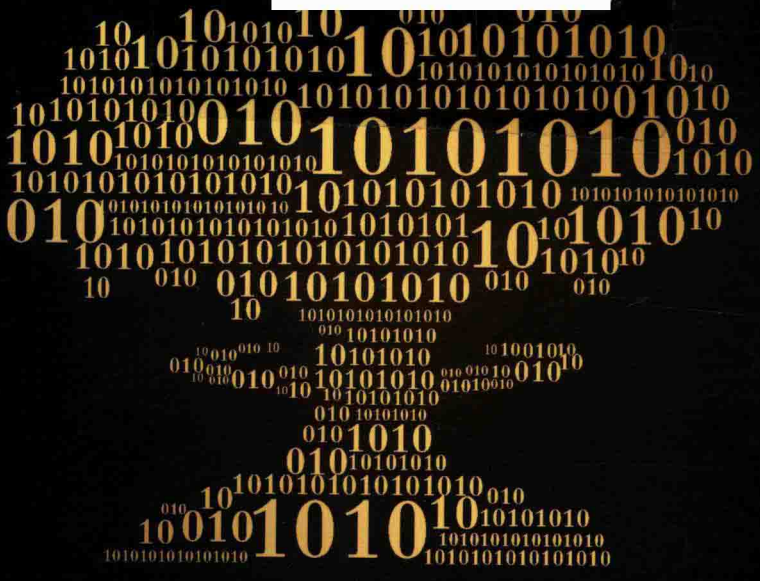
大数据作为一门崭新的学科，尚未形成完整的理论体系，仍存在许多关键问题尚待解决。

本书将与您一同探寻大数据背后的基础理论与核心技术；
并在剖析教育、医疗、金融、交通等典型应用的基础上讨论未来趋势。

The Core Technologies Behind Big Data

大数据背后的核心技术

张桂刚　李超　邢春晓　编著



中国工信出版集团　電子工業出版社 PUBLISHING HOUSE OF ELECTRONICS INDUSTRY http://www.phei.com.cn

大数据背后的核心技术

张桂刚　李　超　邢春晓　编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书分为三大部分，分别为大数据基础理论分析、基于海量语意规则的大数据流处理技术及大数据应用。

第一部分介绍大数据领域的主要基础理论，包括大数据基本概念、可编程数据中心、云文件系统、云数据库系统、大数据并行编程与分析模型、大数据智能计算算法、基于大数据的数据仓库技术、大数据安全与隐私保护，以及基于大数据的语意软件工程方法等。

第二部分介绍基于海量语意规则的大数据流处理技术，包括基于规则的大数据流处理介绍、语意规则描述模型、海量语意规则网及优化、海量语意规则处理算法及海量语意规则并行处理等。

第三部分主要介绍大数据的一些典型应用，包括：文化大数据、医疗健康大数据、互联网金融大数据、教育大数据、电子商务大数据、互联网大数据、能源大数据、交通大数据、宏观经济大数据、进出口食品安全监管大数据、基于大数据的语意计算及典型应用（含语意搜索引擎、语意金融、语意旅游规划、基于海量语意规则的语意电子商务）。最后探讨了大数据未来的研究方向。

本书可供希望较全面、深入地了解大数据及其应用的读者学习参考。

图书在版编目（CIP）数据

大数据背后的核心技术 / 张桂刚，李超，邢春晓编著. —北京：电子工业出版社，2017.1
ISBN 978-7-121-30296-1

Ⅰ. ①大… Ⅱ. ①张… ②李… ③邢… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 269600 号

策划编辑：陈韦凯
责任编辑：陈韦凯　　文字编辑：毕军志
印　　刷：涿州市京南印刷厂
装　　订：涿州市京南印刷厂
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：787×1 092　1/16　印张：21.25　字数：544 千字
版　　次：2017 年 1 月第 1 版
印　　次：2017 年 1 月第 1 次印刷
印　　数：3 500 册　　定价：65.00 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）88254888，88258888。

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。

本书咨询联系方式：chenwk@phei.com.cn，（010）88254441。

前　言

随着 Web 2.0 技术的发展，尤其是移动互联网的飞速发展，每个人、每台手机、每个 iPad 及每台血压计、血糖测量仪等各种智能移动设备无时无刻不在产生数据。大数据（Big Data）正在不断地渗透到人们生活中的每个角落，也在不断地改变人们的生活方式，并引导新兴的产业革命，在给传统行业带来巨大冲击的同时也带来了巨大的新机遇和挑战。一个企业甚至一个国家拥有的数据规模和质量，以及处理和分析数据的能力，已经成为判断一个企业或者一个国家竞争力的最为重要的标志之一，拥有多少大数据资源及如何管理并使用这些大数据资源，已经成为是否具有核心竞争力的关键因素。为了迎接大数据带来的各种挑战和机遇，全球各个国家和企业对大数据的重视程度均达到了一个前所未有的高度。从全球角度来看，很多国家已经把大数据作为一项国家科技意志。例如，美国政府已经制订了大数据研究和发展计划，日本为了增强经济活力提出了大数据战略计划等。不仅如此，一些知名公司如 Google、IBM 及 EMC 等也成立了专门的大数据研究机构，以应对在大数据研究和应用中的各项关键技术挑战及应用实现所面临的问题。

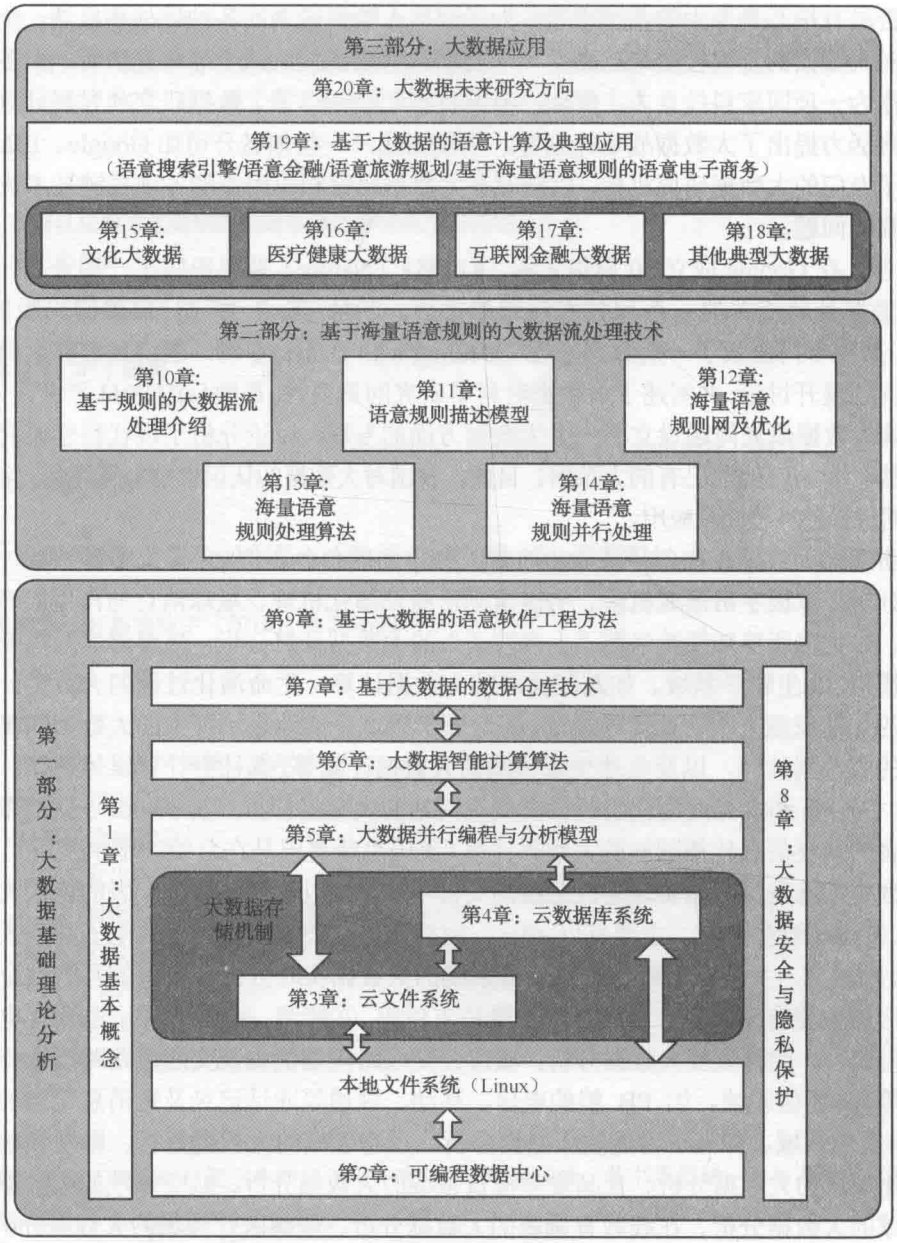
2008 年，在 Google 成立 10 周年之际，《自然》（Nature）杂志出版了一期专刊，专门讨论了未来大数据处理相关的一系列技术问题和挑战。2011 年 2 月 11 日美国出版的《科学》（Science）期刊专门出版了一期数据处理（Dealing with Data）专辑，围绕目前科学研究的海量数据处理问题展开讨论，并阐述了大数据对科学研究的重要性。在随后的 2011 年 9 月 4 日，《自然》再次就大数据研究问题设立了一个大数据方面的专题，讨论分析了现代科学研究面临的一个巨大挑战，即如何处理已有的大数据。目前，我国对大数据的认识也越来越深刻，各行各业均利用大数据进行各种研究及应用。

如上所述，大数据正在各行各业扮演着十分重要的角色，例如：①天文学领域。如通过对大数据的分析，掌握宇宙形成机理、宇宙黑洞形成及演化机理、星球消亡与再生原理等。②物理学领域。如大家所熟知的希格斯“上帝粒子”的大数据计算分析，核弹爆炸及氢弹爆炸的大数据计算模拟。③生物学领域。如基因排序的大数据计算，生命演化过程的大数据计算模拟及生物制药的化学反应大数据计算模拟等。④地理学领域。如地震预警中的大数据计算，海啸预警和防范的大数据计算，以及全球变暖预测的大数据计算等。⑤社会计算媒体领域。主要有以 Facebook、Google 和人人网为代表的社交交友网站的大数据计算，以 Twitter、新浪微博及腾讯微博为代表的社交信息传播网站的大数据计算（美国总统奥巴马在总统选举中采用了对 Twitter 大数据的分析，这是帮助他实现连任总统的关键所在），以天涯论坛为代表的论坛大数据的分析计算等。⑥电子商务领域。主要有以 eBay、阿里巴巴、淘宝网为代表的电子商务大数据计算分析。⑦金融领域。主要有银行及股票交易系统的大数据实时分析，新兴的互联网金融或者大数据金融形态主要有余额宝、百度百发及微信支付等。⑧能源、交通领域。主要有电网的大数据实时分析监控，能源调度大数据分析，城市公交线路规划优化及交通道路路线选择的大数据实时分析等。⑨通信领域。如 PB 级的电信、移动、联通等通话记录及短消息记录的大数据计算分析。⑩其他领域。如人工智能的大数据分析、反恐领域的大数据分析、影视领域的大数据分析、文化领域的大数据分析、食品安全检查领域的大数据分析、航空领域的大数据分析、电子商务领域的大数据分析、在线教育领域的大数据分析、健康医疗领域的大数据分析等。

大数据已经成为全球及全社会各行各业最为重要的战略资源。如何管理好大数据，并从大

数据中挖掘出它的潜在价值将是大数据未来的主要发展方向。大数据将普遍应用于国民生产中的各个领域，包括政府、医疗、经济、社会、教育、航空航天、军事及互联网和物联网等各个领域。本书后面几章将给出一些具体的案例进行初步分析，以期更深入地从应用的角度理解大数据及其在各种应用中的价值所在。

如何处理这些密集型应用所需的大数据显得越来越重要。与其他学科不同，大数据作为一门崭新的学科，尚未形成一套理论体系，依然存在许多关键的问题没有解决，甚至在大数据这门学科中到底有哪些基础理论、关键问题、核心技术等都没有一个完整的概念。鉴于此，本书研究大数据背后的核心技术并对一些具体的应用领域进行了分析。下图展示了本书的总体架构和研究内容。



本书章节关系图

第 1 章：大数据基本概念。本章主要分析大数据的一些基本概念，包括大数据定义、大数据度量、大数据表示、大数据的语意理解及大数据和云计算的关系等。

第 2 章：可编程数据中心。本章设计了一种可编程数据中心模型，该可编程数据中心模型将充分考虑能源消耗、基于各种智能调度的大数据放置方法等。

第 3 章：云文件系统。本章主要分析了现有的常用云文件系统，如谷歌 GFS，Hadoop HDFS 等，并分析了现有云文件系统的缺陷，最后提出了一种新的语意云文件系统的简要设计思路 SCFS。

第 4 章：云数据库系统。本章主要分析了现有的常用云数据库系统，如谷歌 BigTable、Hadoop HBase 等，并分析了现有云数据库系统的缺陷，最后提出了一种新的语意云数据库系统的简要设计思路。

第 5 章：大数据并行编程与分析模型。本章主要分析了现有的常用大数据并行编程与分析模型，如谷歌 MapReduce、Hadoop MapReduce、Hadoop++、Twister 等，并分析了现有大数据并行编程与分析模型的缺陷，最后提出了一种新的大数据并行编程与分析模型的简要设计思路 SemanMR。另外，为了提高大数据实时处理效率，本章设计了一种初步的大数据实时处理方法。

第 6 章：大数据智能计算算法。本章主要总结了当前大数据智能计算常用的一些智能算法，并做了相应的分析。

第 7 章：基于大数据的数据仓库技术。本章分析了现有的常用大数据仓库技术，如 Hive、Pig 等，并提出一种新的基于大数据的数据仓库技术的简要设计思路。

第 8 章：大数据安全与隐私保护。本章介绍了在云环境下的大数据安全与隐私保护机制及相应的各种方法和算法。

第 9 章：基于大数据的语意软件工程方法。本章根据大数据这门新学科的特点，提出了一种基于大数据的语意软件工程的方法，为基于大数据的软件系统的开发提供了一种新的软件工程的研究、设计和开发思路。

第 10 章：基于规则的大数据流处理介绍。本章介绍了基于规则的大数据流处理所涉及的一些基本概念及基础知识。

第 11 章：语意规则描述模型。本章介绍了一种可以表示各种粒度（大粒度、中粒度及小粒度）规则的语意规则描述模型。主要包括语意规则节点表示方法、语意规则节点流量及语意规则节点可计算代价等。

第 12 章：海量语意规则网及优化。本章介绍了基于规则合并及基于规则模块等价替换的海量语意规则网优化方法。本章通过研究语意规则，将不同语意规则中有重复语意规则的节点进行合并，达到语意规则完全合并或部分合并的目的；同时，本章通过分析那些计算功能等价的语意规则模块，用计算代价小的语意规则模块替换计算代价大的语意规则模块。

第 13 章：海量语意规则处理算法。本章在分析现有的各种规则模式匹配处理算法的基础上，针对现有规则模式匹配处理算法的缺陷，介绍了一种适合于海量语意规则的海量语意规则模式匹配处理模型及运行时的处理算法。

第 14 章：海量语意规则并行处理。本章提出并研究了一种海量语意规则并行处理机制 GAPCM。介绍了将海量语意规则生成互相独立的规则子网的方法；任务预分配方法；语意规则子网的合理划分方法；语意规则子网内部通信及处理机之间的外部通信；将任务具体映射到所对应处理机的方法。

第 15 章：文化大数据。本章从大数据在文化领域的应用角度分析了大数据在公共文化、图书馆、博物馆、艺术馆、科技馆、艺术馆及美术馆这种文化领域的数据采集、存储、计算分析及应用方法和典型应用。

第 16 章：医疗健康大数据。本章从大数据在医疗健康领域的应用角度分析了医疗健康领域如何利用大数据进行数据的组织、存储、计算分析及应用方法和典型应用。

第 17 章：互联网金融大数据。本章从大数据在金融领域的应用角度分析了互联网金融领域如何利用大数据进行数据的组织、存储、计算分析及其应用的方法和典型应用。

第 18 章：其他典型大数据。我们在第 15、16 及 17 章中分别介绍了文化大数据、医疗健康大数据及互联网金融大数据。大数据的应用现在已经遍布各个领域，本章对教育大数据、电子商务大数据、互联网大数据、能源大数据、交通大数据、宏观经济大数据、食品安全监管大数据等进行了一个简要的阐述。

第 19 章：基于大数据的语意计算及典型应用。由于大数据的产生，语意计算（Semantic++ Computing）也应运而生。语意计算（Semantic++ Computing）是在语义计算（Semantic Computing）和语意计算（Semantic+ Computing）基础上加上大数据技术的应用而产生的一种新的计算模式。本章分析了基于大数据的各种语意计算的应用，如在社交网络方面的应用、政府方面的应用等，最后又具体介绍了基于大数据的语意计算应用，包括语意搜索引擎、语意金融、语意旅游规划及基于海量语意规则的语意电子商务。

第 20 章：大数据未来研究方向。本章简要描述了大数据未来的发展方向及主要应用方向等。

作　者

目　录

第一部分　大数据基础理论分析 …… (1)
第 1 章　大数据基本概念 …… (2)
1.1　大数据定义 …… (2)
1.2　大数据度量 …… (3)
1.2.1　大数据能耗度量 …… (3)
1.2.2　大数据计算能力度量 …… (4)
1.2.3　大数据的数据中心服务能力度量 …… (4)
1.2.4　大数据商业与社会价值度量 …… (4)
1.2.5　大数据冷热度度量 …… (5)
1.3　语意计算的发展过程 …… (5)
1.3.1　语义计算（Semantic Computing） …… (5)
1.3.2　语意计算（Semantic+ Computing） …… (5)
1.3.3　语意计算（Semantic++ Computing） …… (6)
1.3.4　语意计算和大数据 …… (7)
1.4　大数据的语意理解 …… (8)
1.4.1　大数据资源语意存储 …… (9)
1.4.2　大数据资源语意信息获取 …… (9)
1.4.3　语意资源管理 …… (9)
1.4.4　大数据语意处理 …… (10)
1.4.5　大数据语意服务（语意分析/语意合成等） …… (10)
1.4.6　大数据语意安全与隐私 …… (10)
1.4.7　语意接口 …… (10)
1.4.8　基于语意的大数据应用 …… (10)
1.5　大数据和云计算 …… (11)
1.5.1　云计算 …… (11)
1.5.2　大数据和云计算的关系 …… (11)
本章小结 …… (12)
第 2 章　可编程数据中心 …… (13)
2.1　可编程数据中心体系架构 …… (13)
2.2　数据分配管理 …… (14)
2.2.1　数据分配管理原理 …… (14)
2.2.2　数据分配管理案例 …… (17)
2.3　异构数据节点分配管理 …… (19)
2.3.1　异构数据节点分配管理方法 …… (20)

2.3.2 异构数据节点服务能力计算方法 …… (22)
2.4 规则管理 …… (23)
2.4.1 规则 …… (23)
2.4.2 语意规则 …… (24)
2.4.3 海量语意规则管理架构 …… (24)
2.5 数据放置策略 …… (25)
2.5.1 谷歌的数据放置策略 …… (25)
2.5.2 Hadoop 的数据放置策略 …… (26)
2.5.3 其他常用的数据放置策略 …… (26)
2.5.4 语意数据放置策略 …… (26)
2.6 可编程数据中心机房架构 …… (30)
本章小结 …… (30)
第 3 章 云文件系统 …… (32)
3.1 常用云文件系统综述 …… (32)
3.2 语意云文件系统 SCFS …… (34)
3.2.1 SCFS 系统架构 …… (34)
3.2.2 SCFS 大小文件处理机制 …… (36)
3.2.3 数据一致性保障 …… (40)
3.2.4 元数据集群管理技术 …… (40)
3.2.5 副本管理策略（负载均衡机制） …… (41)
本章小结 …… (44)
第 4 章 云数据库系统 …… (45)
4.1 常用云数据库系统综述 …… (45)
4.2 语意云数据库系统 SCloudDB …… (47)
4.2.1 SCloudDB 系统架构 …… (47)
4.2.2 SCloudDB 设计思路 …… (48)
4.2.3 SCloudDB 的 SRegion 定位机制 …… (50)
4.2.4 多维及海量随机查询机制 …… (51)
4.2.5 支持多维及海量随机查询的语意搜索机制 …… (52)
4.2.6 大表划分方法 …… (54)
4.2.7 基于列族存储及语意的大表划分机制 …… (56)
4.2.8 分布式同步关键技术 …… (57)
本章小结 …… (59)
第 5 章 大数据并行编程与分析模型 …… (60)
5.1 大数据并行编程与分析模型综述 …… (60)
5.2 大数据并行编程与分析模型 SemanMR …… (63)
5.2.1 SemanMR 体系架构 …… (63)
5.2.2 SemanMR 技术思路 …… (64)

5.3 SemanMR 关键技术 …… (66)
5.3.1 基于语意的调度器关键技术 …… (66)
5.3.2 SemanMR 的作业/任务状态交互新规则 …… (68)
5.3.3 语意映射器关键技术 …… (69)
5.3.4 基于语意的作业调度器关键技术 …… (70)
5.3.5 基于语意的任务调度器关键技术 …… (73)
5.3.6 任务跟踪器关键技术 …… (76)
5.4 SemanMR 计算部分框架 …… (78)
5.5 SemanMR 原理分析 …… (82)
5.5.1 SemanMR 原理实现分析 …… (82)
5.5.2 SemanMR 实现原理特点分析 …… (84)
5.6 基于 SemanMR 的大数据实时处理与分析实现技术 …… (88)
5.6.1 SemanMR 实时架构 …… (88)
5.6.2 SemanMR 的 MapReduce 网络优化技术 …… (89)
本章小结 …… (94)
第 6 章 大数据智能计算算法 …… (95)
6.1 大数据智能计算算法架构 …… (95)
6.2 数据采集算法 …… (95)
6.2.1 管理信息系统数据采集 …… (96)
6.2.2 网络信息数据采集 …… (96)
6.2.3 物理信息数据采集 …… (96)
6.3 数据预处理算法 …… (97)
6.4 数据挖掘算法 …… (99)
6.4.1 分类算法 …… (99)
6.4.2 聚类算法 …… (100)
6.4.3 关联挖掘算法 …… (101)
6.4.4 推荐算法 …… (101)
6.5 复杂智能算法 …… (103)
6.5.1 大数据溯源算法 …… (103)
6.5.2 大数据的相关推荐算法 …… (105)
6.5.3 基于大数据的决策管理算法 …… (105)
6.5.4 基于模型的推理及预测算法 …… (106)
6.5.5 基于数据的推理及预测算法 …… (107)
6.5.6 基于规则的推理及预测算法 …… (109)
6.5.7 混合推理及预测算法 …… (109)
本章小结 …… (109)
第 7 章 基于大数据的数据仓库技术 …… (110)
7.1 Facebook 中 Hive 采用的技术思路与存在问题分析 …… (110)

7.1.1 Hive 采用的技术思路分析 …… (110)
7.1.2 Hive 存在的问题分析 …… (111)
7.2 Yahoo!中 Pig 采用的技术思路与存在问题分析 …… (111)
7.2.1 Pig 采用的技术思路分析 …… (111)
7.2.2 Pig 存在的问题分析 …… (112)
7.3 未来数据仓库架构需求分析 …… (113)
7.4 一种基于大数据的数据仓库 SemanDW …… (114)
本章小结 …… (114)
第 8 章 大数据安全与隐私保护 …… (115)
8.1 大数据安全模型 BigData-PKI …… (115)
8.1.1 大数据安全体系结构 …… (115)
8.1.2 大数据安全模型 BigData-PKI …… (116)
8.2 大数据安全协议 BigData-Protocol …… (118)
8.3 大数据隐私 …… (120)
8.4 大数据的隐私提取方法 …… (121)
8.4.1 大数据的直接隐私提取方法 …… (121)
8.4.2 大数据的间接隐私提取方法 …… (121)
8.5 大数据隐私保护模型 BigData-Privacy …… (122)
8.6 大数据共享信息与隐私信息融合技术 …… (122)
8.6.1 大数据的共享信息与隐私信息融合机制 …… (123)
8.6.2 大数据的共享信息与隐私信息融合算法 …… (123)
8.6.3 大数据的共享信息与隐私信息融合质量评价模型 …… (123)
8.7 云环境下医疗大数据安全和隐私保护示范 …… (125)
8.7.1 云环境下大数据安全和隐私保护架构 …… (125)
8.7.2 数据分割及安全机制 …… (127)
8.7.3 数据融合及安全机制 …… (129)
8.7.4 基于隐私数据的查询机制 …… (130)
8.7.5 数据完整性保障机制 …… (131)
8.8 海量电子病历安全保护应用 …… (133)
本章小结 …… (134)
第 9 章 基于大数据的语意软件工程方法 …… (135)
9.1 基于大数据的语意软件工程体系架构 …… (136)
9.2 基于大数据的语意软件编制 …… (136)
9.2.1 基于大数据的语意软件编制方法 …… (136)
9.2.2 基于大数据的语意软件编制方法设计思路 …… (137)
9.2.3 复杂的 SemanPL 程序编程实现原理分析 …… (138)
9.2.4 基于大数据的语意编程语言 SemanPL …… (139)
9.2.5 SemanPL 编译器原理分析 …… (141)

9.3 基于大数据的语意软件测试 …… (143)
9.4 基于大数据的语意软件验证 …… (143)
9.5 基于大数据的语意软件工程方法的语意软件系统应用 …… (144)
本章小结 …… (144)
第二部分 基于海量语意规则的大数据流处理技术 …… (145)
第 10 章 基于规则的大数据流处理介绍 …… (147)
10.1 基于规则的大数据流 …… (147)
10.1.1 基于规则的大数据流应用背景 …… (147)
10.1.2 基于规则的大数据流应用意义 …… (148)
10.2 大数据流的规则处理技术国内外研究现状 …… (149)
10.3 存在的问题总结与分析 …… (153)
本章小结 …… (154)
第 11 章 语意规则描述模型 …… (155)
11.1 规则表示方法 …… (155)
11.2 规则节点图形化符号表示模型 …… (155)
11.2.1 非计算规则节点 …… (156)
11.2.2 计算规则节点 …… (156)
11.3 规则粒度 …… (158)
11.4 规则节点流量分析 …… (159)
11.5 计算规则节点计算代价分析 …… (163)
本章小结 …… (167)
第 12 章 海量语意规则网及优化 …… (168)
12.1 海量语意规则网概述 …… (168)
12.2 海量语意规则网维护 …… (169)
12.2.1 海量语意规则网增量集成 …… (169)
12.2.2 删除规则节点时的规则网维护 …… (170)
12.3 海量语意规则网优化方法 …… (171)
12.3.1 基于规则合并的优化方法 …… (171)
12.3.2 规则模块等价变换的优化方法 …… (173)
本章小结 …… (183)
第 13 章 海量语意规则处理算法 …… (184)
13.1 传统规则处理算法存在的问题 …… (184)
13.2 海量语意规则模式匹配模型 …… (185)
13.2.1 海量语意规则模式匹配模型体系结构 …… (185)
13.2.2 概念与介绍 …… (186)
13.2.3 模式网络存储组织 …… (186)
13.2.4 海量语意规则模式匹配算法 …… (188)
13.3 海量语意规则模式匹配算法特点 …… (195)

13.4 海量语意规则网运行处理机制 …… (195)
本章小结 …… (198)
第 14 章 海量语意规则并行处理 …… (199)
14.1 海量语意规则并行处理面临的问题 …… (199)
14.2 海量语意规则并行处理机制 …… (200)
14.2.1 海量语意规则并行处理机制 GAPCM 概述 …… (200)
14.2.2 海量语意规则子网生成 …… (201)
14.2.3 海量语意规则网计算代价预分配 …… (202)
14.2.4 海量语意规则网通信 …… (219)
14.2.5 映射分配 …… (220)
本章小结 …… (221)
第三部分 大数据应用 …… (223)
第 15 章 文化大数据 …… (224)
15.1 文化大数据的意义 …… (224)
15.2 文化大数据关键技术平台架构 …… (225)
15.3 文化大数据资源层 …… (226)
15.4 文化大数据综合平台层 …… (227)
15.5 基于文化大数据的应用 …… (228)
15.6 文化大数据云管理系统 …… (232)
本章小结 …… (234)
第 16 章 医疗健康大数据 …… (235)
16.1 医疗健康大数据 …… (235)
16.2 医疗健康大数据平台架构 …… (235)
16.3 医疗健康大数据共享平台 …… (237)
16.3.1 集中式医疗健康大数据共享平台 …… (237)
16.3.2 分散式医疗健康大数据共享平台 …… (238)
16.4 医疗健康大数据分散式架构资源集成方法 …… (239)
16.5 医疗健康大数据数据安全保护机制 …… (241)
16.6 医疗健康大数据隐私保护机制 …… (241)
16.7 医疗健康大数据挖掘与分析 …… (242)
16.8 基于可穿戴设备的居家医疗养老大数据分析系统 …… (243)
16.9 医疗健康大数据其他典型应用 …… (244)
本章小结 …… (245)
第 17 章 互联网金融大数据 …… (246)
17.1 互联网金融 …… (246)
17.1.1 互联网金融的概念 …… (246)
17.1.2 互联网金融的产生 …… (246)
17.1.3 互联网金融分类 …… (247)

17.1.4 互联网金融发展历程 …… (248)
17.1.5 互联网金融发展阶段 …… (251)
17.1.6 互联网金融发展趋势 …… (252)
17.2 大数据金融 …… (253)
17.3 金融大数据架构 …… (254)
17.3.1 金融大数据数据源 …… (255)
17.3.2 数据采集/清洗/转换 …… (255)
17.3.3 金融大数据存储 …… (255)
17.3.4 各种金融模型 …… (256)
17.3.5 各种大数据挖掘分析算法 …… (257)
17.3.6 各种大数据并行编程模型 …… (257)
17.3.7 各种大数据金融应用 …… (257)
17.4 大数据金融案例 …… (257)
本章小结 …… (258)
第 18 章 其他典型大数据 …… (259)
18.1 教育大数据 …… (259)
18.1.1 教育大数据平台架构 …… (259)
18.1.2 基于大数据的教育社区学生/教师个性化服务 …… (261)
18.1.3 基于大数据的教育社区学生行为建模与分析 …… (262)
18.1.4 基于大数据的教育社区教学规律分析 …… (262)
18.1.5 基于大数据的教育社区个性化教学 …… (262)
18.1.6 基于教育大数据的语意问答系统 …… (262)
18.2 电子商务大数据 …… (263)
18.2.1 电子商务大数据平台架构 …… (263)
18.2.2 电子商务虚假图片监测 …… (265)
18.2.3 电子商务产品个性化推荐 …… (265)
18.2.4 基于电子商务大数据的消费者行为分析 …… (266)
18.2.5 基于电子商务大数据的物流 …… (266)
18.2.6 电子商务实时大数据流规则处理 …… (266)
18.2.7 电子商务评估管理系统 …… (267)
18.3 互联网大数据 …… (267)
18.3.1 互联网大数据平台架构 …… (267)
18.3.2 互联网热点计算 …… (268)
18.3.3 互联网热点个性化推荐 …… (268)
18.3.4 互联网舆情监测 …… (268)
18.3.5 互联网热点趋势分析预测 …… (269)
18.3.6 互联网舆情预警应用 …… (269)
18.3.7 大型网络软件平台的数据采集与分析方案 …… (269)

18.4 能源大数据 …… (272)
18.4.1 石油大数据 …… (272)
18.4.2 智能电网大数据 …… (275)
18.5 交通大数据 …… (276)
18.6 宏观经济大数据 …… (278)
18.7 进出口食品安全监管大数据 …… (280)
18.7.1 基于大数据的进出口食品安全监管系统总体架构 …… (280)
18.7.2 基于大数据的进出口食品安全监测分析 …… (280)
18.7.3 基于海量语意规则的进出口食品社会应急分析 …… (281)
18.7.4 基于大数据的进出口食品溯源分析 …… (282)
18.7.5 基于大数据的进出口食品安全决策 …… (283)
本章小结 …… (283)
第 19 章 基于大数据的语意计算及典型应用 …… (284)
19.1 基于大数据的应用领域分析 …… (284)
19.1.1 基于大数据的社交网络领域应用分析 …… (284)
19.1.2 基于大数据的医疗领域应用分析 …… (285)
19.1.3 基于大数据的政府领域应用分析 …… (287)
19.1.4 基于大数据的金融领域应用分析 …… (289)
19.1.5 基于大数据的企业计算应用分析 …… (290)
19.2 语意搜索引擎 …… (291)
19.2.1 传统搜索引擎 …… (292)
19.2.2 语义搜索引擎（Semantic Search Engine） …… (293)
19.2.3 语意搜索引擎（Semantic+ Search Engine） …… (293)
19.2.4 语意搜索引擎（Semantic++ Search Engine） …… (295)
19.3 语意金融 …… (296)
19.4 语意旅游 …… (296)
19.5 语意电子商务 …… (297)
19.5.1 案例概述 …… (297)
19.5.2 校园社区网规则举例 …… (298)
19.5.3 优化的带流量的规则网 …… (302)
19.5.4 未经优化的带流量的规则网优化 …… (302)
19.5.5 规则网络代价计算 …… (305)
19.5.6 规则网络任务划分 …… (306)
19.5.7 规则子网划分 …… (308)
本章小结 …… (310)
第 20 章 大数据未来研究方向 …… (311)
参考文献 …… (315)

第一部分　大数据基础理论分析

本部分主要介绍了大数据领域的主要基础理论，包括大数据基本概念、可编程数据中心、云文件系统、云数据库系统、大数据并行编程与分析模型、大数据智能计算算法、基于大数据的数据仓库技术、大数据安全与隐私保护及基于大数据的语意软件工程方法等9章。

本部分主要分析了大数据在存储、管理、处理与分析等方面面临的一些基础的理论问题，例如，如何确保大数据存储中心能够实现智能调度，最优地节省能源；如何确保大数据有一个比较好的放置方法，为大数据的计算提升效率；如何有效实现大数据的存储，从而提升大数据的智能读写效率；如何有效地实现大数据的语意处理，增强其计算能力；如何确保大数据存储过程中文件的安全可靠性，尤其是重要数据的内容安全；如何确保大数据中重要文件的数量方面的安全；为了实现大数据在大规模范围内的共享，如何确保在云环境下的各种大数据的隐私能够得到有效保护；基于大数据的各种应用密集型的大型软件系统如何开发，其软件工程开发方法如何，等等。

第1章　大数据基本概念

大数据已经在社会政治、经济、文化、教育、能源、交通、军事、互联网及科技等各个方面产生了巨大的推动作用，一个企业甚至一个国家拥有大数据资源的规模和存储、管理并使用这些大数据进行各种应用分析并解决问题的能力，已经成为检验一个企业或者国家竞争力的一个重要标志，也是它们是否具有核心竞争力的关键所在。目前，全球很多国家已经将大数据作为一项国家意志进行了确定。首先，本章分析了大数据的一些基本概念，其次，介绍了语意计算（Semantic++ Computing）与大数据的关系，在大数据技术的背景下，语义计算（Semantic Computing）提升到语意计算（Semantic++ Computing）层次已经成为可能，最后，概述了大数据和云计算之间的关系。

1.1　大数据定义

当前，关于大数据的定义已出现多个说法，并没有形成统一的定论，因此，不同的人、公司、机构等可以从不同的角度对大数据进行定义。其中，IBM 公司的观点在一定程度上具有代表性。

IBM 公司关于大数据的表述为 3V 理论，即容量（Volume）、类型（Variety）和速度（Velocity）。其中，容量是指数据的数量和所占用的存储空间十分庞大，需要处理来自服务器、手机、移动设备、传感器、社会媒体等的数据量达到 PB（Petabyte）级别、EB（Exabyte）级别甚至 ZB（Zettabyte）级别。类型是指数据类型各种各样，有结构化的数据、半结构化的数据、非结构化的数据、多媒体数据、文本数据、数据库数据，等等，所有类型的数据都是构成大数据的来源，它们结构各异，都是异构型的数据。速度是指对大数据处理速度的要求越来越高，尤其是对于一些需要实时计算的应用，快速处理数量如此庞大的数据已显得十分困难。对大数据的处理不仅仅是对静态数据进行处理，也需要处理源源不断到达的动态数据，例如，物联网数据流、各种社交媒体产生的信息流等。同时，对大数据的处理需要在瞬间得到结果，对速度的要求达到了实时性的层次。

然而，随着对大数据的理解越来越深入，对大数据计算的需求越来越高，研究越来越广泛，现有的 3V 理论已经很难概括大数据的本质。因此，IBM 在 3V 理论基础上又增加一个新的 V，这个 V 被称为准确度（Veracity），即形成了 IBM 所说的 4V 理论，也就是现在 IBM 对大数据的基本理解[1]。从本质上来说，准确度其实是指大数据的可信性或者其价值所在。可信的大数据才能够真正创造价值，如果大数据不安全或者不可靠，那么无论用什么方式计算得出的大数据决策方案都没有任何价值。因此，准确度（Veracity）在大数据的地位日渐提升，基于此，确保大数据的安全和实施大数据隐私保护就显得十分重要。