贵州大数据发展与建设

张 梅 文静华 刘 振 著

科学出版社

贵州大数据发展与建设

张　梅　文静华　刘　振　著

科学出版社

北　京

内 容 简 介

本书首先系统介绍贵州省和贵安新区大数据发展和建设的理论与技术基础，然后结合贵州实际，深入阐述贵州省和贵安新区大数据发展的典型案例，最后对贵州省大数据发展进行总结与展望。

本书可供企业（机构）信息化管理部门、各类大数据建设与管理人员等参考，也可作为高等院校大数据、信息、管理、经济等相关专业教师和学生的参考书。

图书在版编目（CIP）数据

贵州大数据发展与建设/张梅，文静华，刘振著. —北京：科学出版社，2019.11

ISBN 978-7-03-062563-2

Ⅰ. ①贵… Ⅱ. ①张… ②文… ③刘… Ⅲ. ①数据管理-研究-贵州 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 217621 号

责任编辑：王 哲/责任校对：彭珍珍

责任印制：吴兆东/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

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2019 年 11 月第 一 版 开本：720×1000 B5

2019 年 11 月第一次印刷 印张：8 1/2

字数：150 000

定价：68.00 元

（如有印装质量问题，我社负责调换）

此为试读，需要完整PDF请访问：www.ertongbook.com

前　言

大数据是支撑信息化发展的重要基础，大力发展大数据，无论对信息通信技术行业的发展，还是对于经济社会的转型升级，以及国家政策的落实等方面都具有重要的战略意义。我国南方具有得天独厚的自然生态和气候环境，具备建设和发展大数据的良好条件，贵州省委、省政府优先发展大数据产业的支持政策和产业界的高度认同，使得这一地区成为大数据的集聚地和先行区，引领全国大数据从传统到绿色、从单用到复用、从分散到集中的转变。因此，研究和探讨大数据发展及其在贵州发展建设中的实践具有重要价值和意义。本书将为大数据的发展提供系统性的原理、技术方法与详细的操作指导，并在贵州大数据发展过程中进行创新性探索，对于促进和推广大数据的发展具有重要的参考价值和实践指导意义。

本书根据贵州省和贵安新区大数据发展现状，利用贵州省政府大数据应用典型示范项目、贵安大数据发展典型案例，探讨贵州大数据发展与建设的关键技术，以期帮助贵州大数据更好、更快地发展与建设。本书比较系统地研究了大数据的理论与技术基础、贵州大数据产业发展战略、贵安新区大数据产业发展战略、贵安新区大数据产业布局等关键问题。

目前，对大数据的发展与建设，处于全方位探索阶段，缺乏系统性的理论方法与具体的操作流程。本书依据相关理论及作者多年的实践经验，首先对贵州大数据发展战略和现状、贵州省电子政务云建设实践、贵安新区大数据产业发展任务、贵安大数据发展重点工程等进行了研究；然后，从贵州大数据产业发展战略、贵州省政府大数据应用典型示范项目建设、贵安新区大数据发展典型案例等方面对贵州和贵安新区大数据发展与建设进行了详细分析与深入研究。

本书是作者在从事多年信息化和大数据关键技术研究以及对本科生、研究生教学实践的基础上编写的，内容不但包括一些基础知识，而且较系统地探究

了近年来贵州省大数据发展与建设研究的重要成果。作者在国家级核心期刊和重要学术会议 IEEE、ACM 上发表了相关研究论文 40 余篇，其中 SCI、EI、ISTP 收录 20 余篇，为本书的撰写奠定了一定的理论基础。本书第 3 章由文静华编写，第 4 章由刘振编写，其余各章由张梅编写。

本书的出版得到了贵州财经大学 2018 年度预算经费的资助，在研究工作中还获得了其他基金的资助，包括 2016 年度贵州省科技厅科技基金（项目编号：黔科合基础[2016]1020）和 2016 年度贵州省教育厅自然科学拔尖人才基金（项目编号：黔教合 KY 字[2016]069）。感谢贵安新区管理委员会欧阳武主任，他以深厚的学术功底、超前的产业视角、严谨的治学态度使作者在其指导下受益匪浅。感谢贵安新区大数据产业发展领导小组办公室的肖凌青、吴勇等同志，他们在本书撰写过程中一直给予帮助、鼓励和支持，使得本书的写作得以顺利完成。

感谢冯飞、张志龙、李玉、邢伟琛、程珊、孟丹、王佳、苏慧慧、刘彤等学生，本书介绍的许多工作是作者与他们合作完成的。

大数据发展与建设研究的内容非常宽广，与其相关的学科也很多，由于作者学识有限，书中难免存在不足之处，恳请读者批评指正。

张　梅　文静华　刘　振

2019 年 8 月

目　录

第1章 绪　论

如今，提到大数据（Big Data）几乎是无人不知，其已成为一项业务上优先考虑的工作任务。大数据正在改变人们的生活以及理解世界的方式，而更多的改变正蓄势待发。大数据的应用范围非常广泛，与大数据相关的诸多问题引起了专家和学者的高度重视。本章从大数据的概念出发，阐述大数据的 5V 特点，深入研究大数据的主流技术，分析大数据应用场景，以期为大数据的研究和应用提供有益参考。

1.1 大数据概念

随着信息技术的迅猛发展，“大数据”从网络热点词到今天的大规模广泛运用，仅用了不到 10 年时间。人们也从惊叹“大数据时代”来了，到置身其中且逐步习惯大数据对日常生活各方面的影响。随着计算机技术的发展和互联网应用的普及，数据的产生方式和产生量也发生着质的变化，大数据概念应运而生。

从一般意义上讲，大数据是指传统数据获取、存储、处理、分析和管理技术无法解决的既大又复杂的数据集合。大数据作为一门新兴技术在不同领域产生着巨大作用。一方面，它隐藏着巨大的商业价值。通过对大数据的获取、存储管理以及分析挖掘，各行各业都可以应用大数据技术管理企业、营销产品以及做出决策等。另一方面，政府层面可以依托大数据技术提高管理能力，改变管理方式。比如，通过大数据分析的结果进行舆情监控、犯罪预测，或者通过大数据平台构建全方位立体化的电子政务体系，方便政府和社会群众的互相交流。

大数据是一个很抽象的概念，很多非专业人士提到大数据，也只能从数据

量上去感知大数据的规模。例如，百度公司每天大约要处理几十 PB 的数据；Facebook 每天生成 300TB 以上的日志数据；根据著名咨询国际数据公司（International Data Corporation，IDC）的统计，2011 年全世界被创建和复制的数据总量为 1.8ZB（10^{21}），但仅从数据量上并不能区别大数据和传统的海量数据。

2008 年 9 月，在 *Nature* 发表的 *Big data: science in the petabyte era* 一文[1]中，大数据被定义为“代表着人类认知过程的进步，数据集的规模是无法在可容忍的时间内用目前的技术、方法和理论去获取、管理、处理的数据”。由此，“大数据”便开始在学术界迅速扩散传播，成为学术界和产业界的研究热点。

大数据并非一个确切的概念。起初，这个概念是指等待处理的信息总量过大，已经超出了日常生活中使用的一般的电子计算机处理数据的能力，即数据的大小超出了电子计算机运作时可以使用的内存量，因此工程人员必须改进处理数据的工具[2]。现在大数据被普遍认为除了用来表述数据的规模巨大，而且指需要借助先进的计算技术支持以完成有效数据挖掘和推算新数据的计算过程，从信息技术发展的意义上说，大数据是目前人类已经掌握的信息技术应用的最高形态。大数据不仅仅使人们基于数据创造出新的价值，也改变了现代商品市场的组织结构，改进了政府的治理方式[3]。

无所不在的各类传感器将产生越来越多的数据，数据量级将从现在的 GB、TB 级逐步增长到 PB、EB 和 ZB 级。迫切需要通过分析这些结构复杂、数量庞大的数据，以云端运算整合分析，快速地将其转化成有价值的信息，从中探索和挖掘自然和社会的变化规律。利用大规模有效数据分析预测建模、可视化和发现新规律的时代已经到来。

对于“大数据”，研究机构 Gartner 给出了这样的定义：“大数据”指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新的处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产[4]。大数据就是“未来的新石油”。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。在《大数据时代：生活、工作与思维的大变革》[2]中，大数据指不用随机分析法（抽样调查）这种捷径，而采

此为试读，需要完整PDF请访问：www.ertongbook.com

用所有数据进行分析处理的新的数据处理方式。

1.2 大数据特点

现在普遍认为大数据特性为5V[5-8]，即大量（Volume）、高速（Velocity）、多样（Variety）、低价值密度（Value）、真实性（Veracity）。

（1）数据容量巨大作为大数据的头号特征，不但描述了大数据的规模，也从另一个角度说明了大数据的发展态势，大数据只会增加，不会减少。据IDC统计预测，到2020年全球数据总量预计会达到40ZB，相当于平均每人拥有5247GB数据。具体到我国，由相关研究报告显示，目前我国的数据总量正在以每年平均50%的速度保持持续增长，在这种前提下，预计到2020年，我国数据总量在全球的占比将达到21%，而这一比例在2011年仅为13%。

（2）处理速度超快体现了大数据极高的数据生长率和转换率。大数据时代每天都有海量的数据更新，因此对数据的处理速度有很高的要求，需要响应以秒甚至毫秒计的流数据。我国的数据总量正在以每年平均50%的速度保持持续快速增长，处理这样体量的数据，需要极快的处理速度，这也与我国超级计算机技术的发展相辅相成。

（3）数据类型众多作为大数据区别普通数据的一个重要特征，从数据结构和类型上判断，大数据不仅拥有众多的数据类型，还存在着结构与非结构化数据。相对于过去以文字信息为主的结构化数据，大数据时代的数据类型明显增多，除了文本信息，还包含了图片信息、音频信息、视频信息、地理位置信息等半结构化和非结构化数据。如何在一个系统平台中处理多种类型的数据，是大数据的核心挑战之一。

（4）较低价值密度并不是指数据的价值高，反而是数据价值低，甚至没有价值。正是由于大数据的价值密度低，因此，要挖掘出大数据潜藏的价值，犹如大浪淘沙。例如，股票市场就是典型的应用大数据的市场，在对股票市场的全天交易信息的监测中，只有很少一部分可以用来分析某一只股票的市场表现。在浩如烟海的数据中，有效数据变得更加碎片化，不再是集中的大块信息，而是分散的零碎的信息群。大数据通常都是自动采集的，天然具有噪声，如何在

有噪声的情况下还能被有效地运用？这不是传统的查询操作能够完成的，需要发展更复杂的数据治理、数据分析和机器学习技术。

（5）数据真实可靠是指数据的准确性和可信赖度，即数据的质量。数据的有效性和真实性依赖于数据的质量，高效地对数据和数据中的知识进行评估对此至关重要，质量较好的数据对后期提取大知识和做出个性化服务具有重要意义，高质量的数据和知识也能够体现大数据的价值。

1.3　大数据主流技术

在大数据时代，对大数据进行统一表示，实现大数据处理、查询、分析和可视化是亟须解决的关键问题。互联网点击数据、传感数据、日志文件、具有丰富地理空间信息的移动数据和涉及网络的各类评论，成为了海量信息的多种形式。海量的电子政务数据、移动终端数据、网站日志、社交媒体数据、来自物联网传感器的流式数据、企业长期积累的业务数据等也都是大数据的主要来源。现有面向大数据的研究主要针对存储、处理、分析、可视化等某一方面的关键技术。当今主流的大数据技术如图 1.1 所示。

1.3.1　分布式计算存储技术

分布式计算存储平台基于 PC Server X86 服务器集群部署，提供分布式数据存储、分布式计算框架，同时整个生态圈提供了大量外围组件满足各类应用场景需求，主要采用 Hadoop 和 Spark 技术[9-12]。

Hadoop 基于 Java 语言开发，以分布式文件系统和 MapReduce 为核心，具有如下特点。

1. 可扩展性

Hadoop 运行在基于 X86 结构的普通 PC 服务器或刀片服务器上，硬件和软件松耦合在一起，可以很方便地增加计算节点。

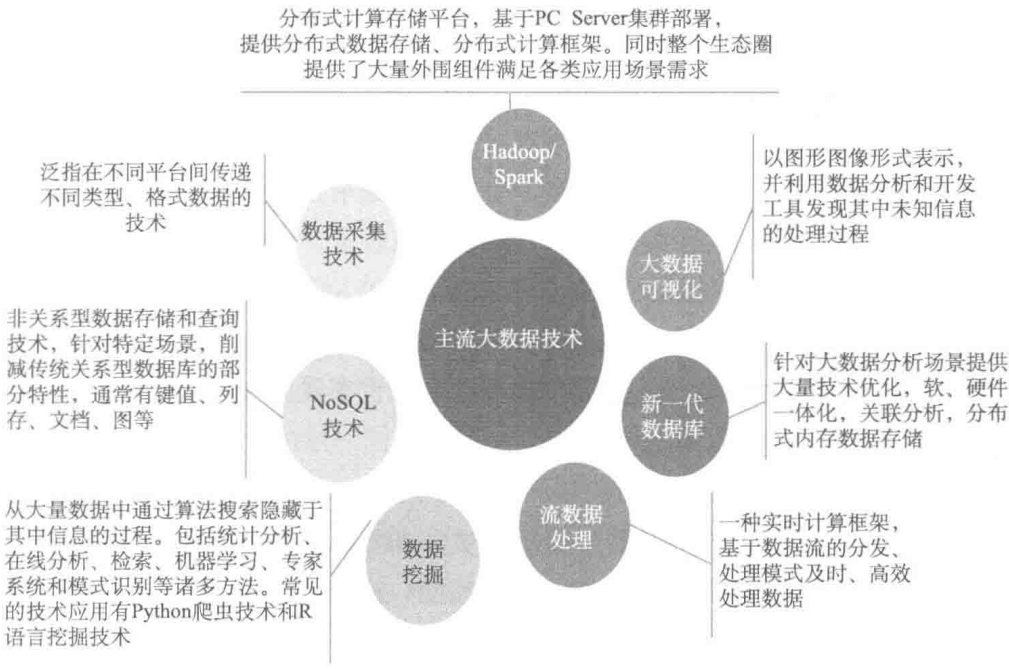

图 1.1 大数据主流技术

2. 可靠性

Hadoop 能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配，确保能够针对失败的节点重新分布计算。

3. 低成本

Hadoop 架构在廉价的硬件服务器上，不需要昂贵的硬件做支撑。其软件是开源产品，不需要授权费用。

4. 高效性

相比传统并行计算结构，Hadoop 的计算和存储是一体的，实现任务之间无共享，I/O 开销小。

Hadoop 是目前大数据平台中应用率最高的技术，特别是针对诸如文本、社交媒体以及视频等非结构化数据。Hadoop 可以部署在价格低廉的服务器上，形成分布式系统，它提供高吞吐量来访问应用程序的数据，适合那些有着超大数

据集的应用程序。

Hadoop 的核心是 HDFS（Hadoop Distributed File System）和 MapReduce。HDFS 具有高容错性和高扩展性等优点。MapReduce 分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序。因此，通过 Hadoop 可以轻松地组织计算机资源，搭建自己的分布式计算平台，完成海量数据的处理。相对当前应用较多的 SQL 关系型数据库，HDFS 提供了一种通用的数据处理技术，它用大量低端服务器代替大型单机服务器，用键值对代替关系表，用函数式编程代替声明式查询，用离线批量处理代替在线处理，以高容错的方式并行处理大量的数据集。

Spark 拥有 MapReduce 的优点，但不同于 MapReduce 的 Job 中间输出，其结果可以保存在内存中，从而不再需要读写 HDFS。Spark 是新兴的分布式计算平台，使用基于内存的计算框架以提升性能，适用于迭代计算场景，同时提供应用工具降低使用复杂度。其有以下特点。

（1）速度快。Spark 支持内存计算，对于小数据集能达到亚秒级的延迟。

（2）易于使用。Spark 支持 Scala、Java 和 Python 编写程序，提供了超过 80 个高级运算符，以便于更容易地构建并行应用程序。

（3）与 HDFS 底层兼容。Spark 能够运行在 Hadoop 2.x 的 YARN 集群管理器上，并且能够读取任何存在 Hadoop 数据。

此外，Spark 提供了拥有完善的容错机制，具备准实时性和高吞吐量的大规模弹性可扩展的流处理框架 Spark Streaming[13,14]。Spark 中的 HBase/Hive 能支持海量网络流量的存储查询。因此很适合用 Spark 来构建大型、低延迟的交互式数据分析型运用。

1.3.2　大数据可视化技术

大数据可视化技术是以图形图像形式表示大数据，并利用数据分析和开发工具发现其中未知信息的处理过程。大数据可视化技术已经涵盖科技和生活的各个方面，现有的大数据可视化涉及自然科学现象及计算领域、计算机网络、政治商业金融、工程管理及艺术表现学等众多领域。目前大数据可视化研究的重点是如何将复杂多维的数据进行图形表征，包括抽象的、具象隐喻的或是仿

真的表征方法以及优化算法[15]。一般来说，大数据可视化技术包括文本可视化技术、多维数据可视化技术、网络可视化技术、时空可视化技术等。

大数据可视化中经典的二维可视化方法是热力图，热力图作为一种直观的可视化方法，具有综合展示数据地理空间特征和属性特征的良好特性，可帮助各个领域的研究人员获取地理空间知识[16]，因此深受欢迎。赵婷等使用微软内部发布的 Heat Map 并结合 K-means 聚类算法针对地理标签数据的可视化表达进行了研究[17]。Spakov 和 Miniotas 通过数据聚集热图可视化技术根据人眼凝视调整热图的透明度，实现了轨迹实验验证和产品的可用性研究[18]。Bojko 提出针对不同的需求选择不同的热力图可视化技术，并在轨迹可视化方面符合人类视觉分辨要求[19]。

三维可视化可以更加精细地表现数据特征，交互界面较受用户青睐，可视化效果也更为清晰明了。李新维等提出了基于四叉树结构的大规模倾斜摄影模型三维可视化方法，利用开源 OSG 库将倾斜摄影模型重建后进行场景可视化[20]。Kwan 等通过三维 GIS 可视化技术来研究人类活动模式，就居民的家庭分布、职业、种族和出行等方面进行了研究。

大数据可视化技术作为一种可以有效地简化与提炼数据流，将海量复杂的数据直观可视化呈现的工具逐步发展起来。知识图谱是其中之一，它是以科学知识为研究对象，描述科学知识的发展进程与结构关系的一种图形[21]。科学知识图谱涉及数理统计学、计算机科学、社会学、信息科学、图像学等多学科的理论，并且与科学计量学的共词分析、共引分析等方法结合，通过对科学知识的挖掘和处理，绘制一系列可视化的图形，将学科知识发展进程和结构关系直观形象地展示出来[22]。这种将数据通过可视化技术变成直观图形的方法不仅让冰冷枯燥的数据变得亲切和易于理解，更是激发了人的形象思维与想象力，从而为科学新发现创造新的手段和条件[23]。

目前，大数据可视化工具主要包括：开源的、可编程的工具，如 R 语言、D3.js、Leaflet、Python、Processing.js 等；商业化软件工具，如 Tableau、Qlikview、SAS、水晶易表、IBM Cognos 等。

1.3.3　新一代数据库

在大数据推动行业发展的时代，为了满足不同业务需求，大型企业级应用往往选择多种数据库产品，这种组合式解决方案需要精细的控制数据流转和一致性，使用难度颇高，系统间的数据同步和冗余带来了很高的成本开销，限制了企业级应用的发展。

针对以上问题，Gartner 在 2014 年提出融合数据库技术 HTAP（Hybrid Transactional/Analytical Processing）架构，即一个数据库既支持在线事务处理（OLTP），又支持在线分析处理（OLAP），能满足所有数据模型的需求，能处理所有工作负载（OLTP 和 OLAP），支持高并发、高吞吐量事务和分析混合的任务流。

新一代非关系型数据库有以下 5 个主要类型。

（1）面向文件存储：适用于存储海量文件，代表产品为 MongoDB。

（2）列存储（Wide Column Store/Column-Family）数据库：快速查找相关数据，相关数据被放在同一列中，代表产品为 Cassandra。

（3）搜索引擎：适用于存储文件索引，代表产品为 Solr。

（4）键值（Key-Value）数据库：快速访问非相关数据。可以通过 Key 来添加、查询或删除数据，代表产品为 Redis。

（5）图（Graph）数据库：访问以图片方式存储的数据，如社交网络，代表产品为 Neo4j。

MPP 数据库是基于 X86 服务器集群部署的并行关系型数据库，针对分析型使用场景提供了大量技术优化，以充分发挥关系数据库的数据关联分析能力和 MPP 架构的性能优势。适合于深度分析与挖掘、即席查询与自助分析等应用场景。内存数据库通过将数据存储在内存中以提高数据库性能，同时通过额外的数据保护机制保障内存数据库安全。适合数据访问提速、为流处理提供数据存储和查询、系统间或组件间数据传递平台（高速访问）。

Apache Ignite 是新一代数据库缓存系统，将数据存储在缓存中能够显著地提高应用的速度，因为缓存能够降低数据在应用和数据库中的传输频率。Apache Ignite 允许用户将常用的热数据储存在内存中，支持分片和复制两种方

式，使开发者可以均匀地将数据分布式到整个集群的主机上。同时，Ignite 还支撑任何底层存储平台，不管是 RDBMS、NoSQL，又或是 HDFS。在集群配置好之后，数据集增加只需在 Ignite 集群中增加节点而不需要重启整个集群，节点数目可以无限增加。

1.3.4 流数据处理技术

大数据变化快这一特征具体体现在数据实时到达、规模庞大、大小无法提前预知，并且数据一经处理，除非进行存储，否则很难再次获取。在金融应用、网络监控、社交媒体等诸多行业领域，都会产生这类变化极快的数据。为了解决这一问题，人们提出了流数据处理技术[24]。流处理技术是针对实时性非常强的流式数据进行处理的利器，适用于基于多元化数据采集、实时分析处理、复杂规则匹配等过程的实时营销、监控类场景。

流处理系统通过同一时间对系统传输的每一条数据项或微批处理数据操作，实现对数据的实时处理和显示，适合用来处理关注一段时间变化并对变动或峰值做出响应趋势的数据。目前典型的流式大数据处理方案如下。

（1）Spark Streaming[25]是用于处理流式数据的组件，数据流以时间片（秒级）为单位进行拆分，以批次的方式处理每个时间片数据，因此相比于 Storm 存在延迟高、吞吐量较小等缺点。Spark Streaming 是构建在 Spark 基础上的流式大数据处理框架，同时支持多种数据输入源和输出格式。

（2）Yahoo S4（Simple Scalable Streaming System）[26]是一个分布式流处理引擎，具有通用性、可扩展性、容错性等特点。开发者能够在 S4 引擎基础上开发面向无限的、不间断的流数据处理应用。S4 的不足之处在于数据传输可靠性差，可能丢失数据，同时由于数据暂存在内存中，一旦节点出现故障，节点上的数据就会丢失。

（3）Samza[26]是一款开源的、分布式的、基于 Hadoop 架构的流处理系统。Samza 对 Kafka 的依赖是一种限制，批处理的每个计算之间对 HDFS 的依赖导致了一些严重的性能问题，语言支持方面不如 Storm 灵活。Samza 的工作模式像 MapReduce 的过程，使用 YARN 进行资源分配和任务调度，如图 1.2 所示。

此为试读，需要完整PDF请访问：www.ertongbook.com