门却经常不能提供客户所要求的整合性信息，或者无法按客户要求的时间提供其所需的信息，这导致公司错过了利用这些收集和储存的信息的最佳时机。更糟糕的是，竞争对手却能迅速反应，战略性地应用类似的信息。

很多组织长期以来都面临这些问题。数据质量问题还可能表现在其他方面，具体包括：

- 许多跨国公司难以管理其全球的数据，虽然这些数据可以用于解决公司当前以及未来全球性的、区域性的业务问题。
- 外部检查使组织内部的数据质量问题浮出水面，正如前面提到的监管机构对医院的医保报销和患者投诉进行审查的事例。
- 信息系统项目也能够揭示出存在的数据问题，特别是一些涉及跨业务的、多数据来源的数据质量问题。
- 组织成员在工作中发现了数据质量问题，却只使用某些变通方法临时满足数据需求，而不是使用或创建永久性的、持续性的解决方案。

1.2 新系统不是解决办法

每一个组织都希望自己拥有高质量的数据，但是常常不知道如何实现这个目标。一类常见的做法是开发一个新系统来取代旧系统，然而常常会在实施之后立即后悔。这是因为公司实施此类方案时，总是重建一套全新的系统，却很少在第一时间考虑原系统存在困难的真正原因——数据质量问题。比如信息系统部门往往热衷于使用最新的技术，开发更流行或更常见的软、硬件解决方案，我们将这种方法称为系统驱动型解决方案。此时，公司采取的方案的真实目标退化为开发新系统，而非修正数据质量问题以提供高质量的数据。显然，这种舍本逐末的新系统非但不能解决原有问题，而且很有可能加剧数据质量问题。即使某个解决方案偶尔会有成效，通常真正造成问题的原因却更容易被掩盖或进一步隐藏。

许多公司误以为使用了最新的软件，比如企业资源规划（enterprise resource planning，ERP）系统或者紧跟潮流地引入一个数据仓库（data warehouse，DW）就会坐享更高质量的数据。公司希望通过这些系统更好地实现公司范围内的信息共享。然而，信息技术部门在整合多来源数据的过程中越来越清醒地认识到，数据定义、数据格式以及数据的值都可能存在大量的不一致现象，但是时间等各方面的压力迫使他们依然继续使用之前存在的同样糟糕的数据。

许多公司感到失望的是，在数据仓库上付出大量努力却没有得到较好的商业价值。在众多案例中，许多采用了 ERP 系统和数据仓库的公司并没有获得最初承诺的预期商业价值。

依然以前面提到的全球制造公司为例，公司试图整合全球范围的销售信息，尽管公司有全部的原始数据，却依然需要花费几个月的时间才能实际提供某个指定客户商业需求的一套有用的数据。存在的问题包括：同一个客户对应多个标识，以及多个客户被赋予同一个标识。此外，子系统中储存的数据在公司层面没有合理的定义和记录，物理数据库并不是一直可以访问的，公司内部没有对概念和术语的定义实行标准化，与标准不同的内容没有被记录或不能被共享，等等。

在此阶段，公司已经在这个项目上花费了大量的预算，管理层还愿意增加预算吗？或者管理层会终止这个项目吗？如果管理层增加预算，但是依旧没有对基础业务和数据质量问题予以足够的关注，上述状况会有实质性的改变吗？公司是否应该致力于创建另一个业务流程，但是与新业务流程相关的费用会不会导致商业价值没有增加？如果管理层终止项目，那么公司共享信息的努力也将终止。此时无论增加预算还是终止项目，公司整体的数据质量都将不会有任何提高。

大多数组织总是狭隘地关注系统层面的问题，却一再忽视数据层面的问题。解决数据质量问题可以增加组织内部共享信息的能力；反之，如果忽视数据问题，大多数系统层的解决方案最终都将失败。那么，如果公司发现自身可能存在严重的数据问题应该怎么办呢？

一些组织通过使用基础性的数据清理软件来尝试改进数据质量。在全球制造公司的案例中，通过初始的数据清理建立了一个可用的数据仓库。然而，随着时间的推移，数据仓库内的数据质量再次急剧下降。在更普遍的案例中，企业通常会指派个别人员去解决特定的数据质量问题，或者某个、某些对数据质量问题感兴趣的人主动解决了其中的问题。但是，不论这些问题是怎样被发现的，无论某个人如何成为问题的负责人，最初的调查和解决方案通常都是临时方案。

许多企业已经应用了多种多样的临时方法，却依然得不到尽如人意的结果。此时，数据质量项目可能会被迅速终止。这种情况下，重新开始新的数据质量方案将变得非常困难。本书中，我们将提供更系统、更全面的基础性解决思路和方案。

1.3 开启数据质量之旅

让我们重温上述案例的场景，看看他们采取了什么后续行动。全球制造公司的 Jane Fine 做了某些调研，在自己的数据库管理经验和数据质量领域知识的基础上，她开始了解业界和学术界的前沿发展和行业状况并参加了多个研讨会。她广泛地搜索外部的资源，试图获取解决问题所需的知识，通过部署技术和流程来改善公司的这一问题。当然，她仍然面临着很多挑战。

在医院负责人的支持下，Jim Brace 编写了一个独立的内部软件来尝试解决问题。在此基础上，他获得了一些反馈建议并得以实施。但是数据质量问题依然存在，所以 Brace 通过查阅全面数据质量管理方面的文献，采用测量的方法实现了数据质量的改善。该方法取得了一定的成功，得益于外部的技术和知识，Brace 采取更主动、更全面的方法设计出一个可持续并切实可行的数据质量方案。

人力资源部门的 Dean Gary 利用经典数据库理论中数据整合的概念和数据挖掘技术解决了他的问题。他使用一种技术手段力求识别出不同类型的数据错误，这可以帮助他解决当前的问题。在数据分析前清理所接收的数据，使之能够提供有效的报告。然而，他无法识别并消除数据不一致现象发生的根源，所以他在每次收到新的数据时，都不得不重复进行数据清理，这促使他开始寻求其他改进数据质量的方法。

三位管理者都有意或无意地踏上了数据质量之旅。有许多不同的路可供选择。如果选择了某条合适的路径，伴随着旅程，数据质量将会不断提升，即使在旅程中会不断遇到新的数据质量问题。这种发现问题、提高质量、解决问题的过程形成一个周而复始的循环。在经历了几个循环之后，低质量的数据对组织的影响将会快速降低。然而，重复过程仍将会继续。问题的解决方案会产生新的问题，这些问题会激发新的需求，而新的需求又会产生新的问题。

1.4 成功开始的故事

许多组织已经走上了数据质量之旅，但多数仅仅是为了尽早地完结这一旅