多根层次数据分布模型
——论大数据时代的数据管理

张建英　著

科学出版社

多根层次数据分布模型

——论大数据时代的数据管理

张建英　著

科学出版社

北京

内 容 简 介

人类进入信息社会大数据时代，传统数据管理面临很多挑战，数据管理正面临一场科学革命。本书从大数据发展现状出发，在人类 DIKW 知识层次中认识“数据”，阐述大数据时代以数据为中心的必然性，进而提出数据管理的新范式，即以系统科学及开放复杂巨系统为主要特征的范式，并论述数据管理正在向新范式转换；为解决数据系统中众多管理问题，从数据语义出发给出数据分布模型概念，并论述其是大数据时代数据管理的核心与基础；定义了一种数据分布模型——MHM；另外，本书还涉及数据管理的几个主要方面，包括数据一致性、事务处理、访问控制、扩展性等，实验表明 MHM 在性能、可靠性方面的优势，同时讨论 MHM 潜在的适用范围。

本书可以作为高等院校数据管理相关专业研究生的教学参考书，也可供相关领域的科研人员参考。

图书在版编目(CIP)数据

多根层次数据分布模型：论大数据时代的数据管理/张建英著. —北京：科学出版社, 2017.5

ISBN 978-7-03-052571-0

Ⅰ.①多… Ⅱ.①张… Ⅲ.①数据管理—研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2017) 第 083060 号

责任编辑：王 哲 赵微微 / 责任校对：郭瑞芝

责任印制：张 倩 / 封面设计：迷底书装

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 5 月第 一 版 开本：720 × 1000 1/16

2017 年 5 月第一次印刷 印张：15 3/4

字数：310 000

定价：89.00 元

(如有印装质量问题，我社负责调换)

前　言

本书的工作可以追溯到 9 年前我参加的一次学术会议。大学毕业后我做了几年软件开发，然后重返高校攻读硕士学位。这期间我研发了一个主存数据库管理系统原型，实现了存储管理、T 树索引、查询处理、乒乓检查点、事务管理、SQL 访问等功能，这激发了我继续从事数据管理研究的兴趣，进而留校任教。2007 年参加在燕山大学举办的 VLDB School，王珊老师讲授的模式子图与数据子图概念启发了我，使我开启了数据分布模型研究工作。

我最初的想法是利用数据之间的这种语义关系研究数据复制技术。随着研究的深入，我认识到这种语义关系不仅可以用来进行数据复制，而且在分布式数据库中，对如查询优化、访问控制等都会产生影响。2008 年校在职博士研究资助计划也促使我整合、提高自己的想法，有了数据分布模型的初步想法。

为了验证数据分布模型想法的创新性，我查阅了数据库领域大量的经典文献。然而，除了 1974 年的数据库界著名的大讨论中提及在关系模型上构建一层数据模型，以及 ER 模型论文中提及的数据模型 (视图) 分为四个级别之外，所获不多。但通过阅读这些经典文献，对关系模型、层次模型、网状模型、数据分布等基本问题有了更为深入的认识，也更加认识到所从事研究的基础性。

数据分布模型研究在挫折与希望之间缓慢推进。2009 年我参加了华中科技大学举办的“分布式计算与系统”全国研究生暑期学校之后，受到云计算概念的影响，开始将数据分布模型与云计算结合起来，就扩展性、安全性展开研究。2010 年我参加中国人民大学举办的“数据密集型计算和非结构化数据管理”会议，期间与王会举博士进行了交流。根据他的建议，我后来实现了基于 32 个计算节点的性能、扩展性的对比实验。2011 年我在博士学位论文答辩时，王希诚教授建议我进一步查阅下钱学森的开放的复杂巨系统理论。之后的几年中，我阅读了系统哲学、开放的复杂巨系统、科学革命的结构以及信息哲学等方面的一些文献，进一步拓展了我的视野，让我从不同的视角来思考问题与看待眼前的困难，也更加坚信自己所做的工作。

2015 年“十一”长假期间，同于红教授的一次深入交流中发现，自己对数据分布模型工作的意义、重要性、理论依据还一直未能阐述清楚。而要将这些都阐述，哪怕只是泛泛之谈，也非一篇论文所能承载得了的，也难成体系，因此有了整理数据分布模型研究成果的初步想法。该月下旬，第 32 届中国数据库学术会议最后一天，在“大数据对科研与教学带来的问题与思考”的讨论会上，几位专家各抒己见：

高宏教授提出“大数据管理系统”，王晓阳教授对于数据语义做了强调，杜小勇教授明确提出“数据库系统面临一场革命”“模型是主线”“系统是核心”“应用是动力”等。这些都激起我的共鸣，也促使我着手整理本书。

数据分布模型的最终确立不会仅取决于理论研究，更非性能的简单比较，这其中有着非常复杂的哲学、社会、系统、管理、技术等综合的因素。正如当初关系模型与网络模型的争议一样，其最终确立是要靠市场的选择。人类进入信息社会的时间相对人类社会漫长的历史来说还太短，信息管理的技术也必将随着人类的发展而发展。信息管理也在不断提出新的需求，这些造就了过去数据库系统的成绩，今天的需求又高了一个层次，同样也要求今天的数据管理再上一个台阶。本书的目的并非提供一个完整的数据分布模型，而是借用 MHM 对数据管理进行展望与勾勒，激发业界对数据分布模型的关注，引起大数据时代数据管理方面更多的思考。本书涉及数据管理以及信息哲学、系统科学、管理等多个学科，数据管理本身也关乎其核心内容，所需知识的广度、深度、综合性可想而知，限于个人的能力，粗浅之处请多包涵，存在的不妥之处恳请读者批评指正!

感谢王秀坤教授给予我生活上的关怀和工作上的帮助以及学业上的教导；感谢刘洪波教授、于红教授认真倾听我的想法并不厌其烦地帮助我斟酌论文；感谢课题组的孟军副教授、杨南海老师对于我研究工作的建议；感谢杨元生教授、王希诚教授、滕弘飞教授、申彦明教授给予我的鼓舞与帮助。感谢郭崇慧教授对于我书稿的建议；感谢朱明华教授、郜激扬老师为我提供实验条件；感谢王宇新副教授、孙永奇教授，在我情绪低落时给我的勉励，懈怠时给我的鼓舞，迷茫时给我的警醒；感谢王会举博士对于 MHM 性能实验的建议；感谢课题组的研究生刘淼、林敏泓、王铁存、李鑫 (男)、李鑫 (女)、刘健男、李苗苗、孙永洁的支持；感谢在研究之路上给我启迪、默默关心与帮助我的所有的人；最后要把我的感谢留给家人，他们的关爱与付出，最终使得我完成本书。

作　者

2016 年 10 月于大连理工大学

目　录

插图目录

表格目录

第1章 导 论

本章围绕数据管理，从时代背景、未来展望、技术途径、章节组织结构几方面对全书概述。首先论述大数据时代数据管理面临的挑战；接着提出社会数据管理的必然及特征，得出基于数据语义进行层次化数据管理的结论；然后引出数据分布模型概念，作为大数据时代的数据管理核心问题 —— 数据分布情况的假设，这样数据一致性、事务管理、访问控制、扩展、自适应自组织、模式演化等方面就有望在一个框架内解决，降低总的社会数据管理成本。

1.1 数据管理面临着一场科学革命

人类社会经历了农业社会、工业社会，现在进入了信息社会，信息所附加的活动已开始成为人类重要的活动之一。1946 年出现第一台电子计算机以来，特别是随着互联网的出现，信息技术得到了飞速的发展，激起了一轮又一轮的信息技术革命。互联网、物联网、移动互联、网格、云计算 [1−6]、传感器网络、互联网 + 等概念层出不穷，社会信息化程度不断深入。随着大数据 [7] 概念的兴起，信息社会开始进入大数据时代。

信息技术、信息应用、信息需求都发展得太快，以至于我们在信息社会高速发展的洪流中，被裹挟着，被一个又一个新的概念所冲击，很难慢下来进行冷静思考：信息、数据究竟在人类的认识层次中处于何等地位；我们究竟应该怎样做技术的主人，而非数据的奴隶；作为个体的人怎样在大数据时代找到个体的生存价值与意义；作为从事信息学科工作的人，如何从其他的学科吸收营养，如何从几千年的历史文化中获得养料，而非将自己囿于学科的小天地中。信息技术的快速发展将人、物、各种社会活动都组织成一张巨大的网，这张网越来越大，涉及越来越多的领域，也越来越厚重。如何认识这张网，如何构建这张网都是今天的人类要关注的。既需要从哲学层面认识其本质，也需要从社会影响、思维方式上关注，作为 IT 界的人士，更要在技术层面探讨其构建，让其服务于人类、造福于人类，而非被其所控制。

信息社会进入到了大数据时代，“数据”在社会的重要性不言而喻，数据量飞速增长给传统的数据处理技术带来了挑战。通常数据业务不尽相同，数据处理需要编写不同的应用程序加以解决。数据管理技术作为数据处理的基本环节，是所有数据处理过程中的必有部分。数据处理与数据管理是相联系的，数据管理技术的优劣将对数据处理的效率产生直接影响。由于可利用的数据呈爆炸性增长，且数据的种

类繁杂，要有效地管理数据非常复杂。因此客观需要一个通用、使用方便、高效的数据管理软件，把数据有效地管理起来为数据处理使用。数据库技术就是针对该目标进行研究并发展和完善起来的计算机应用的一个分支。但是，大数据时代给以关系数据库为核心的数据管理技术带来了新的课题。

在大数据时代，数据库的前提，特别是集中式数据库技术的前提发生了很大的变化，这些变化主要体现在以下几点。

(1) 数据类型、数据量、数据变化速度、应用场景都发生了一些变化。数据类型，不只是关系数据，还有其他结构化数据、XML 等半结构化数据，乃至非结构化数据。由于产生数据不只是手工输入，还有仪器采集、社交网络生成、其他应用的输出等来源，数据产生变得更为快捷，产生的数据量也更为巨大。同时，社会活动节奏的加快、实时控制等需求使得数据变化的速率提高。数据的关注点也从以信息管理为主转变为信息管理与数据分析并重。

(2) 与外围系统的关系发生巨大的变化。传统的数据管理主要以企业、行业各类组织的自有系统为核心，外围的数据由人工输入或者通过程序转入，虽然效率不高，但能够满足企业的基本需求。由于企业主要将关键、核心的功能进行信息化，通过信息系统来管理，这样数据不一致性等问题不是很突出。而随着互联网的深入影响，不但人与人之间逻辑上通过网络连接，人与物通过物联网也可以连接起来。进一步来说，其他的各类设备、生产制造等社会活动的很多方面都会相互关联。系统之间的关联在数量上变多，在复杂程度上也大大增加。

(3) 计算模式发生很大的变化。在传统的信息管理系统中，经常是自建计算中心、数据中心、信息中心、网络中心等来支撑自有的信息系统。计算模式先后出现主机模式、客户机/服务器（Client/Server，C/S）模式、浏览器/服务器（Browser/Server，B/S）模式、云计算等。随着网格、虚拟化技术，特别是云计算概念的兴起，可以通过公有云、私有云来组织企业的计算。大数据时代数据管理对于计算的弹性、可扩展性方面的需求更为迫切。现在人们面临多种计算模式的选择。取长补短、相互融合必将成为发展的方向，各种应用系统必将易于在各种计算模式间便捷地迁移、整合，私有云、公有云、各片云之间的壁垒也终将被打破。

(4) 数据与计算的关系带来变化。在数据与计算的关系上，曾主要是以计算为中心来组织计算过程。早期的计算机，只能做计算就是一个好的例子：数据通过读卡机输入，再通过打孔机输出，本身不具有数据管理的功能。随着硬盘为代表的外存的出现，开始使用文件来管理数据，当然是以计算为中心。数据库管理系统出现后，情况略有变化，在一定程度上是以数据管理系统为核心来组织计算。进入大数据时代以来，Hadoop 等技术的兴起，实质上仍是以计算为中心的批处理思想。大数据时代，客观上要求以数据为中心来组织计算。

(5) 从企业以自有信息系统为核心，到社会活动以社会信息系统为核心。早期

的信息系统，以企业自有的信息系统为主要特征，经过多年快速的信息化进程，人类正在进入到基于各企业信息系统，各行业、政府、各领域融合的社会信息系统为核心时代。这个社会信息系统，既包含上述多个相对独立的系统，又对其有一定的影响、控制能力。在企业数据管理时代，我们以一种超越个别应用的视角组织、存储、管理企业的数据；在全社会数据管理时代，需要一种超越个别企业、行业的视角来组织、存储、管理数据。

上述数据管理前提的变化，自从出现了局域网就开始了，只是随着互联网、物联网的发展，整个社会信息化程度加深、广度加大，变得更为明显。就数据库来说，集中式的数据管理技术也在不断地进行技术升级与更新以适应新的数据管理环境。从集中式数据库到分布式数据库，从分布式数据库到大规模分布式数据库，再到云数据库等反映的就是这样的变化。在这些变化中，一些传统的数据管理技术，如串行化、强一致性、扩展性等变得越来越难以适应新的形势，一些以大数据为背景的新技术、新理论开始出现。正如信息管理领域的一些专家所断言的那样，数据管理面临着一场科学革命。

1.2 社会数据管理

库恩于 1962 年出版的《科学革命的结构》[8] 一书中就科学革命做了翔实的论述。他提出了“范式”的概念，指的是常规科学所赖以运作的理论基础和实践规范，也有中文译本 [8] 将“paradigm”译为“规范”。“范式”是从事某一科学的科学家群体所共同遵从的世界观和行为方式。这样，进行研究的人们，受同样的科学实践规则和标准所制约。这种制约以及由此所造成的表面上的一致，正是常规科学的前提，也是某一种研究传统形成和延续的起源。书中还指出，一种范式经过科学革命向另一范式逐步过渡，正是成熟科学的通常发展模式。范式的变革不可能是知识的直线积累，而是一种创新和飞跃，一种科学体系的革命。当今的数据管理面临的就是这样的问题，在原有的范式框架内修修补补已无法适应大数据时代数据管理的要求，需要的是一种范式的变革。

只要简单地思考下传统的数据库系统与当今的数据库系统，乃至大数据时代现在的数据管理系统，不难得出这样的结论，系统变得更大了，系统也变得更复杂了，系统之间的关系也越来越丰富了。将来的系统又会是怎样的系统？我们怎样认识这些系统的异同？贝塔朗菲的《一般系统论：基础、发展和应用》[9]、拉兹洛的《系统哲学引论 —— 一种当代思想的新范式》[10]、哈肯的《协同学：大自然构成的奥秘》[11]、魏宏森与曾国屏的《系统论：系统科学哲学》[12]，以及钱学森等的开放的复杂巨系统 [13–15]方面的论述可以帮助回答这些疑问。传统的数据库管理系统，元素规模较小，元素间结构简单，可以划归到简单系统，最多是简单巨系

统，而大数据时代的数据管理，面临的是各种层次下的数据，数据不但量大，而且结构复杂，与外界进行数据、信息交互又非常的频繁，这应该归为开放的复杂巨系统。开放的复杂巨系统是所有系统中最为复杂的一种，还有相当多理论没有搞清楚，目前还没有形成从微观到宏观的完整理论。开放的复杂巨系统研究需要有新的方法论，一方面要吸收已有的方法论的长处，同时也要有新的发展。钱学森于1989年提出了研究开放的复杂巨系统的方法论，这就是从定性到定量综合集成方法（meta-synthesis），简称综合集成方法。1992年，钱学森又提出“从定性到定量综合集成研讨厅体系”思想。这里的定性与定量相结合的综合集成方法，被认为是研究处理开放的复杂巨系统的当前唯一可行的方法。

贝塔朗菲的《一般系统论：基础、发展和应用》开辟了从系统角度认识事物、对象之间的相互联系、相互作用的共同本质，内在规律性的研究领域。之后，系统论以一种崭新的科学方法论活跃于国际学术论坛。林康义与魏宏森所译的《一般系统论：基础、发展和应用》译者序中，称“系统论是继相对论和量子力学之后，又一次改变了世界的科学图景和当代科学家的思维方式”“是思想领域大变动的一个重要标志”。贝塔朗菲将一般系统论看作科学思维的新范式。现代科学思维正由机械论的范式，转变到一般系统论的范式。

在传统的数据库管理系统中，我们自发地、凭直觉地使用“系统”这一概念，由于系统相对简单，问题不大。随着问题规模的扩大，问题越来越复杂，特别是在大数据时代，我们需要自觉地在数据管理中应用一般系统论方面的理论以及开放的复杂巨系统研究成果。数据管理系统与自然系统、人类社会不同，不是经历漫长的自然、社会过程演化而来，它出现很晚，自然演化可能会很慢。当认识到一般系统论的基本原理后，我们可以自觉、主动地人为设计、构建、影响这个人工系统的演化进程。

我们知道，信息管理以数据库系统为核心与基础。数据库系统的演化发展反映了信息管理演化与发展的脉络。从层次数据库、网络数据库到关系数据库；从集中式数据库到分布式数据库、多数据系统，还有与云计算结合的云数据库等技术，以至于现在的NoSQL数据管理趋势，数据管理技术总体上是从集中到分布，从小规模分布到大规模分布；从单一结构化数据的管理，到结构化、非结构化、半结构化数据的管理，到正在形成不同数据质量层次的数据管理；从企业自有数据管理，到跨企业、跨应用、跨行业的全社会数据管理系统；网络通信技术的快速发展，以及社会对于高效、准确、弹性、稳定、统一、多质量层次的社会数据管理的需要与现实的数据管理技术存在一个不小的差距。数据管理技术，脱胎于集中式数据管理时代，现在很多的前提都发生了很大的变化，是时候考虑未来的数据管理了。

本书认为未来的数据管理应是全社会数据管理，应同时具有高效、准确、弹性、稳定、统一、多质量层次的特征。社会的数据管理，指的是我们每个人、每个企业、