DASHUJU SHIDAI
GAOXIAO XINXIHUA ZHANLUE YU SHIJIAN

大数据时代
高校信息化战略与实践

主编　王继成　李竹林

東北大學出版社
Northeastern University Press

大数据时代高校信息化战略与实践

主编　王继成　李竹林

东北大学出版社
·沈　阳·

图书在版编目（CIP）数据

大数据时代高校信息化战略与实践 / 王继成，李竹林主编. — 沈阳：东北大学出版社，2016.5
ISBN 978-7-5517-1266-8

Ⅰ. ①大… Ⅱ. ①王… ②李… Ⅲ. ①信息技术—应用—高等学校—研究—中国
Ⅳ. ①G649.2

中国版本图书馆 CIP 数据核字（2016）第 100416 号

出 版 者：东北大学出版社
地址：沈阳市和平区文化路三号巷 11 号
邮编：110819
电话：024－83687331（市场部）　83680267（社务部）
传真：024－83680180（市场部）　83687332（社务部）
E-mail: neuph@ neupress. com
http://www. neupress. com
印 刷 者：沈阳市新友印刷有限公司
发 行 者：东北大学出版社
幅面尺寸：185mm × 260mm
印　　张：24
字　　数：614 千字
出版时间：2016 年 5 月第 1 版
印刷时间：2016 年 5 月第 1 次印刷
责任编辑：汪彤彤　　责任校对：木　卫
封面设计：刘江旸　　责任出版：唐敏志

ISBN 978-7-5517-1266-8　　定　　价：50.00 元

《大数据时代高校信息化战略与实践》

编委会

序 言

我们身处信息化、数字化的时代，各种数据纷繁杂乱，如何利用这些数据，如何将这些数据进行专业化处理，使之对我们现有的工作、生活、学习形成有指导性的决策，如何挖掘这些数据，使之形成有效的数据模型，从而对我们未来的工作、生活、学习提出前瞻性的预测，是亟待解决的问题。云技术的诞生，为大数据的采集、计算、分析提供了强有力的支撑。

而教育行业的数据更像是一个蕴藏巨大能量的宝库，如何收集整理这些数据，并对之进行行之有效的分析、建模，使之服务于教学、科研、管理等，实现学校的个性化管理，成为很多专家学者乐于探讨的问题。

2015 年 5 月，国家主席习近平在致国际教育信息化大会的贺信中指出："当今世界，科技进步日新月异，互联网、云计算、大数据等现代信息技术深刻改变着人类的思维、生产、生活、学习方式，深刻展示了世界发展的前景。"

为了研究高校信息化与大数据的关系，首先关注 3 个问题。

1. 国家对高校信息化的重视

教育部在 2012 年发布的《教育信息化十年发展规划（2011—2020 年）》中明确指出："我国教育改革和发展正面临着前所未有的机遇和挑战。以教育信息化带动教育现代化，破解制约我国教育发展的难题，促进教育的创新与变革，是加快从教育大国向教育强国迈进的重大战略抉择。"

2. 国家对大数据发展的重视

2015 年 9 月，国务院印发《促进大数据发展行动纲要》（以下简称《纲要》），系统部署大数据发展工作。

《纲要》明确指出，推动大数据发展和应用，在未来 5 ~ 10 年打造精准治理、多方协作的社会治理新模式，建立运行平稳、安全高效的经济运行新机制，构建以人为本、惠及全民的民生服务新体系，开启大众创业、万众创新的创新驱动新格局，培育高端智能、新兴繁荣的产业发展新生态。

3. 大数据时代下的高校信息化

利用云计算技术从海量数据中寻找出有意义的规律，并为教育信息化平台管理与发展的决策提供技术支持，是信息化数据分析的目标，没有数据的留存和深度挖掘，教育信息

化可能只会流于形式。

《大数据时代高校信息化战略与实践》一书首先从大数据、云计算、高校信息化建设的概念出发，通过前 8 章的内容阐述了高校信息化建设的关键技术，并且详细论述了用于大数据计算的云平台 Hadoop 的相关技术，最后援引几所高校的实例进一步说明如何在大数据时代开展高校信息化建设工作。

本书的作者包括高校信息化建设的管理者、信息化工作的具体实施者和相关研究的一线教师，本书涵盖的内容包括技术理论、管理论述和多所高校的具体实例，是一本难得的著作，相信本书一定能对我国高校未来的信息化建设起到指导作用。

中国教育和科研计算机网
东北地区网络中心总经理

前 言

本世纪初期，随着互联网络的不断普及，我国高校的校园网络陆续建设和完善，全国各大高校都建立了自己的校园网络，建设了各自的信息化系统，进而建设了数字化校园，并努力把其建设为智慧校园。

近几年来，云计算、大数据、物联网等概念相继提出，建设高校的数字化校园开始向建设智慧校园转变，本书正是写于这个转变的时期。书中所倡导的“统一规划、软硬并重、分步实施、以点带面、重点突破、持之以恒”的信息化校园建设理念和所涉及的网络基础设施建设、基本网络服务系统建设、网络应用支持系统开发、网络信息服务系统开发等及智慧校园所涉及的一些基本概念，对于高校建设智慧校园有着一定的指导作用，也是高校学生学习和了解高校信息化的入门教程。

本书由沈阳农业大学王继成、李竹林统稿；第 1 章由沈阳农业大学刘伟、陈杨编写；第 2 章由沈阳农业大学王继成、辛彦军、潘飞，内蒙古农业大学李东泽，辽宁工程职业学院刘家庚编写；第 3、4 章由沈阳农业大学刘占文，沈阳工程学院谷彩连，沈阳仪表科学研究院张文良编写；第 5 章由沈阳农业大学李娜、高振东编写；第 6、7、8 章由沈阳农业大学李竹林、蔡明知、江红霞、王兴阳、赵洋旭编写；第 9 章由沈阳航空航天大学宛岳，沈阳农业大学潘飞编写；第 10 章由沈阳农业大学王铁良、马殿荣、郑伟编写；第 11 章由沈阳农业大学李竹林、霍春梅、李迎、潘飞、高振东，北京林业大学黄国华，沈阳大学周强编写。

本书的出版得到了“东北老工业地区大学农业科技服务关键技术集成与示范”（编号：2013BAD20B08）课题的资助。本书在编写过程中得到了赛尔网络东北区分公司、北京林业大学、东北大学、东北林业大学、沈阳航空航天大学、内蒙古农业大学、沈阳工程学院、沈阳大学的大力支持，在此一并表示感谢。

因时间仓促、水平有限，书中难免存在错误和不足，敬请广大读者、专家和同行不吝赐教，竭诚致谢并请各位读者海涵。

编 者

2016 年 4 月

目 录

第 1 章　大数据与云计算

1.1　大数据概述

1.1.1　大数据的概念

“大数据”这个术语最早期的引用可追溯到 apache. org 的开源项目 Nutch。当时，大数据用来描述为更新网络搜索索引需要同时进行批量处理或分析的大量数据集。随着谷歌 MapReduce 和 Google File System（GFS）的发布，大数据不再仅用来描述大量的数据，还涵盖了处理数据的速度。早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。不过，大约从 2009 年开始，“大数据”才成为互联网信息技术行业的流行词语。美国互联网数据中心指出，互联网上的数据每年将增长 50%，每两年便将翻一番，而目前世界上 90% 以上的数据是最近几年才产生的。此外，数据又并非单纯指人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、振动、温度、湿度乃至空气中化学物质的变化，也产生了海量的数据信息。

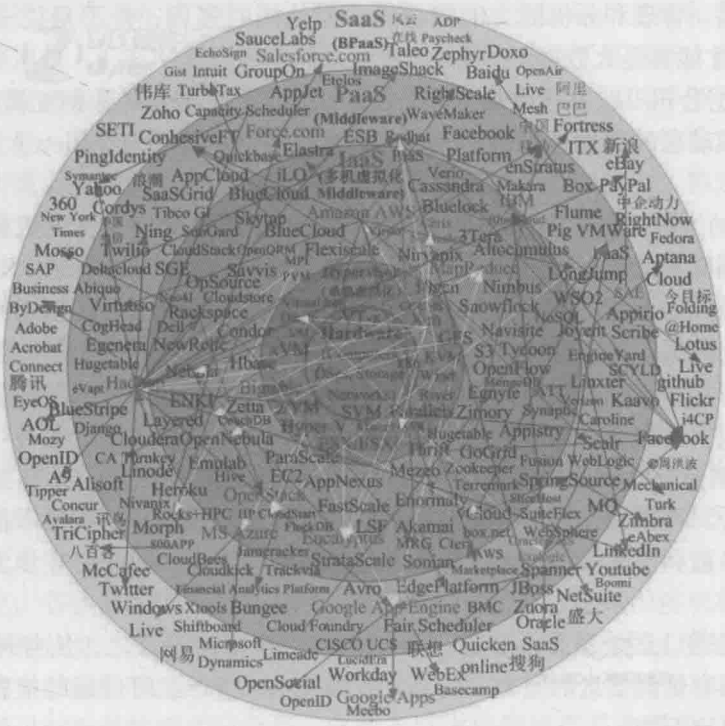

图 1-1　著云台大数据

大数据（big data，mega data）是指一个体量特别大、数据类别特别大的数据集，并且这样的数据集无法用传统数据库工具对其内容进行抓取、管理和处理。大数据首先是指数据体量（volumes）大，指代大型数据集，一般在10TB规模左右，但在实际应用中，很多企业用户把多个数据集放在一起，已经形成了PB级的数据量；其次是指数据类别（variety）大，数据来自多种数据源，数据种类和格式日渐丰富，已冲破了以前所限定的结构化数据范畴，囊括了半结构化和非结构化数据；再次是数据处理速度（velocity）快，在数据量非常庞大的情况下，也能够做到数据的实时处理；最后一个特点是数据真实性（veracity）高，随着社交数据、企业内容、交易与应用数据等新数据源的兴起，传统数据源的局限被打破，企业愈发需要有效的信息之力以确保其真实性及安全性。

关于大数据的概念和准确定义仍有不同的解释。

① 大数据，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理，并整理成为帮助企业经营决策更积极目的的资讯。

② 在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中，大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。

③ 研究机构Gartner给出的定义为，大数据是需要新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。从数据的类别上看，大数据指的是无法使用传统流程或工具处理或分析的信息。它定义了那些超出正常处理范围和大小、迫使用户采用非传统处理方法的数据集。

④ 麦肯锡是研究大数据的先驱，在其报告 *Big data: The next frontier for innovation, competition, and productivity* 中给出大数据的定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。但他同时强调，并不是说一定要超过特定TB值的数据集才能算是大数据。

⑤ 国际数据公司IDC从大数据的4个特征来定义，即海量的数据规模（Volume）、快速的数据流动和动态的数据体系（Velocity）、多样的数据类型（Variety）、巨大的数据价值（Value）。

⑥ 亚马逊的大数据科学家John Rauser给出了一个简单的定义：大数据是任何超过了一台计算机处理能力的数据量。

⑦《著云台》的分析师团队认为，大数据通常用来形容一个公司创造的大量非结构化数据和半结构化数据，这些数据在下载到关系型数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像MapReduce一样的框架来向数十、数百甚至数千的电脑分配工作。

⑧ 大数据就是互联网发展到现今阶段的一种表象或特征而已，没有必要神话它或对它保持敬畏之心，在以云计算为代表的技术创新大幕的衬托下，这些原本很难收集和使用的数据开始容易被利用起来了，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

⑨ 从某种程度上说，大数据是数据分析的前沿技术。简言之，从各种各样类型的数据中，快速获得有价值信息的能力，就是大数据技术。世界经济论坛的报告认定大数据为新财富，价值堪比石油。

⑩ 大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的

数据集合。

大数据的概念远不止大量的数据（TB）和处理大量数据的技术，或者所谓的“4个V”之类的简单概念，而是涵盖了人们在大规模数据的基础上可以做的事情，而这些事情在小规模数据的基础上是无法实现的。换句话说，大数据让我们以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见，最终形成变革之力，核心就是预测。大数据将为人类的生活创造前所未有的可量化的维度。

2009年出现了一种新的流感病毒——甲型H1N1流感，结合了导致禽流感和猪流感的病毒的特点，在短短几周之内迅速传播开来。全球的公共卫生机构都担心一场致命的流行病即将来袭。有的评论家甚至警告说，可能会暴发大规模流感，类似于1918年在西班牙暴发的、影响了5亿人口并夺走了数千万人性命的大规模流感。更糟糕的是，我们还没有研发出对抗这种新型流感病毒的疫苗。公共卫生专家能做的只是减慢它的传播速度。但要做到这一点，他们必须先知道这种流感出现在哪里。美国，和所有其他国家一样，都要求医生在发现新型流感病例时告知疾病控制与预防中心（CDC）。但由于人们可能患病多日，实在受不了了才会去医院，同时这个信息传达回疾控中心也需要时间，因此，通告新流感病例时往往会有一两周的延迟。而且，疾控中心每周只进行一次数据汇总。然而，对于一种飞速传播的疾病，信息滞后两周的后果将是致命的。这种滞后导致公共卫生机构在疫情暴发的关键时期反而无所适从。在甲型H1N1流感暴发的几周前，互联网巨头谷歌公司的工程师们在《自然》杂志上发表了一篇引人注目的论文，它令公共卫生官员们和计算机科学家们感到震惊。文中解释了谷歌为什么能够预测冬季流感的传播：不仅是全美范围的传播，而且可以具体到特定的地区和州。谷歌通过观察人们在网上的搜索记录来完成这个预测，而这种方法以前一直是被忽略的。谷歌保存了多年来所有的搜索记录，而且每天都会收到来自全球超过30亿条的搜索指令，如此庞大的数据资源足以支撑和帮助它完成这项工作。发现能够通过人们在网上检索的词条辨别出其是否感染了流感后，谷歌公司把5000万条美国人最频繁检索的词条和美国疾控中心在2003年至2008年间季节性流感传播时期的数据进行了比较。其他公司也曾试图确定这些相关的词条，但是它们缺乏像谷歌公司一样庞大的数据资源、处理能力和统计技术。虽然谷歌公司的员工猜测，特定的检索词条是为了在网络上得到关于流感的信息，如“哪些是治疗咳嗽和发热的药物”，但是找出这些词条并不是重点，他们也不知道哪些词条更重要，更关键的是，他们建立的系统并不依赖于这样的语义理解。他们设立的这个系统唯一关注的就是特定检索词条的频繁使用与流感在时间和空间上的传播之间的联系。谷歌公司为了测试这些检索词条，总共处理了4.5亿个不同的数字模型。在将得出的预测与2007年、2008年美国疾控中心记录的实际流感病例进行对比后，谷歌公司发现，他们的软件发现了45条检索词条的组合，一旦将它们用于一个数学模型，他们的预测与官方数据的相关性高达97%。和疾控中心一样，他们也能判断出流感是从哪里传播出来的，而且他们的判断非常及时，不会像疾控中心一样要在流感暴发一两周之后才可以做到。所以，2009年甲型H1N1流感爆发的时候，与习惯性滞后的官方数据相比，谷歌成为了一个更有效、更及时的指示标。公共卫生机构的官员获得了非常有价值的数据信息。惊人的是，谷歌公司的方法甚至不需要分发口腔试纸和联系医生——它是建立在大数据的基础之上的。这是当今社会所独有的一种新型能力：以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见。基于这样的技术理念和数据储备，下一次流感来袭的时候，世界将会拥有一种更好的预测