大数据智慧管理与分析之技术和实践

——从数据仓库/OLAP到NoSQL和NewSQL

朱 焱 编著

西南交通大学出版社

大数据智慧管理与分析之技术和实践

——从数据仓库/OLAP 到 NoSQL 和 NewSQL

朱　焱　编著

西南交通大学出版社
·成　都·

图书在版编目（CIP）数据

大数据智慧管理与分析之技术和实践：从数据仓库/OLAP到NoSQL和NewSQL/朱焱编著. —成都：西南交通大学出版社，2019.8

ISBN 978-7-5643-7076-3

Ⅰ.①大… Ⅱ.①朱… Ⅲ.①数据库系统 Ⅳ.①TP311.13

中国版本图书馆CIP数据核字（2019）第179111号

Dashuju Zhihui Guanli yu Fenxi zhi Jishu he Shijian

大数据智慧管理与分析之技术和实践

——从数据仓库/OLAP到NoSQL和NewSQL

朱焱 编者

责任编辑	李华宇
封面设计	何东琳设计工作室
出版发行	西南交通大学出版社 （四川省成都市金牛区二环路北一段111号 西南交通大学创新大厦21楼）
发行部电话	028-87600564 028-87600533
邮政编码	610031
网址	http://www.xnjdcbs.com
印刷	四川煤田地质制图印刷厂
成品尺寸	170 mm × 230 mm
印张	14.5
字数	242千
版次	2019年8月第1版
印次	2019年8月第1次
书号	ISBN 978-7-5643-7076-3
定价	68.00元

前 言

1. 编写的背景

首先，基于大数据的互联网电子商务、智能交通系统、大数据金融、智慧城市等渗透进我们生活的各个层面，已经从“概念”走向了“价值”和“应用”。为了理解、分析和研究大数据，有效地掌握和应用大数据关键技术，撰写一本讨论大数据智慧管理和分析的核心概念、技术要领、技术关联和实现机制的书是迫切需要的。

其次，人们希望学习和掌握如何在近真实的应用环境中运用关键技术，如何建设大数据管理与分析环境，研发大数据基础系统，从而从理论和实践两个方面更好地理解和使用大数据技术。从而能有效地解决实践中的问题，推动应用领域快速向前发展。因此，一本涵盖这些内容的书是喜闻乐见的。

最后，大数据覆盖了结构化、半结构化和非结构化数据类型，发展出了5V、6V甚至是7V特点。大数据技术从关系型数据库和数据仓库技术发展到现在的非关系数据管理技术，如NoSQL和NewSQL。新发展的NoSQL和NewSQL技术与关系数据管理与海量数据仓库技术有着相辅相成、继承再发展的关系。沿着这条“继承再发展”脉络讨论大数据智慧管理和分析是新颖的和十分必要的。

2. 本书的特点

编著本书历时4年，具有较大的挑战性：

其一，大数据技术的发展十分迅速，新的技术、技术组合和应用领域不断涌现，现有的技术不断更新和完善。著作内容应体现技术的发展和新颖性。

其二，大数据智慧管理和分析技术不仅继承了成熟丰富的结构化数据管理技术的优势，还继承了海量数据智慧管理与分析计算（数据仓库）的技术优势和先进思路，而且正站在这些“巨人”的肩膀上，逐步发展出能应对大数据新特点、新挑战、较全面和完善的技术体系。整个技术体系、

架构和机制具有相当的深度和广度。因此，著作内容应具有继承性。

其三，大数据与互联网、分布式架构、数据分析、故障恢复等技术紧密相关，技术间相互作用，技术集成产出功能十分复杂。因此，著作内容应涵盖技术的关联性。

其四，要理解、掌握和运用大数据智慧管理与分析技术，理论联系实际是关键途径。但目前阐述这些技术及其在应用中如何发挥作用的相关内容还较零散，不够具体和全面。因此，著作内容必须具有实践性。

作者力图通过翔实的内容、清晰的层次、重点突出的阐述、丰富的案例讨论、形象生动的图表，以及具体的应用实例来解释枯燥的技术原理，从而达到使读者易读、易懂、易使用的目标。

3. 要点、作用与适用面

作者在查阅、分析和研究文献资料的基础上，介绍了大数据智慧管理与分析技术的历史沿革和最新发展；总结了大数据的 5V 特点和技术挑战；重点讨论了大数据智慧管理与分析技术对关系数据模型、海量数据仓库机制与 OLAP 分析技术的继承与发展。本书聚焦数据仓库、Hive 和 HBase 三大技术，重点阐述了基本概念、数据模型、技术原理、实现平台环境；厘清并归纳总结了相关核心技术；分析了技术特点、技术之间的重要关联和技术集成的优势；深入讨论了三大技术在实践应用中的实现；针对海量数据管理、NoSQL 和 NewSQL 实践项目开发给予详细、有效、可行的指导。本书示范了如何应用数据仓库、Hive 和 HBase 三大技术完成不同的实践应用的全过程，从而突出了理论与实践相结合的特点。

希望读者通过阅读本书，能了解数据仓库、NewSQL、NoSQL 技术的发展和技术特点；能学习掌握数据仓库机制/OLAP 分析、Hive 和 HBase 等大数据管理和智慧分析的技术原理、功能、架构和实践应用；厘清各个关键技术的相互关系、优势与不足，了解技术集成与所解决的问题。特别是能在本书的指导下，掌握开发基于 Hive 和 HBase 大数据管理和分析的应用，提高数据管理、分析和应用的能力。

本书既可面向具有计算机科学、数据科学、IT 专业知识的技术人员、研发者和学习大数据知识与技术的研究生与本科学生，作为他们学习掌握大数据智慧管理和分析技术、践行理论联系实际的参考用书或教材；也适用于从事大数据工作的非专业领域人士，包括经济分析、交通大数据管理与分析、医疗健康大数据管理及分析等领域。本书给出了完整的应用示例作

为技术运用的参考，为上述人士，特别是非专业技术人员提供了较好的帮助。

作者采用全英语讲授的“数据仓库与数据挖掘”是一门备受学生欢迎的课程，是正在开展的“计算机科学与技术全英文研究生专业建设”的重要课程之一。大数据管理与分析也是本科教学内容的深化，是许多教学改革项目的支撑。作者通过10多年相关的科学研究、项目研发和教学，以及对学生进行硕士论文和课内外实践项目的指导，对大数据管理与分析有了较深的领悟，同时积累了知识和经验，践行了关键技术和项目的开发，从而为编著本书创造了很好的条件。

4. 本书的主要内容

本书由两大部分组成，第一部分是技术篇，重点介绍大数据发展的特点和面临的挑战，讨论了经典的海量数据智慧管理与分析技术——数据仓库和OLAP；介绍了大数据、分布式、并行计算环境下，数据智慧管理技术的新发展——NoSQL数据库和NewSQL数据仓库；重点阐述了这两大新技术的基本数据模型、系统架构和性能优势；重点对比了OldSQL（经典关系数据模型）与NoSQL数据库（HBase）、关系数据仓库与NewSQL数据仓库（Hive）的技术异同和优劣势；讨论了这些技术的组合集成方案、组合技术的优势以及实际应用案例。第二部分是实践应用篇，分别从关系数据仓库建设与OLAP分析、基于Hive的数据仓库技术与OLAP应用、HBase大数据管理技术实践三个方面进行了详细阐述；根据相关技术原理和应用需求，详细讨论了基于Hadoop的环境配置、大数据管理模型建立与实现；着重给出了应用三大关键技术构建大数据智慧管理与分析系统、开发实践项目的指导内容。全书采取由概念到技术、理论到实践的顺序编写。

第一部分技术篇：

第1章介绍了大数据的发展与影响，以及5V特点，简要讨论了大数据的应用，分析了对大数据技术的误解，简要阐述了大数据形态与关系数据模型在基本原理和管理机制上的异同。

第2章介绍了大数据的生命周期，分析了大数据面临的技术挑战和应对挑战的策略与方法。

第3章分析讨论了关系数据模型在大数据管理的局限性，阐述了面向大数据特点的技术革新，讨论了NoSQL中几个关键技术的原理和特点。

第4章重点介绍列式数据管理技术，讨论了HBase的数据模型和集群架构等，分析了HBase的作用与局限。

第 5 章首先讨论了数据仓库建模与 OLAP 分析技术，由此引出 NewSQL 数据仓库——Hive 技术。重点讨论了 Hive 的数据模型和系统结构，基于 Hive 的 OLAP 功能，分析了 Hive 与其他技术的比较与集成。

第 6 章阐述了新老技术的继承与发展关系，讨论了技术组合的作用，分析了相关案例。

第二部分实践应用篇：

第 7 章重点讨论了面向大数据的数据仓库建模和实现，示范了一个基于 ROALP 数据仓库技术的海量数据管理应用与基于 OLAP 技术的数据分析。

第 8 章重点讨论了 Hive 技术及其实践项目开发。针对大数据 5V 特点，设计与实现了一个基于 Hive 数据仓库技术的大数据管理与分析示范系统。该系统包括基于 Hadoop 平台的应用环境搭建、参数设置、基于 HQL 的 OLAP 分析计算，以及分析结果可视化展示。

第 9 章着力于讨论 NoSQL 家族中的 HBase 技术如何与实践相结合，设计与实现了一个基于 HBase 列式数据库技术的大数据管理示范系统。该系统包括分布式 Hadoop 平台的搭建和组件配置、HBase 数据库设计与实现、数据访问功能与结果可视化展示。

5. 致　谢

本书的编写工作得到了四川省科技计划项目（No. 2019YFSY0032）的支持。西南交通大学“扬华学者”计划和研究生院对作者的教学改革与实践给予了大力支持，从而促成了本书的编著。再者，衷心感谢西南交通大学出版社对本书的出版所给予的帮助。

作者在繁重的教学、科研和学生指导培养工作中，能完成本书的编著，离不开亲人们的照顾、扶持和帮助，在此向亲人们表达深深的谢意。

全书由朱焱编著，陶霄、颜仕雄、杜强、张人之、何欢实现并优化了第二部分实践应用篇中的示范性实例系统。限于作者的水平和时间，书中难免存在不当之处，恳请读者及专家批评指正。

朱　焱

2019 年 6 月于成都

目 录

技术篇
基于 NoSQL 和 NewSQL 新技术的大数据管理与分析

实践应用篇
大数据智慧管理与分析之实践指南

技术篇

基于 NoSQL 和 NewSQL 新技术的大数据管理与分析

第 1 章　大数据及其特点

【本章要点】

- ✧ 大数据的定义与 5V 特点
- ✧ 大数据的应用场景
- ✧ 对大数据的误解
- ✧ CAP 理论与 BASE

1.1　大数据时代当前的状态

大数据（海量数据的全面升级版）已经成为风靡全世界的 IT（信息技术）新领域，毫不夸张地说，我们已经进入了一个大数据时代。Kantar Media CIC 每年都会通过一张信息图整理出中国互联网发展的数据。图 1.1 展示了 2017 年中国社交媒体、电子商务、共享经济等领域的大数据爆炸式增长。

类似地，图 1.2 展示了 2017 年短短 60 s 内，美国主要互联网公司产生的数据量。

在大数据时代，商务与科技人士对大数据及其发展的认知是十分积极的[1]。IBM（国际商业机器公司）调查了来自 70 个国家的 900 个商务和 IT 经理，这些商务领导者认为大数据带来的效应是[2]：

（1）他们基于数据进行绝大多数的决策的可能性增加到 166%。

（2）以数据分析为职业发展道路的可能性提升了 2.2 倍。

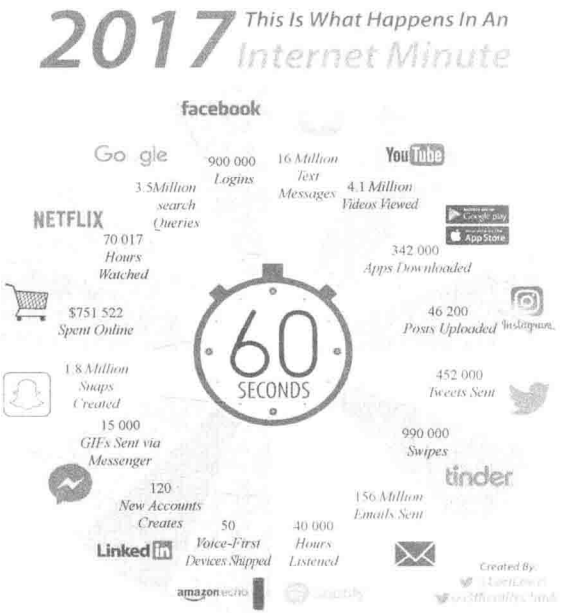
（3）他们在使用来自数据分析的关键价值资源方面增长了 75%。

（4）他们中 80%的人要衡量大数据对分析投资的影响力。

（5）他们中 85%的人拥有这样或那样的共享大数据分析资源。

中国 2017 年社会媒体 60 s 信息量（扫码查看彩图）

图 1.1 Kantar Media CIC 发布的 2017 年中国社会媒体 60 s 信息量[1]



美国 2017 年互联网 60 s 信息量（扫码查看彩图）

图 1.2 美国 2017 年互联网 60 s 内的信息生成量

① Kantar Media CIC（中国社会化商业资讯提供商）2017 年发布。http://www.ciccorporate.com/index.php?option=com_content&view= article& id=1379%3Akantar-media-cic-released-2017-every-60-seconds-in-china-infographic-big-data-for-understanding-chinese-social-media&catid=112%3AarcHives-2017&Itemid=223&lang=zh（2018-07-18 可访问）

TEK 系统针对大数据调查了 2 000 多名 IT 专业人士和 1 500 多名 IT 领导，得到了以下的统计数据[2]：

（1）90%的 IT 领导者和 84%的 IT 专家相信在大数据上投入时间、金钱和资源是值得的。

（2）14%的 IT 领导者认为，在他们的组织中大数据的概念会经常应用。

（3）66%的 IT 领导者和 53%的 IT 专家报告，他们的数据存储在完全不同的系统中。

（4）60% 的 IT 领导者和 53%的 IT 专家报告，他们的组织缺少对数据质量的责任感。

（5）多于 50%的 IT 领导者质疑他们的数据的有效性。

（6）81%的 IT 领导者认为他们的组织缺少必需的专业人员，这些人应该能计划、建设和执行大数据行动。

1.2　大数据定义与特点

1.2.1　什么是“大数据”（Big Data）？

按照全球最具权威的 IT 研究与顾问咨询公司 Gartner 的定义[3]，数据是海量、高增长率和多样化的信息资产，它需要性价比高并具有创新性的处理模式，才能具有更强的决策力、洞察力和流程自动化。在 Gartner Group 的定义中首次定义了大数据 3V 的特点。

大数据包含三类数据：无结构化数据、结构化数据、半结构化数据。无结构化数据是指数据没有预定义的结构、类型、模式或数据模型等，如 PDF、email、文本式数据。网页的 HTML 数据虽然有标签，但只是用于面向浏览器的文档显示样式渲染，并没有捕捉、存储和自动处理信息内容的功能，所以仍然是无结构化的。结构化数据是数据具有预先定义的符合规则的结构、类型和模式等，具有可处理、存储、使用的元数据信息，如传统的关系数据库数据。半结构化数据具有很有限的结构、数据类型或模式定义，如 XML。

1.2.2　大数据的 5V 特点

大数据的 5V 特点是 IBM 提出的，分别是数据量（Volume）、多样性

（Variety）、高速（Velocity）、价值（Value）和真实性（Veracity），具体要点如图 1.3 所示。

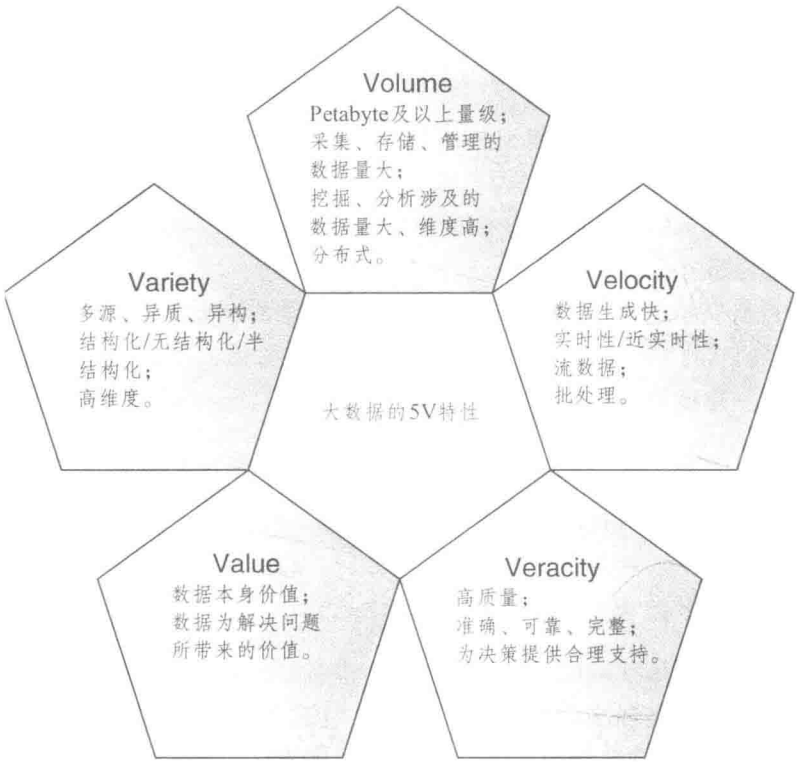


图 1.3 大数据的 5V 特性和具体要点

（1）Volume：表示数据量巨大，包括数据采集、存储和计算的量都非常大。大数据的起始计量单位至少是 PB（1 000 个 TB）、EB（100 万个 TB）或 ZB（10 亿个 TB）级别。

（2）Variety：数据的来源、类型、格式、语义等是多种多样的，具有多源异质异构性。80%的大数据是半结构化和非结构化的，例如：网络日志、社交网络平台数据、网页文件、电子商务交易数据、设备传感器数据、音频、视频、图片、地理位置信息等。多源异质异构数据对处理能力提出了更高的要求。

（3）Velocity：数据增长速度快，I/O 速度快，从而要求数据处理速度要快，时效性高。例如：天气大数据的动态性，提出了快速处理要求；“双十一”电商平台对高速增长的交易数据管理和处理；以毫秒级速率产生的

各种传感器数据；等等。

（4）Value：数据应被用于解决特定的问题或完成特定的任务，因此对大数据本身及其带来的价值要求高。大数据具有巨量性，但原始未加工的数据因其量大而杂乱稀释了本身的价值；大数据的巨量性也造成了数据挖掘算法的低效，甚至失效，挖掘出的结果并未带来期望的价值，不能满足各类应用对大数据大价值的要求，这也是大数据的特点和挑战。

（5）Veracity：大数据质量不高，体现在准确性、可靠性、完整性等方面的挑战。因为期望从大数据中获得决策，所以需要高质量的大数据。

现在大数据的 5V 特性又被扩展为 6V、7V 等。其中可视化（Visualization）作为大数据所呈现出的特点比较勉强，但作为理解大数据的基本处理要求是迫在眉睫的，因为可视化技术是大数据分析最直观、易懂，也是最理想的方法。这是对信息的一种新的阅读和理解方式。

1.3　沃尔玛应用大数据的案例

沃尔玛（Walmart Inc）开始使用大数据的时间甚至早于这个词汇享誉整个业界。2012 年沃尔玛已将 10 个节点的 Hadoop 平台集群升级到了 250 个节点，同时将存放在 Oracle、Netezza 和 Greenplum 硬件上的数据迁移到其自己的系统上，目标是将 10 个不同网站集成为一个网站，以便在新的 Hadoop 集群上存放所有新产生和流入的数据。

沃尔玛的实验室研发了许多大数据工具，如 Social Genome，ShoppyCat 和 Get on the Shelf。

按照资料[4]的介绍，Social Genome 等大数据分析工具，能够分析百万至十几亿的 Facebook 信息、推文、YouTube 视频、博文等，使得沃尔玛可以与在网络上提到某类商品的顾客或顾客的朋友们联系，告知他们特定的商品信息，包括降价等。这个软件工具组合了来源于万维网的公开数据、社交数据和个人专有数据，如顾客购买数据和合同信息。这样便构建了一个巨大的、持续变化和更新的知识库，库中管理了上亿的实体和关系，从而使得分析师和决策者能更好地理解顾客在线表达的上下文。例如，一位女士定期在推特上讨论电影，当她某次发推说“我喜欢盐”时，沃尔玛能够理解她正在谈论著名的好莱坞电影“盐”（Salt，中文翻译为“特工绍特”），而不是调味料“盐”。