大数据与农业应用

王文生　陈　明　编著

科学出版社

大数据与农业应用

王文生　陈　明　编著

科学出版社

北　京

内 容 简 介

本书结合大数据技术发展现状，对大数据在农业领域应用进行了探讨。全书共七章，第 1 章概述了国内外有关大数据的理论包括大数据产生背景和概念，第 2～4 章介绍了大数据的存储与管理、处理、分析等理论技术，第 5、6 章总结了所在科研团队近年的部分研究成果，从农业领域的角度，对大数据农业科研创新、大数据农业应用进行了阐述，第 7 章对农业大数据未来发展进行了展望。

本书适合从事农业大数据研究与应用的科技工作者、研究生等参考阅读。

图书在版编目 (CIP) 数据

大数据与农业应用/王文生，陈明编著. —北京：科学出版社，2015.1

ISBN 978-7-03-042618-5

Ⅰ. ①大… Ⅱ. ①王… ②陈… Ⅲ. ①信息技术－应用－农业－研究 Ⅳ. ①S126

中国版本图书馆 CIP 数据核字(2014)第 277687 号

责任编辑：于海云 / 责任校对：郑金红

责任印制：徐晓晨 / 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

北京厚诚则铭印刷科技有限公司 印刷

科学出版社发行 各地新华书店经销

*

2015 年 1 月第 一 版 开本：720×1 000 B5

2015 年 1 月第一次印刷 印张：10

字数：188 000

定价：30.00 元

（如有印装质量问题，我社负责调换）

序

ICT 快速发展正深刻地改变着世界，大数据科学作为新一代信息技术发展的后起之秀，对引领科学认知世界、促进产业发展方式和商业经营模式转变、社会公共管理等的重要变革起着重要的作用。数量庞大、种类广泛、迅速产生和更新的大数据，蕴含着前所未有的经济社会价值和商业价值，在推动经济发展、改善公共服务、保障国家安全、支撑前沿基础科技创新具有重大意义。

大数据科学的研究与应用，还处于萌芽起步阶段。从需求来看，很多产业对其使用价值仍缺乏意识；从供给来看，技术和人才储备滞后，仍然缺乏深厚的数据分析手段来支撑应用需求。大数据科学是一种应用驱动性很强的服务，需要从战略上重视大数据的开发利用，把它作为推动经济发展方式和社会公共管理等的有力抓手。抓住机遇，将拥有的数据资源转化为经济社会发展的动力，是摆在科技界、产业界面前的紧迫课题。

农业大数据涉及到农业水、土、光、热、气候资源以及作物育种、种植、施肥、植保、过程管理、收获、加工、存储、机械化等各环节多类型复杂数据的采集、挖掘、处理、分析与应用等问题。数据类别多样、结构复杂、内容广泛。从领域来看，以农业领域为核心（涵盖种植、林业、畜牧水产养殖、产品加工等子行业），逐步拓展到相关上、下游产业（饲料、化肥、农药、农机，仓储、屠宰业，肉类加工业等），并需整合宏观经济背景数据，包括统计数据、进出口数据、价格数据、生产数据、气象、灾害数据等；从地域来看，以国内区域数据为核心，借鉴国际农业数据作为有效参考，不仅包括全国层面数据，还应覆盖省市数据，甚至地市级数据，为区域农业发展研究提供基础；从粒度来看，不仅包括统计数据，还包括涉农经济主体基本信息、投资信息、股东信息、专利信息、进出口信息、招聘信息、媒体信息、地理空间坐标信息等；从专业性来看，需要分步构建农业领域的专业数据资源，进而逐步有序规划各专业领域的数据资源，如针对畜品种的生猪、肉鸡、蛋鸡、肉牛、奶牛、肉羊等专业检测数据等。

农业是生命之源、发展之基。农业资源、环境、多样化的生产经营方式、保障舌尖上的安全，都时刻面对不断产生大批非结构化数据的信息获取、挖掘、存储、处理与智慧应用的问题。利用大数据技术和理念，加快推进农业现代化与信息化的深度融合，将为建设现代农业，推进社会主义新农村建设，提高农民的社会经济地位，构建社会主义和谐社会和建设资源节约、环境友好的可持续农业提供重要科学技术支撑。

《大数据与农业应用》一书，结合大数据技术与理论，阐述了大数据农业应用的新思路、新方法，对于新形势下农业大数据的应用发展研究具有重要的指导意义和借鉴作用。相信本书的出版对于从事相关理论、技术研究的科技工作者和相关管理工作者，都会大有裨益。

农业大数据蕴含着巨大的机遇，作为世界上第一人口的农业大国，我国农业科技工作者在利用大数据技术创新农业发展上任重而道远。包括本书编著者在内的一大批仁人志士正在不懈努力，积极思考大数据农业应用的模式，为我国农业、农村信息化发展，进而为加快推进我国农业现代化建设做出贡献。希望从事新一代信息技术和大数据农业应用研究的科技界、产业界的朋友们，积极关注并支持我国农业大数据的应用发展研究，为农业转方式、调结构，产业提质增效，农民持续增收，坚持走出一条生产技术先进、经营规模适度、市场竞争力强、生态环境可持续的中国特色农业现代化道路做出更大的贡献！

王樊华

2015 年 1 月

前　言

继物联网、云计算之后，大数据已经成为当前信息技术产业最受关注的概念之一。大数据正在成为国家竞争的前沿，以及产业竞争力和商业模式创新的源泉。发达国家的大数据产业发展步入大规模商业化阶段，已广泛渗透到经济、政治、教育、安全和社会管理等众多领域。我国有关部门正在组织各方力量进行专项研究，从国家层面通盘考虑我国大数据发展战略，推动大数据的收集、分析和应用，引导和推动各行业对大数据进行研究与利用，推动各个领域的大数据应用工作，提升科学决策能力。

农业是国民经济的基础，关系到国计民生。党和国家高度重视三农工作，推动实施了一系列重大举措，农业农村发展取得了举世瞩目的巨大成就。然而，面对复杂严峻的国内外经济形势，农业发展面临多方面巨大变化和多重挑战，人口、耕地、环境等问题越来越突出。推动现代农业发展，保障国家粮食安全，必须依靠科技创新，走一条符合中国国情农业现代化道路。大数据是与自然资源、人力资源一样重要的战略资源，也是一种重要的科技创新资源。正如半个多世纪以前石油作为一种化石能源投入农业奠基当代绿色革命发展一样，大数据在农业上的应用也将带动农业产生一次新的革命，在今后几十年将极大地提升农业生产力、创造新的农业价值，造福亿万农户，并提高政府管理效率和服务水平。

我们科研团队近年主持和参加了国家有关部门组织的与大数据相关的研究项目，对农业大数据进行了初步研究与利用，开展了农业大数据的挖掘、分析、应用，并以之为基础面向公众提供各种类型的信息服务。为了促进我国农业大数据研究的学术交流、推动学科发展、培养专业人才，我们通过概述国内外有关大数据理论，总结所在科研团队近年部分研究成果，撰写完成此书。

在本书撰写的过程中，我们得到了许多专家的帮助和支持，在这里我们无法一一列举，谨向他们表示诚挚的感谢。我们尤其要感谢中国工程院汪懋华院士的指导并为本书作序。我们还要感谢中国农业科学院李金祥副院长和中国农业科学院农业信息研究所原所长梅方权研究员为本书的撰写提供很多宝贵建议。中国农业科学院农业信息研究所的李秀峰主任、谢能付副研究员、杨勇副研究员、孙志国副研究员、郭雷风博士、尹国伟博士、杨永前等为本书的编写做了大量细致的工作。

本书引用了许多国内外学者的研究成果，在此一并表示感谢。由于作者水平所限，书中某些内容可能有不妥之处，恳请读者批评指正。

王文生

2014 年于中国农业科学院

目　录

第 1 章　大数据概述

需求是科学技术发展的原动力。目前，大数据问题的出现与研究已经成为了计算机科学与技术研究的新热点，并显示出日益强大的吸引力，科学大数据催生了数据密集型知识发现的科学研究第四范式。对于信息领域，大数据带来的不仅是机遇，还有一系列的困难和挑战。目前，大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界寄予无比的厚望。大数据成为继 20 世纪末、21 世纪初互联网蓬勃发展以来的又一轮 IT 工业革命。

1.1　问题的提出

在全世界范围内，以电子方式存储的数据(又简称为电子数据)总量空前巨大。在 2011 年电子数据总量已达到 1.8ZB(1ZB=1024PB)，较 2010 年同期提高超过 1ZB，统计结果表明，每经过 2 年就可以增加一倍，预计到 2020 年可达到 35ZB，如图 1-1 所示。面对数据增长的速度迅猛地提升，数据量的飞速增加，对大量电子数据的高效存储、高效传输与快速地处理是必须面对的研究问题。

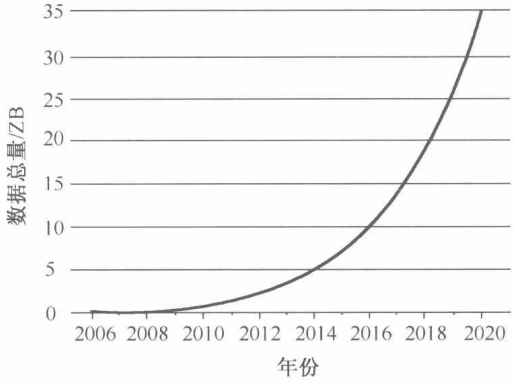

图 1-1　全球数据创建及复制的数据总量预测

1.1.1　电子数据迅速增加

物联网、云计算、移动互联网、车联网、手机、平板电脑、个人计算机(PC)、气候信息、公开的信息(例如杂志、报纸)、交易记录、网络日志、病历、军事监控、

视频和图像档案，以及大型电子商务和在地球各地的各种传感器是数据来源，再加上承载的方式不断更新与发展、大型科学研究设备产生的数据、以及社交媒体的快速发展，这些构成了大数据持续产生的生态环境。尤其是近年来，随着互联网技术的发展，来自人们的日常生活，特别是来自互联网服务而产生的大量数据迅猛增加。据不完全统计，互联网当前包含 93 亿多个页面，80%～85%的数据是存储在数据库的文本中。互联网一天产生的全部内容可以用 1.68 亿张 DVD 存储，发出的邮件有 2940 亿余封，发出的社区帖子达 200 万个(这相当于《时代》杂志 770 年的文字量)，卖出的手机为 37.8 万台，比全球每天出生的婴儿数量 37.1 万还多。从数据统计角度来看，电子数据量迅速增加。预计中国数据技术和服务市场将来 5 年的复合增长率为 51.4%，其中增长率最高的是存储市场为 60.8%，服务器市场的增长率则为 38.3%，远高于其他产品相关的市场。

1.1.2 数据蕴藏巨大的经济价值

数据本身是无意义的，而通过统计、分类、萃取、特征抽取等一系列技术手段，可以从数据中产生信息与知识。数据是重要的战略资源，蕴含巨大的经济价值，因此已经引起科技界和和企业界的高度重视。有效地组织和使用数据，将对经济发展产生巨大的推动作用。大数据出现孕育着前所未有的机遇。对大数据的采集、整合和分析，可以发现新的知识，创造新的价值，带来知识、科技和利润的大发展。

越来越多的企业等机构意识到数据是最重要的资产，数据分析能力逐渐成为核心竞争力。企业将远离服务与咨询，更多地专注于因数据分析而带来的全新业务增长点。数据将成为决定胜负的基本因素，最终将成为人类至关重要的自然资源。各著名的大型公司已经开始开发自己的大数据处理和存储系统，如何整合这些数据成为未来的关键任务。

在互联网、电信、金融等行业，几乎已经到了数据就是业务本身的地步。物联网、社交网络等新的互联网技术在改变人们生活方式的同时，也产生了大量的数据。有效地存储和查询这些数据，通过数据挖掘，从数据中获得有用的信息，为用户提供好的用户体验，进而增强企业的竞争力，这是一个新挑战。目前，两天就能产生出自人类文明诞生以来到 2003 年所产生的数据的总量。大数据已经成为重要的时代特征，充分使用大数据和挖掘大数据商业价值将为行业企业带来强大经济效益与竞争力。

大数据全生命周期可以分为“数据产生、数据采集、数据传输、数据存储、数据处理、数据分析、数据发布、展示和应用、产生新数据”等阶段。已经形成了大数据的“生产与集聚层、组织与管理层、分析与发现层、应用与服务层”的产业链，而 IT 基础设施为这各环节提供基础支撑。

据统计，2012 年市场规模达到 4.5 亿元，2016 年估计可达到百亿规模，如图 1-2 所示。

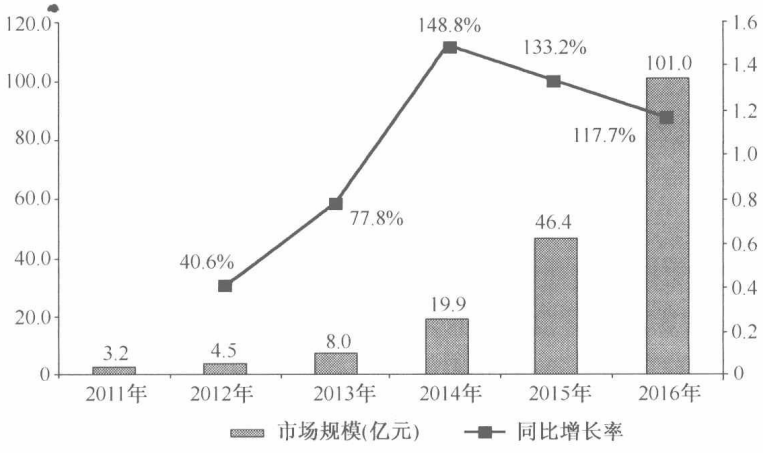

图 1-2　中国大数据应用市场规模与增长

1.1.3　数据是国家的核心资产

一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，对数据的占有和控制将成为陆权、海权、空权之外的另一种国家核心资产。联合国也在 2012 年发布了大数据政务白皮书，指出大数据对于联合国和各国政府是一个历史性的机遇，通过使用极为丰富的数据资源，对社会经济进行前所未有的实时分析，帮助政府更好地响应社会和经济运行。

数据为王的大数据时代已经到来，对数据的占有和控制也将成为国家间和企业间新的争夺点。大数据技术的专业人才，特别是数据分析复合型人才将会影响该市场的发展。

在技术层面上，大数据、海量数据与超大规模数据无本质的区别，它们都是指用传统处理方法无法处理的大量数据。通过对大数据的高速有效地处理，可以发现数据中蕴藏的规律与规则，进而为各种关键决策提供依据与指导，正确的预测与决策将导致巨大财富的产生。技术与工具密不可分，目前常用的数据处理技术与工具是小数据处理技术与工具，一些海量数据处理方法与工具是一种过渡性的方法与工具，大数据处理技术与工具的研究是一项有理论意义和实际价值的工作。大数据技术就是从各种类型的数据中快速获得智慧的技术。信息要能转化成智慧，至少要满足以下三个标准。

1. 可破译性

可破译性是大数据时代特有的问题，但非结构化的数据不是一定都可破译。例

如，记录了某客户在网站上三次翻页的时间间隔分别是 5s，4s，15s，却忘记标注这三个时间代表什么，也就是说，知道这些数据是信息，却不可破译，所以不可能成为知识。

2. 关联性

关联性即是相关性。无关的信息可以被看作是噪声。

3. 新颖性

新颖性是指无法仅仅根据拥有的数据和信息进行判断。例如，某电子商务公司通过一组数据/信息，分析出了客户愿意为当天送货的产品多支付 10 块钱，然后又通过另一组完全独立的数据/信息得到了同样的内容，这样的情况下，后者就不具备新颖性。但是，很多时候，只有在处理了大量的数据和信息以后，才能判断它们是否具有新颖性。

1.1.4 大数据的产生源泉

大数据是人类活动的产物，来自于人们改造客观世界的过程中，是生产与生活在网络空间的投影。信息爆炸是对信息快速发展的一种逼真的描述，形容信息发展的速度如同爆炸一般席卷整个地球。在 20 世纪 40～50 年代，信息爆炸主要是指科学文献的飞快增长；而到 20 世纪 90 年代，由于计算机和通信技术广泛应用，信息爆炸主要是指社会信息飞快增长，包括正式交流过程和非正式交流过程所产生的电子式的和非电子式的信息，而到 21 世纪的今天，信息爆炸是由于数据洪流的产生和发展所造成信息飞快增长。在技术方面，新型的数据中心、分布式计算、云计算、大容量数据存储与处理技术、社会网络、移动终端设备、多种数据采集方式使大数据的产生和存储成为可能。在用户方面，日益人性化的用户界面、信息行为模式都容易作为数据而存储，用户即可成为数据的制造者，也可以成为数据的使用者。可以看出，随着云计算、物联网计算和移动计算的发展，世界上所产生的新数据都能够汇入数据洪流，导致数据洪流席卷互联网。

归纳起来，大数据主要来自于物理世界与互联网世界。

1. 互联网世界

大数据时代，需要更加全面的数据来提高预测的准确度，因此需要更多廉价、便捷、自动的数据生产工具。

随着互联网无处不在地渗透到人们的工作和生活，移动互联网、物联网、可穿戴联网设备的普及，数据正在以指数级别的加速度产生，目前世界上 90%的数据是由互联网出现以后迅速产生的。

大数据来自人类社会，尤其互联网的发展为数据的存储、传输与应用创造了基础与环境。依据基于唯象假设的六度分割理论而建立的社交网络服务(Social Network Service，SNS)，以认识朋友的朋友为基础，扩展自己的人脉。基于 Web 2.0 网站建立的社交网络，用户既是网站信息的使用者，也是网站信息的制作者。社交网站记录人们之间的交互，搜索引擎记录人们的搜索行为和搜索结果，电子商务网站记录了人们购买商品的喜好，微博网站记录了人们所产生的即时的想法和意见，图片视频分享网站记录了人们的视觉观察，百科全书网站记录了人们对抽象概念的认识，幻灯片分享网站记录了人们的各种正式和非正式的演讲发言，机构知识库和开放获取期刊记录了学术研究成果等。归纳起来，来至互联网的数据可以划分为下述 6 种类型。

1)视频

视频图像是大数据的主要来源之一，电影、电视节目可以产生大量的视频图像，各种室内外的视频摄像头昼夜不停地产生巨量的视频图像。视频图像以每秒几十帧的速度连续记录运动着的物体，一个小时的标准清晰视频经过压缩后，所需的存储空间为 GB 数量级，对于高清晰度视频所需的存储空间就更大了。

2)图片与照片

图片与照片也是大数据的主要来源之一，截至 2011 年 9 月，用户向脸谱(Facebook)上传了 1400 亿张以上的照片，脸谱是美国最大的一个社交网站，类似于中国的新浪微博。如果拍摄者为了保存拍摄时的原始文件，平均每张照片大小为 1MB，则这些照片的总数据量就是 140G × 1MB=140PB，如果单台服务器磁盘容量为 10TB，则存储这些照片需要 14000 台服务器，而且这些上传的照片仅仅是人们拍摄到的照片的很少一部分。此外，许多遥感系统一天 24 小时不停地拍摄并产生大量照片。

3)音频

DVD 光盘采用了双声道 16 位采样，采样频率为 44.1kHz，可达多媒体欣赏水平。如果某音乐剧的长度为 5.5min，计算其占用的存储容量为

$$\begin{aligned}\text{存储容量} &=(\text{采样频率}\times\text{采样位数}\times\text{声道数}\times\text{时间})/8\\&=(44.1\times1000\times16\times2\times5.5\times60)/8\\&=12.6\text{MB}\end{aligned}$$

4)日志

网络设备、系统及服务程序等，在运作时都会产生 log 的事件记录；每一行日志都记载着日期、时间、使用者及动作等相关操作的描述。网络操作系统设有各种日志文件，如应用程序日志，安全日志、系统日志、Scheduler 服务日志、FTP 日志、WWW 日志、DNS服务器日志等，并且根据系统开启服务的不同而有所不同。用户在系统上进行操作时，日志文件通常记录了用户操作的相关内容，这些内容被系统安全工作人员使用。例如，有人对系统进行了 IPC探测，系统就会在安全日志中迅

速地记下探测者探测时所用的 IP、时间、用户名等；用 FTP 探测后，就会在 FTP 日志中记下IP、时间、探测所用的用户名等。

网站日志记录了用户对网站的访问，电信日志记录了用户拨打和接听电话的信息，假设有 5 亿用户，每个用户每天呼入呼出 10 次，每条日志占用 400B，并且需要保存 5 年，则数据总量为 5×10×365×400×5B = 3.65PB。

5) 网页

2000 年，谷歌索引了大约 10 亿个网页，如果平均每个网页用 25KB，一万亿个网页的数据总量为 25PB。

6) 结构化数据

视频、图片与照片、音频和日志都是非结构化数据，网页是半结构化数据。结构化数据只占 15%左右，结构化数据可以在结构数据库中存储，并可用二维表来逻辑表达数据。这类数据是先定义结构，然后才有数据。这类数据在大数据中所占比例较小，但却应用广泛、起到关键作用，例如银行财务系统、股票与证券系统、信用卡系统等。

2. 物理世界

来自物理世界的数据主要是指来自大型国际实验；跨实验室、单一实验室或个人观察实验所得到的科学实验数据或传感数据，最早提出大数据概念的学科是天文学和基因学，这两个学科从诞生之日起就依赖于基于海量数据的分析方法。科学实验由科技人员设计，数据采集与集成、数据处理都是事先设计，无论是检索还是数据挖掘，都有科学规律可循。例如，希格斯粒子又称为上帝粒子的寻找，采用了大型强子对撞机实验。这是一个基于大数据的科学实验，至少要在 1 万亿个事例中才可能找出一个希格斯粒子。可以看出，科学实验的大数据处理是整个实验的一个预定过程，所以是否发现有价值的信息可在预料之中。大型强子对撞机每秒生成数据量约为 1PB。建设中的下一代巨型射电望远镜阵每天生成的数据大约在 1EB。波音发动机上的传感器每小时产生 20TB 左右的数据。

随着科研人员获取数据方法与手段的变化，科研活动产生的数据量激增，科学研究已成数据密集型活动。科研数据因其数据规模大、类型复杂多样、分析处理方法复杂等特征，已成为大数据的一个典型代表。大数据所带来的新科学研究方法反映了未来科学的行为研究方式，数据密集型科学研究将成为科学研究的普遍范式。

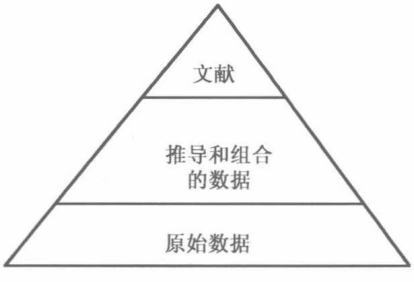

图 1-3　在线科学数据

利用互联网可以将所有的科学数据与文献联系在一起，创建了一个文献与数据能够交互操作的系统，即在线科学数据系统，如图 1-3 所示。

对于这些在线科学数据，许多领域互相交叉，可以使用其他领域的数据。互联网能够将所有文献与数据集成在一起，可以实现从文献计算到数据，再回到数据。这样可以提高科技信息的速度，进而大幅度地提高生产力。也就是说，在阅读某人的论文时，可以查看他们的原始数据，甚至可以重做分析，也可以查看某些数据时查处所有关于这一数据的文献。

1.2 大数据的概念与特性

1.2.1 大数据的概念

大数据是指数据规模大，尤其是因为数据形式多样性、非结构化特征明显，导致数据存储、处理和挖掘异常困难的那类数据集。大数据需要管理的数据集规模很大，数据的增长快速，类型繁多，如文本、图像、视频等。处理包含数千万个文档、数百万张照片或者工程设计图的数据集等，如何快速访问数据成为核心挑战。大数据是指无法用常规的软件工具捕捉、处理的数据集合，即大数据 Big-Data={S_1,S_2,S_3}，其中 S_1 代表结构化数据集，S_2 代表非结构化数据集，S_3 代表半结构化数据集。而大数据是指{S_1,S_2,S_3}所占的存储空间达到 PB 数量级。

大数据是人类活动的产物，来自人们认识世界与改造世界的过程中，是生产与生活在网络空间的投影。通常将其归纳为 5 个“V”：Volume(数据量)，Variety(多样性)，Value(价值)，Velocity(速度)，Value(真实性)，如图 1-4 所示。Value 代表价值密度低，商业价值高，以视频为例，连续不间断监控过程中，可能有用的数据仅占几秒钟；Velocity 代表数据变化速度快。

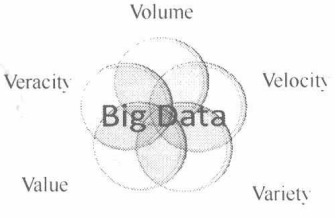

图 1-4 大数据的 5 个 V

1. 数据容量巨大

Volume 代表数据量巨大，存储容量单位的定义如表 1-1 所示。

表 1-1 存储容量单位的定义

单 位	定 义	字节数(2 进制)	字节数(10 进制)
Kilobyte(千)	1024Byte	2^{10}	10^3
Megabyte(兆)	1024 Kilobyte	2^{20}	10^6
Gigabyte(吉)	1024 Megabyte	2^{30}	10^9
Terabyte(太)	1024 Gigabyte	2^{40}	10^{12}
Petabyte(拍)	1024 Terabyte	2^{50}	10^{15}
Exabyte(艾)	1024 Petabyte	2^{60}	10^{18}
Zettabyte(泽)	1024 Exabyte	2^{70}	10^{21}
Yottabyte(尧)	1024 Zettabyte	2^{80}	10^{24}