HZ BOOKS
华章IT

Big Data Integration

大数据管理丛书

大数据集成

[美] 董 欣（Xin Luna Dong）
戴夫士·斯里瓦斯塔瓦（Divesh Srivastava） 著
王秋月 杜治娟 王硕 译

机械工业出版社
China Machine Press

大/数/据/管/理/丛/书

Big Data Integration

大数据集成

[美] 董 欣（Xin Luna Dong）
戴夫士·斯里瓦斯塔瓦（Divesh Srivastava） 著
王秋月 杜治娟 王硕 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据集成/（美）董欣（Xin Luna Dong），（美）戴夫士·斯里瓦斯塔瓦（Divesh Srivastava）著；王秋月，杜治娟，王硕译．—北京：机械工业出版社，2017.3
（大数据管理丛书）
书名原文：Big Data Integration

ISBN 978-7-111-55986-3

I. 大… II. ① 董… ② 戴… ③ 王… ④ 杜… ⑤ 王… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 029113 号

本书版权登记号：图字：01-2016-5929

本书作者在多年研究传统数据集成的基础上，着重分析了大数据背景下的大数据集成。和传统的数据集成相比，大数据集成具有一些新的挑战，例如数据和数据源的海量性、数据的多样性和数据的动态性等。本书共分 6 章，包括大数据集成的挑战和机遇、模式对齐、记录链接、数据融合、出现的新问题和结论，系统地讨论了解决大数据集成中关键问题的一些重要研究成果和方法，对大数据集成的研究者和实践者都很有帮助。另外本书也可以作为学生学习该领域的入门读物。

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：朱秀英　　责任校对：李秋荣
印　刷：北京诚信伟业印刷有限公司　　版　次：2017 年 5 月第 1 版第 1 次印刷
开　本：170mm×242mm　1/16　　印　张：12.75
书　号：ISBN 978-7-111-55986-3　　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）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

当下大数据技术发展变化日新月异，大数据应用已经遍及工业和社会生活的方方面面，原有的数据管理理论体系与大数据产业应用之间的差距日益加大，而工业界对于大数据人才的需求却急剧增加。大数据专业人才的培养是新一轮科技较量的基础，高等院校承担着大数据人才培养的重任。因此大数据相关课程将逐渐成为国内高校计算机相关专业的重要课程。但纵观大数据人才培养课程体系尚不尽如人意，多是已有课程的“冷拼盘”，顶多是加点“调料”，原材料没有新鲜感。现阶段无论多么新多么好的人才培养计划，都只能在20世纪六七十年代编写的计算机知识体系上施教，无法把当下大数据带给我们的新思维、新知识传导给学生。

为此我们意识到，缺少基础性工作和原始积累，就难以培养符合工业界需要的大数据复合型和交叉型人才。因此急需在思维和理念方面进行转变，为现有的课程和知识体系按大数据应用需求进行延展和补充，加入新的可以因材施教的知识模块。我们肩负着大数据时代知识更新的使命，每一位学者都有责任和义务去为此“增砖添瓦”。

在此背景下，我们策划和组织了这套大数据管理丛书，希望能够培

养数据思维的理念，对原有数据管理知识体系进行完善和补充，面向新的技术热点，提出新的知识体系/知识点，拉近教材体系与大数据应用的距离，为受教者应对现代技术带来的大数据领域的新问题和挑战，扫除障碍。我们相信，假以时日，这些著作汇溪成河，必将对未来大数据人才培养起到“基石”的作用。

丛书定位：面向新形势下的大数据技术发展对人才培养提出的挑战，旨在为学术研究和人才培养提供可供参考的“基石”。虽然是一些不起眼的“砖头瓦块”，但可以为大数据人才培养积累可用的新模块（新素材），弥补原有知识体系与应用问题之前的鸿沟，力图为现有的数据管理知识查漏补缺，聚少成多，最终形成适应大数据技术发展和人才培养的知识体系和教材基础。

丛书特点：丛书借鉴 Morgan & Claypool Publishers 出版的 Synthesis Lectures on Data Management，特色在于选题新颖，短小精湛。选题新颖即面向技术热点，弥补现有知识体系的漏洞和不足（或延伸或补充），内容涵盖大数据管理的理论、方法、技术等诸多方面。短小精湛则不求系统性和完备性，但每本书要自成知识体系，重在阐述基本问题和方法，并辅以例题说明，便于施教。

丛书组织：丛书采用国际学术出版通行的主编负责制，为此特邀中国人民大学孟小峰教授（email：xfmeng@ruc.edu.cn）担任丛书主编，负责丛书的整体规划和选题。责任编辑为机械工业出版社华章分社姚蕾编辑（email：yaolei@hzbook.com）。

当今数据洪流席卷全球，而中国正在努力从数据大国走向数据强国，大数据时代的知识更新和人才培养刻不容缓，虽然我们的力量有限，但聚少成多，积小致巨。因此，我们在设计本套丛书封面的时候，特意选择了清代苏州籍宫廷画家徐扬描绘苏州风物的巨幅长卷画作《姑苏繁华图》（原名《盛世滋生图》）作为底图以表达我们的美好愿景，

每本书选取这幅巨卷的一部分，一步步见证和记录数据管理领域的学者在学术研究和工程应用中的探索和实践，最终形成适应大数据技术发展和人才培养的知识图谱，共同谱写出我们这个大数据时代的盛世华章。

在此期望有志于大数据人才培养并具有丰富理论和实践经验的学者和专业人员能够加入到这套书的编写工作中来，共同为中国大数据研究和人才培养贡献自己的智慧和力量，共筑属于我们自己的“时代记忆”。欢迎读者对我们的出版工作提出宝贵意见和建议。

大数据管理丛书

主编：孟小峰

大数据管理概论

孟小峰　编著

2017 年 5 月

异构信息网络挖掘：原理和方法

[美] 孙艺洲（Yizhou Sun）　韩家炜（Jiawei Han）　著

段磊　朱敏　唐常杰　译

2017 年 5 月

大规模元搜索引擎技术

[美] 孟卫一（Weiyi Meng）　於德（Clement T. Yu）　著

朱亮　译

2017 年 5 月

大数据集成

[美] 董欣（Xin Luna Dong）　戴夫士·斯里瓦斯塔瓦（Divesh Sriva-tava）　著

王秋月　杜治娟　王硕　译

2017 年 5 月

短文本数据理解

王仲远　编著

2017 年 5 月

个人数据管理

李玉坤　孟小峰　编著

2017 年 5 月

位置大数据隐私管理

潘晓　霍峥　孟小峰　编著

2017 年 5 月

移动数据挖掘

连德富　张富峥　王英子　袁晶　谢幸　编著

2017 年 5 月

云数据管理：挑战与机遇

［美］迪卫艾肯特·阿格拉沃尔（Divyakant Agrawal）　苏迪皮托·达斯（Sudipto Das）　阿姆鲁·埃尔·阿巴迪（Amr El Abbadi）　著

马友忠　孟小峰　译

2017 年 5 月

‖ 译者序

近年来，随着信息技术的迅猛发展，各行各业产生和积累的数据量急剧增大。在生产和日常生活过程中，人们不仅需要管理和操作大量的数据，更重要的是，将这些跨领域的大量异构数据进行关联和融合之后，进行相应的分析能产生巨大的价值，如科学发现、商业决策、政府管理、精准医疗等。大数据+深度学习催生了智能革命，正在改变着各行各业，并影响着社会的方方面面。

大数据的关联分析离不开大数据集成，即将多个数据源的数据链接融合在一起。数据集成技术在传统数据库界已经被研究多年，主要针对结构化的关系数据，在模式对齐、记录链接和数据融合等方面取得了许多进展。大数据集成是在大数据背景下的数据集成，具有一些新的挑战，例如数据和数据源的海量性、数据的多样性（即不单单是结构化数据，同时还有许多非结构化和半结构化数据）、数据的动态性等。

本书的作者 Xin Luna Dong 和 Divesh Srivastava 在传统数据集成和大数据集成领域有多年的研究经验，在书中系统地梳理和讨论了该领域中关键问题的一些重要研究成果和方法，对大数据集成的研究者和实践者

都很有帮助，另外本书也可以作为学生学习该领域的入门读物。

本书第 1、2 章由王秋月翻译；第 3 章由杜治娟翻译；第 4～6 章由王硕翻译。最后由王秋月统稿并校订一些关键译法。

由于译者水平有限，书中难免有不当之处，敬请各位读者批评指正。

王秋月

2016 年 9 月

‖前　言

大数据集成是两大重要工作的结合：一个是相对较老的“数据集成”工作；另一个是相对较新的“大数据”工作。

只要存在人们要将多个数据集链接并融合起来以提升它们价值的情况，数据集成就必不可少。早在计算机科学家开始研究这一领域之前，统计学家们就已经取得了许多进展，因为他们迫切需要关联和分析随时间不断积累的普查数据集。数据集成具有很大的挑战性是由多种原因造成的，不仅仅因为我们表示现实世界中实体的方式多种多样。为了有效地应对这些挑战，在过去几十年里，数据集成研究者们已经在一些基础问题（如模式对齐、记录链接和数据融合），尤其是结构化数据的研究上，取得了巨大进步。

近年来，我们在将现实世界中的每个事件和交互都捕获成数字化数据方面的能力增长十分显著。伴随着这种能力的增长，我们渴望从这些数据中分析和抽取出价值，从而迎来了大数据时代。在大数据时代，数据的数量和异构性以及数据源的数目，都极大地增长了，而且许多数据源是非常动态的并且质量千差万别。不同数据进行链接和融合会使数据的价值爆炸性地增大，因而大数据要能使我们做出改变社会各方面的有

价值的、数据驱动的决策，数据集成是关键。

大数据上的数据集成称为大数据集成。本书探讨数据集成研究界在应对大数据集成带来的新的挑战方面已经取得的进展。它的目的是可以作为研究者、从业者和学生想要了解更多关于大数据集成的一个起点。我们试图覆盖该领域内各种各样的研究问题和工作，但显然要全面覆盖这样一个动态发展的领域是不可能的。我们希望本书的读者能对这个重要领域有所贡献，帮助发展大数据的美好愿景。

致谢

本书在成书过程中得到了许多人的帮助。衷心感谢 Tamer Özsu 邀请我们写这本书，感谢 Diane Cerra 管理整个出版过程，并感谢 Paul Anagnostopoulos 制作本书。没有他们温和的提醒、定期的推动和提示编辑，本书的完成将花费长得多的时间。

本书的大部分内容从我们在以下学校开的讲习班和会议上做的大会报告演化而来，这些会议和学校包括：ICDE 2013、VLDB 2013、COMAD 2013、苏黎世大学、ADC 2014 和 BDA 2014 的博士学校。感谢许多同行在报告进行中或之后所给的建设性的反馈。

我们也想感谢许多合作者，他们多年来影响了我们对该研究领域的思考和理解。

最后，感谢我们的家人，他们持续的鼓励和爱的支持使所有的付出更加值得。

Xin Luna Dong 和 Divesh Srivastava

2014 年 12 月

董欣（Xin Luna Dong） 公司高级科学研究员。加入谷歌公司之前，曾在 AT&T 公司研究实验室工作。她拥有美国华盛顿大学博士学位、北京大学硕士学位和南开大学学士学位。研究兴趣主要包括数据库、信息检索和机器学习，特别是在数据集成、数据清洗、知识库和个人信息管理等方面有浓厚的兴趣。已在数据集成方面的顶级会议和期刊上发表 50 多篇论文，并获得 2005 年 SIGMOD 的最佳展示奖（前三名之一）。曾担任 2015 年 WAIM 会议的联合主席，以及 2015 年 SIGMOD 会议、2013 年 ICDE 会议和 2011 年 CIKM 会议的区域主席。

戴夫士·斯里瓦斯塔瓦（Divesh Srivastava） AT&T 公司研究实验室的数据库研究负责人，ACM Fellow，VLDB 基金理事会委员，VLDB 基金会论文集（PVLDB）执行主编，《ACM Transactions on Database Systems》副主编。他拥有威斯康辛大学麦迪逊分校博士学位和印度理工学院孟买分校学士学

位。研究兴趣和发表的著作涵盖了数据管理的各个主题。已在顶级会议和期刊上发表 250 多篇论文。曾担任多个国际会议的主席或联合主席，包括 2015 年 ICDE 会议和 2007 年 VLDB 会议等。