大 数 据 管 理 丛 书

大数据管理概论

孟小峰 编著

机械工业出版社
China Machine Press

大/数/据/管/理/丛/书

大数据管理概论

孟小峰　编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据管理概论 / 孟小峰编著 . —北京：机械工业出版社，2017.3
（大数据管理丛书）

ISBN 978-7-111-56440-9

I. 大… II. 孟… III. 数据处理 - 概论 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 065498 号

本书涵盖大数据管理的理论、方法、技术等诸多方面，集成了大数据融合、存储、分析、隐私和系统等方面的工作。本书共分 7 章：第 1 章描述大数据的概念、演变过程和处理模式；第 2 章提出大数据融合的概念，分析大数据融合的独特性和任务，给出大数据融合的方法论；第 3 章介绍大数据存储与管理方法；第 4 章描述大数据分析技术，包括实时分析、交互分析、智能分析等；第 5 章讲述大数据涉及的隐私问题，主要介绍不同领域中的隐私保护问题及其隐私保护技术；第 6 章介绍大数据管理系统，并分析其体系结构；第 7 章是基于大数据的交叉学科研究，介绍在线用户行为演化的相关研究。本书适合对大数据管理领域有兴趣的学生、研究人员和相关从业人员阅读参考。

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：余　洁　　责任校对：殷　虹
印　　刷：北京文昌阁彩色印刷有限责任公司　　版　　次：2017 年 5 月第 1 版第 1 次印刷
开　　本：170mm × 242mm　1/16　　印　　张：13.25
书　　号：ISBN 978-7-111-56440-9　　定　　价：69.00 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

前 言

陈寅恪先生说：“一时代之学术，必有其新材料与新问题。取用此材料，以研求问题，则为此时代学术之新潮流。治学之士，得预于此潮流者，谓之预流（借用佛教初果之名）。其未得预者，谓之未入流。”对今天的信息技术而言，“新材料”即为大数据，而“新问题”则是产生于“新材料”之上的新的应用需求。

对数据库领域而言，真正的“预流”是 Jim Gray 和 Michael Stonebraker 等大师们。十三年前面对“数据库领域还能再活跃 30 年吗”这一问题，Jim Gray 给出的回答是：“不可能。在数据库领域里，我们已经非常狭隘。”但他转而回答到：“SIGMOD 这个词中的 MOD 表示‘数据管理’。对我来说，数据管理包含很多工作，如收集数据、存储数据、组织数据、分析数据和表示数据，特别是数据表示部分。针对数据查询已经做了相当多的工作，但这些工作仅仅围绕查询画了个‘艾普西龙球面’，而没有真正超越它。所以，如果我们还像以前一样把研究与现实脱离开来，还继续保持狭隘的眼光审视自己所做的研究，数据库领域将要消失，因为那些研究越来越偏离实际。现在人们已经拥有太多数据，而我对许多人说我们仅仅希望拥有更多的时间。所以，整个数据收集、数据分析和数据简单化的工作就是能准确地给予人们所要的数据，而不是把所有的数据都提供给他们。

这个问题不会消失，而是会变得越来越重要。如果你用一种大而广的眼光看，数据库是一个蓬勃发展的领域；如果采用审视的眼光看，现在做的很多研究对 30 年后的人们不会产生任何影响”（见《数据库大师访谈录》）。

最近人们提出了“数据湖”，以区别传统的“数据库”技术。两者的差别到底何在呢？偶读了费孝通先生所著的《乡土中国》后，笔者略有所悟。费老分析总结了中国乡土社会结构，指出中国社会呈现出所谓的“差序格局”，而西方社会呈现的是“团体格局”。传统数据库结构关系单一，呈现状态犹如“团体格局”，即以单个实体为本位，实体之间的关系好比一捆柴，几根成一把，几把成一扎，条理清楚，有共同的模式可循。而当下大数据来源广泛，关系复杂，远近亲疏各不同，这种关系就好比“差序格局”，以语义主题为本位，每类实体都以自我为中心按照与其他实体的语义关系为主线结成网络，这个网络按照语义关系的紧密亲疏呈现“差序”状态，就如同湖面丢下的石子形成的水波纹依中心扩散开去的样子。这种状态随着实体间关系的变化而动态演化，并且每个网络的大小不同，体现的语义关系也不同，蕴含的价值也不同。

数据库的“团体格局”本质上是先有模式后有数据，因此数据集成可以采用中介模式（GAV 和 LAV）以自顶向下的方式实现集成。数据湖的“差序格局”是先有数据后有模式，因此需要按照自底向上的方式以一种大数据融合的方法实现集成。大数据融合即建立数据间、信息间、知识片段间多维度、多粒度的关联关系，实现更多层面的知识交互，从而聚敛出数据湖中一个个维系我们社会的“水波纹”（即语义关联的紧密程度）。

本书集成了大数据融合、存储、分析、隐私和系统等方面的工作，其组织结构如下：第 1 章描述大数据的概念、演变过程和处理模式；第 2 章提出大数据融合的概念，分析大数据融合的独特性和任务，给出大数据融合的方法论；第 3 章介绍大数据存储与管理方法；第 4 章描述大数据分析技术，包括实时分析、交互分析、智能分析等；第 5 章讲述大数据涉及的隐私问题，主要介绍不同领域中的隐私保护问题及其隐私保护技术；第 6 章介绍大数据管理系统，并分析其体系结构；第 7 章是基于大数据的交叉

学科研究，介绍在线用户行为演化的相关研究。

本书中涉及的研究工作得到众多科研项目的支持，其中包括：国家自然基金重点项目——“大规模关联数据管理的关键技术研究”（编号：61532010）；国家自然基金重点项目——“面向大数据内存计算的计算机体系结构”（编号：61532016）；国家重点研发项目——“科学大数据管理系统”（编号：2016YFB1000600）；中国人民大学重点科学研究基金重大基础研究项目——“社会计算若干关键问题研究”（编号：11XNL010）；高等学校博士学科点专项科研基金优先领域课题——“云计算环境下的在线聚集技术研究”（编号：20130004130001）；国家自然基金重大研究计划重点项目——“大数据开放与治理中的隐私保护关键技术研究”（编号：91646203）。

本书架构的安排以及统稿、审校工作由孟小峰组织完成，这里要特别感谢王春凯、杜治娟、郭崎、杨晨、王硕、叶青青和李勇，在本书的编写过程中他们给予了极大的帮助。

本书涉及面广，内容丰富，术语量大，如果在阅读过程中发现有不当之处，恳请读者批评指正；如果有任何建议或意见，欢迎发邮件与作者（xfmeng@ruc.edu.cn）联系。

孟小峰

2016 年 9 月 28 日于北京

作者简介 ‖

孟小峰 中国人民大学信息学院教授，博士生导师。现为中国计算机学会会士、中国保密协会隐私保护专业委员会副主任，《Journal of Computer Science and Technology》《Frontiers of Computer Science》《软件学报》《计算机研究与发展》等编委。先后获中国计算机学会“王选奖”一等奖（2009 年），北京市科学技术奖二等奖（2011 年）等奖励，入选“第三届北京市高校名师奖”（2005 年）。发表论文 200 余篇，近期结合近十年的研究工作出版了“网络与移动数据管理三部曲”（《Web 数据管理：概念与技术》《XML 数据管理：概念与技术》《移动数据管理：概念与技术》，清华大学出版社），获得国家专利授权 12 项。近期主要研究领域为网络与移动大数据管理，包括 Web 数据管理、云数据管理、面向新型存储器的数据库系统、大数据隐私管理、社会计算等。