第一章 大数据的理论演化

近几十年来，随着信息技术发展的日新月异，数据分析已广泛运用于国民经济、商业实践、国家治理和社会生活等各个领域，深刻影响着人们的日常生活，“大数据”概念开始备受社会各界关注。本章从大数据概念提出的背景出发，按时间顺序梳理大数据理论的演变过程，对当前的大数据概念进行明确界定，并指出大数据的趋势变革。

第一节　大数据的提出背景

“大数据”概念的提出建立在信息技术进步的基础上，有其清晰的社会历史发展脉络，迎合着现代产业结构转型升级的需要。硬件存储性能、光纤传输带宽等基础设施的完善，互联网、云计算与物联网技术的发展，网络社交及智能终端的普及都为“大数据”概念的提出奠定了基础，并推动“大数据”这一概念不断渗透到更多相关领域。

一、技术进步

（线上阅读）

（一）信息基础设施的完善

作为英特尔的创始人之一，戈登·摩尔（Gordon Moore）于 1965 年提出了著名的“摩尔定律”。该定律阐述了计算机存储器的未来发展趋势，即每隔 18 个月，计算机存储器的性能便会提升一倍，即计算机的计算、存储能力将相对于时间周期呈指数式上升。与此同时，计算机软件系统也会随之升级，从而使计算机的信息处理和存储功能在短期内得以迅速提升，单位信息存储的成本大幅下降。当 IBM 于 1955 年推出第一款商用硬盘存储器时，其价格是 6000 多美元/兆，1960 年下降到 3600 美元/兆，1993 年大概约为 1 美元/兆，2000 年再降至 1 美分/兆，截至 2010 年则约为 0.005 美分/兆。而自 1977 年美国芝加哥率先投入使用光纤通信系统以来，光纤传输带宽实现迅猛增长，其信息传输能力也得到大幅跃升，甚至超越了摩尔定律下芯片性能的提升速度。信息基础设施的持续完善，包括数据存储性能的不断提升、数据传输带宽的持续增加，为大数据的存储和传播提供了物质基础，使数据信息的大

规模存储、传输与分析得以实现。目前，硬件存储性能与网络带宽不再是制约大数据应用的主要因素，且它们的高速发展将持续为大数据时代提供廉价的存储与传输服务。

（二）互联网领域的发展

人与人之间的交流沟通由于互联网的出现得到极大便利，互联网的广泛运用改善了人们的日常生活，并逐渐渗透到人们生活的方方面面。当人们在互联网的海洋里徜徉时，会留下海量的数据，越来越多的重要数据保存在无数台计算机上。为了保证数据存储的安全与数据传递的高效，要求计算机之间相互传递数据、互为备份的通信机制具有更高的性能标准。目前，在使用互联网数据时，一般都是通过“请求 + 响应”的模式，即只有在客户端发出请求的情况下，服务器终端才会发送所需要的数据。这种数据传递模式在一定程度上保证了数据传递的安全和高效，也使得人们在使用网络时的每一个搜索请求、每一个访问请求、每一个交易记录等数据信息都被忠实准确地记录在各类服务器的日志上。互联网的广泛普及积累了巨量的数据信息，使大数据分析过程中的数据采集成为可能，也大大降低了数据采集的成本，提高了数据信息记录的真实性和可靠性。

二、产业升级

从哲学意义上说，世界处于永续变动之中，万事万物在其运动过程中产生了大量的数据信息。近年来，随着互联网、云计算、物联网等信息技术的飞速发展，各行各业的产业结构不断升级，这无时无刻不产生海量的数据，形成大数据雏形。目前，我国经济本质上仍处于传统经济的阶段，缺乏具有国际竞争力的现代产业，产业结构升级已经迫在眉睫，这无疑为大数据的滋生提供了肥沃的土壤。

随着互联网的普及、信息技术的进步及电子化时代的到来，人们以更快捷、更容易、更廉价的方式获取和存储数据，使得数据及信息量以指数方式增长。据粗略估计，一个中等规模企业每天要产生 100MB 以上的商业数据。而电信、银行、大型零售业随着产业结构的不断调整和升级，每天产生的数据量都可以用 TB 来计算（数据的最小计量单位是字节 Byte，具体换算标准为：1KB=1024B；1MB=1024KB；1GB=1024MB；1TB=1024GB；1PB=1024TB；1EB=1024PB；1ZB=1024EB；1YB=1024ZB；1BB=1024YB；1NB=1024DB；1DB=1024NB）。据《数字中国发展报告（2021 年）》，2021 全年我国数据产量达 6.6ZB，已位居全球第二。产业结构升级所带来的数据越来越多，激增的数据背后隐藏着许多重要的信息，人们希望对其进行更高层次的分析，以便更好地利用这些数据。现有的数据库系统虽然拥有高效地完成数据的输入、统计、查询等功能，却不能发现数据中的关系与规则，不能在现有数据的基础上推断今后的发展趋势。大数据技术背后隐藏的知识手段的不足使得“数据爆炸但知识匮乏”这一现象浮现出来。自此人们纷纷提出“学会选择、提炼、舍弃信息”，并思考怎样才能不被海量的信息淹没，怎样才能及时发现有用的知识、提高信息利用效率？如何从浩如烟海的资料中选择性地搜集有价值的信息？这为数据分析带来了一些挑战：第一是信息过量，难以消化；第二是信息真假难以辨别；第三是信息安全难以保

证；第四是信息形式不一致，难以统一处理。为应对这些挑战，计算机数据仓库处理技术随之走向成熟，从数据中发现知识及其核心技术——大数据技术便应运而生，并得以蓬勃发展，显示出越来越强大的生命力。

三、社会进步

1998 年，《科学》杂志刊登的一篇名为“大数据的处理程序”的文章中第一次明确使用了大数据（big data）一词。2008 年 9 月，《自然》杂志刊登了名为“Big Data”的专题，“大数据”概念开始受到广泛关注。大数据的产生和发展有其特定的社会历史发展脉络。其实大数据存在的历史非常悠久，“大数据”概念的提出标志着人们已经开始意识到大数据的客观存在，而且已经感受到了大数据应用的重要性。

各种各样的海量数据构成了大数据的基石。悠久的社会历史文化为大数据的产生提供了充足的时间条件。从人类历史发展脉络来看，数据的产生与人类自身的生存、生活密切相关，也正是这种内在需求促使数据发展为大数据。从数据的观察到数据的收集和使用，到处都是功利主义的性质。大数据分析是一种非常实用的技术，古希腊的哲学家真正让数据从实用走向抽象。哲学家们第一次抛弃实用主义的桎梏，把数据当作事物的本源，这种独特的思维模式为自然哲学的研究开辟了一条崭新的道路，也为大数据的诞生奠定了哲学历史基础。纵观数据的发展历史，数据和其他语言文字一样，都是人类文明的产物，是用于记录事物性质和互相交流的工具。从广义上看，数据可以被看作语言的一部分，但与文字语言的差别在于数据的表达形式更简单、更加有利于交流。所以虽然不同人类文明有着不同的记数方式和数制，但随着不同文化的相互交流融合，数据形式的高度统一超出了所有文字语言，这离不开数字的简单精确的属性。回顾科学技术的发展史，科学技术的迅猛发展离不开科学数据的支撑，科学数据具有客观性、精确性、一致性和易交流性等特征。所以说，数据不仅是连接事物客观性和人类主观性的纽带，还是人类认识世界的桥梁。但从数据产生的那一刻起，人类主观因素无时无刻不在影响着数据的客观性。大量数据构成的集合形成了一种重要的研究素材，激发着科学家和哲学家们进行深入研究，他们在研究过程中越发意识到数据的重要性，所以大数据便应运而生。

在这里，我们简要介绍一下数据科学的发展历史。

20 世纪中期以来，生物学领域的基因组测序技术发展迅猛，累积了海量的生物学数据，如何理解这些数据，是生物学家们面临的一项新挑战。同样的数据分析问题也存在于其他领域（如气象学、社会学等）和复杂系统的研究之中。值得注意的是，国际科技数据委员会（下称 CODATA）于 1966 年成立，旨在提升数据的质量、可信度、可达性，并加强对数据的管理，从而在世界范围内实现共享科技数据的目标。1984 年 6 月，中国科学院以国家会员的身份加入 CODATA。

基于数据的相关研究已得到学术界的广泛关注。数据科学是一门以大量观测数据、理论数据和计算机模拟数据为研究对象，通过挖掘、提取等手段寻求其内在规律的学科。1974 年，彼得·诺尔（Peter Naur）首次提出“数据科学”（data science）这一术语。1996 年，在日本

东京召开的分类国际联合会（The International Federation of Classification Societies，IFCS）上，第一次将数据科学用于会议题目——数据科学，分类和相关方法（Data Science，Classification and Related Methods）。美国普渡大学统计学教授 William S. Cleveland 于 2001 年首次倡导应该将数据科学建设成一门独立的学科，他认为数据科学是统计学与数据的结合，并建立了数据科学的六个细分技术领域：多学科研究、数据模型和方法、数据计算、教育、工具评估、理论。

2001 年，CODATA 创办了学术刊物 *CODATA Data Science Journal*，标志着数据科学的诞生。2003 年，由中美两国学者共同创办的 *Journal of Data Science* 在哥伦比亚大学正式出版，*Journal of Data Science* 主要发表一些关于数据的研究成果，如数据的搜集、分析及建模等。

2012 年，Springer 出版集团创建了期刊 *EPJ Data Science*。该期刊的主办方认为，21 世纪出现的“数据驱动科学”是传统“假说驱动科学”研究方法的重要补充。数据科学的出现促进了科学研究范式的变革。利用电子计算机，在对密集型数据进行深度挖掘后获取有用信息，由此催生了不同学科领域的新的研究方向，如生物信息科学、地理信息科学等。这种发展伴随着科学范式从“还原主义”到“复杂系统”的转变，不仅极大地丰富了自然科学的研究范式，而且对技术—社会—经济科学研究也产生了非常重大的影响。

学者们从超级计算、互联网经济、生物医药等多个方面重视“大数据”引发的技术挑战及今后的发展趋势。2010 年，Bollier 提出计算机存储技术、产生数据流的设备（如望远镜、摄像机及交通监视设备）、云计算、面向消费者的应用（如 google earth 和 map quest）等成为大数据产生的几个重要因素，并首次提出“一种新的知识基础设施正在实现，大数据时代正在出现”的观点。

第二节　大数据的发展进程

“大数据”一词来源于英文“Big data”，其概念起源于美国。“大数据”最早在统计领域得到应用，并在计算机通信领域引发了一场革命，随后蔓延至经济、社会、科学、环境等各个领域，并成为现代国家发展战略的重要组成部分。在互联网热潮的推动下，“大数据”技术迅速渗透到人们生活的方方面面，吸引着人们的目光。

一、大数据的主要发展阶段

大数据发展的主要阶段如图 1-1 所示。

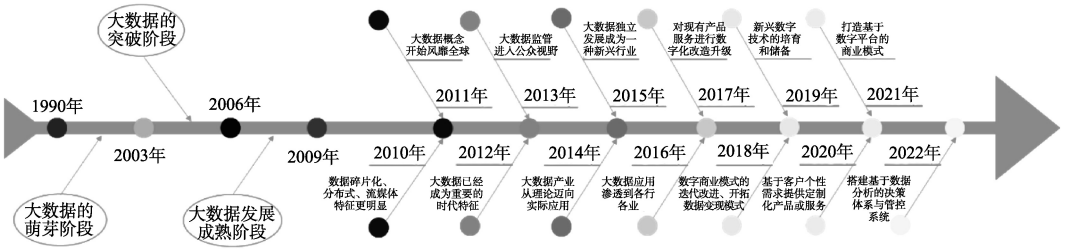

图 1-1　大数据发展的主要时间轴

20 世纪 90 年代，随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术（如数据仓库、知识管理系统等）开始被广泛应用，大数据概念开始萌芽。此时，关于大数据的相关研究主要聚焦于数据挖掘技术方面，其他方面涉及较少。

2003—2006 年是大数据发展的突破阶段，是非结构化数据的自由探索阶段。2004 年脸书网（Facebook）的创立使得大量非结构化数据涌现，大数据技术的快速突破得益于非结构化数据的爆发。

2007—2009 年为大数据发展的成熟期，大数据技术并行运算与分布式系统基本形成。

2010 年，随着智能手机的应用日益广泛，数据的碎片化、分布式、流媒体特征更加明显，移动数据量急剧增长。

2011 年，麦肯锡全球研究院发布《大数据：下一个创新、竞争和生产力的前沿》研究报告。之后，经 Gartner 技术炒作曲线和 2012 年维克托·舍恩伯格《大数据时代：生活、工作与思维的大变革》的宣传推广，大数据概念开始风靡全球。

2012 年，《大数据研究和发展提倡》的发布标志着大数据已经成为时代特征，这一倡议也意味着大数据从商业行为上升到国家科技战略这一更高层面。联合国在纽约总部发布了一份与“大数据政务”相关的白皮书，概括性地阐述了各国政府如何利用大数据更好地服务和保护人民，实现“与数俱进”，快速应变。

2013 年，“大数据”监管进入公众视野。中国证券监督管理委员会（简称为“中国证监会”）利用各个异动指标数据，将交易异常数据和股价异动联系起来，构建证券市场监控的综合数据模型，全面提升对内幕交易、市场操纵、证券欺诈等文本信息的挖掘和监管。大数据也成为政府监管对象之一。

2014 年，云计算的爆发推动智能科技加速发展，大数据产业从理论迈向实际应用。2014 年 12 月，中国计算机协会等发布了《中国大数据白皮书》，第一次全面深入且系统完整地阐述了我国大数据产业发展与学术研究的大方向，从国家主权、政府政策、产业发展、数据科学、投资理念、公司战略等层面分析了我国大数据市场当前发展现状及未来发展趋势，这是我国大数据行业逐步迈向产业化、系统化的重要一步。

2015 年，大数据逐步迈向独立发展阶段，其市场化和规模化程度进一步提升，已经成为一种新兴行业，数据租售服务大量出现，数据分析企业更加专业化，数据决策外包服务企业更加高效，推动更多传统企业向科技智能化转型。

2016 年后，大数据应用渗透到各行各业，大数据价值不断凸显，数据驱动决策和社会智能化程度大幅提高，大数据产业迎来快速发展和大规模应用实施。

2017—2022 年，中国数字经济规模从 31.3 万亿元扩大到 45.5 万亿元，占 GDP 的比重从 34.8%上升到 39.8%。除了规模的快速扩大，数字技术的创新及应用还引发了产业形态和生产组织方式的深刻变革。六年间，数字化浪潮席卷各行各业，行业边界日益模糊，颠覆与创新成为常态，新可能、新机遇不断涌现。借力数字技术重塑业务、拓展边界，是企业

基业长青的不二选择。

当前，地缘政治冲突仍存在不确定性，全球主要经济体的发展都已经进入放缓阶段。2019 年全球大数据市场规模为 496 亿美元，同比增长 17.3%，2021 年全球大数据市场规模已达到 649 亿美元，2019—2021 年的复合增长率约为 14.4%。我国大数据市场规模则由 2019 年的 619.7 亿元人民币增长到 2021 年的 863.1 亿元人民币，复合年增长率达到18.0%。面对“俄乌冲突”的不确定性、持续的供应链挑战及不断上升的通胀率和利率，我国在大数据市场中率先复苏并总体保持恢复态势。根据预测，全球整体市场规模有望在 2024 年超过 800 亿美元，2019—2024 年的复合增长率约为 11.8%[①]。

二、主要国家的发展现状

19 世纪 80 年代，“大数据”概念开始萌芽。1887—1890 年，美国统计学家赫尔曼·霍尔瑞斯为了统计 1890 年的人口普查数据，发明了一台电动器，将原本耗时 8 年的人口普查活动缩短为 1 年，由此在全球范围内引发了数据处理的新纪元。1944 年，卫斯理大学图书馆管理员弗莱蒙特·雷德预见了大数据时代的到来，他估计美国高校图书馆的规模每 16 年就会翻一番。1961 年德里克·普赖斯通过观察科学期刊和论文的增长规律来研究科学知识的增长，得出“指数增长规律”，即新期刊数量的增长方式为指数型而非线性型，每 15 年翻一番，每 50 年以 10 为指数成倍增长。这些规律发现都表明数据知识未来将呈爆炸式增长，大数据时代即将来临。

在信息通信领域，数据的大规模增长与存储首先引发关注。1980 年 4 月，I. A. 特詹姆斯兰德在第四届美国电气和电子工程师协会（IEEE）“大规模存储系统专题研讨会”上发表了题为“我们该何去何从？”的报告，指出所有数据都正在被无选择地保存下来以避免丢失有价值的信息。1986 年 7 月，哈尔·B. 贝克尔在《数据通信》上发表了《用户真的能够以今天或者明天的速度吸收数据吗？》一文，预计数据记录密度将大幅增长。1997 年 10 月，迈克尔·考克斯和大卫·埃尔斯沃思在文章《为外存模型可视化而应用控制程序请求页面调度》中较早使用了“大数据”这一术语。2001 年，美国一家在信息技术研究领域具有权威地位的咨询公司 Gartner 首次开发了大数据模型。同年 2 月，梅塔集团分析师道格·莱尼发布了题为“3D 数据管理：控制数据容量、处理速度及数据种类”的研究报告，文中提及的 3V 特征定义了大数据的三个维度，如今被广泛接受。从此，“大数据”这一概念在信息通信领域被普遍接受、研究和使用。

Nature 杂志于 2008 年推出的一份专刊，从互联网科技、自然与环境、网络经济和金融等多个方面介绍了海量数据带来的挑战。2008 年年末，一些著名的美国计算机科学研究者开始认同“大数据”这一概念。业界组建起计算社区联盟，并发布了一份影响深远的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。该白皮书使得大众对大数据

① 数据来源：Wikibon。

的解读发生了显而易见的变化——从单一的数据处理机器这一角度扩展到了商业、科学、社会等各个领域，极大地丰富了“大数据”的内涵和价值，而计算社区联盟也因此被称为最早提出“大数据”概念的机构。2011 年 5 月，全球知名咨询公司麦肯锡发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，该机构第一次全方位地介绍和展望大数据，具体论述了大数据的应用价值与发展前景，“大数据”概念逐渐进入公众视野。

（一）美国

作为大数据概念发源地的美国，不仅在大数据理论研究方面引领全球风潮，也在大数据应用实践中占得先机。2009 年，美国政府通过启动 data.gov 网站的方式进一步开放了数据的大门，这个网站向公众提供各项政府数据。

2010 年 1 月，美国总统办公室下属的科学技术顾问委员会（PCAST）和信息技术顾问委员会（PITAC）提交了一份《规划数字化未来》的战略报告，第一次把大数据收集和使用的工作提升到体现国家意志的战略高度。在 2012 年美国总统选举中，竞选团队开创性地利用“大数据”来规划这次选举。例如，利用房产记录、选举记录，甚至是期刊的订阅注册等预测人们对候选人的看法、这些看法是否能被改变，以及为此要采取怎样的措施等。这次实践充分说明了大数据的潜在价值只有通过数据挖掘才能显现。由此可见，如何有效分析和利用巨大的原始数据，将其转化成有价值的信息，成为日后研究大数据的重要意义。2012 年 3 月，奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。3 月 22 日，奥巴马政府宣布 2 亿美元投资大数据领域，是大数据技术从商业行为上升到国家科技战略的重要标志。美国政府将数据定义为“未来的新石油”，表明了大数据技术领域的竞争事关一个国家未来的发展。在此基础上，于 2016 年 5 月，美国政府进一步发布了《联邦大数据研究与开发战略计划》，为后续战略的完善奠定了基础。特朗普执政后，美国加快了数据战略布局，2019 年 12 月，美国联邦数据战略团队发布了《联邦数据战略》，该战略描述了美国 2020 年以后十年的数据愿景，与之一起发布的还有推进这一战略的《2020 年行动指南》。自此，美国数据战略体系已经初具雏形。

虽然大数据应用的广阔前景引发了广泛关注，但在实际应用中如何科学、规范、公正地使用大数据也成为各相关主体议论的焦点。2014 年 5 月，美国白宫发表的 2014 年全球“大数据”白皮书研究报告《大数据：抓住机遇、守护价值》指出，应当鼓励利用数据来促进社会进步，同时，还需要制定相应框架结构来保护个人隐私、反歧视或保证公平。随后，美国政府在 2016 年发布的《联邦大数据研发战略计划》中也提到，为了强化数据驱动的体系和能力建设，应当形成涵盖技术研发、数据可信度、基础设施、数据开放与共享、隐私安全与伦理、人才培养及多主体协同七个维度的系统顶层设计，打造面向未来的大数据创新生态。2020 年 10 月，美国国防部发布《国防部数据战略》，明确将数据定位为战略资产，提出“使国防部成为以数据为中心的机构，通过快速规模化使用数据来获取作战优势和提高效率”的发展愿景，并提出了推进数据发展的七大目标、八大指导原则、四个基

本能力和三个重点应用领域。

（二）法国

2011 年 7 月，法国启动“Open Data Proxima Mobile”项目，挖掘公共数据价值。该项目旨在通过实现公共数据在移动终端的使用，最大限度地发掘公共数据的应用价值。该项目涉及文化、旅游、环境、交通等多个领域。待结项后，所有的公共数据都可以免费使用，法国公民及在法国旅游的欧洲公民都将能使用个人移动终端获得法国的公共数据。应用程序操作简单，不仅方便公民使用，而且为私人企业提供巨大商机。2011 年 12 月，法国政府推出公开信息线上共享平台 data.gov，该网站的所有数据都经过政府部门专员统计和收集，后期会不断实时更新。这个平台大大方便了公民自由查询和下载公共数据。

2013 年 2 月，法国政府发布《数字化路线图》，明确了未来将大力支持大数据这一战略性高新技术。法国政府将以工程师、信息系统设计师、新兴企业、软件制造商为主体，进行一系列投资计划。2013 年 4 月，法国投入专项资金促进大数据技术发展。法国经济、财政和工业部预计投入 1150 万欧元投资七个项目，目的在于通过创新性解决方案确保法国在大数据领域的领先地位。同年 7 月，发布《法国政府大数据五项支持计划》，制定了引进数据科学家教育项目、设立大数据发展资金等举措，为大数据建立良好的生态环境。

2022 年 3 月，马克龙在一次介绍其竞选纲领的活动中提到，“将为建设欧洲的元宇宙而战”。这句雄心勃勃的口号某种程度上体现着马克龙在数据领域布局的决心。2022 年 1 月，LVMH 集团任命曾在丝芙兰担任数字产品总监的奈莉·门萨为数字创新副总裁兼元宇宙负责人，3 月 LVMH 便发布了自己在元宇宙中的虚拟形象大使。此外，她还计划在 6 月中旬巴黎举行的 Viva Technology 大会上亮相，为 LVMH 集团及其旗下品牌发表有关数字创新的声明。

（三）德国

德国在大数据发展早期重点关注的是数据保护，通过立法为大数据的发展提供安全保障。1977 年，德国联邦层面的《数据保护法》生效。德国凭借自身较高的信息化水平，通过大型基础数据库和地方数据库的建设，逐渐在政府管理中运用数据资源服务公众和服务决策。对政府管理而言，大数据的价值在于提供尽可能多的详尽信息，并对信息进行有效分析，实现科学化决策和精细化管理。

2000 年德国发布了《2005 年联邦政府在线计划》，该计划要求联邦政府到 2005 年向公众提供所有可用的在线服务。2003 年 6 月，德国启动了整合电子政务的“德国在线”计划，加强基础数据库及地方数据库的建设力度，整合大量分散的信息资源，以公众需求为导向，为公众提供更方便的数据服务。

2004 年生效的德国《电信法》也涉及电子通信领域的数据保护。2006 年，德国开始将其拥有的 GESTIS 等七个有毒有害物质官方数据库及本国气候变化预测图免费公开。德国电信和 Vodafone 也通过开放 API 的方式，向数据挖掘公司等机构提供一些客户的匿名定位

数据，从而捕捉公众出行的特征和规律。德国在云计算与大数据技术的支持下发展人工智能技术，研发本国的“谷歌眼镜”“智能农场”“交通监测”等技术。2009 年，德国对现行的《联邦数据保护法》进行修改并生效，约束范围包括互联网等电子通信领域，目的是防止因个人信息泄露而引发的侵犯隐私行为。政府内部须设立“联邦数据保护与信息自由专员”，实时监控政府机构在保护个人数据方面的行动。除了联邦层面，德国各州也都设立了各自的数据保护专员，以类似的方式监督各州政府机构的行为。

同时，德国也重视信息资源共享。例如，2013 年 1 月，为了改善教学和科研中的数字信息支撑水平，德国科学组织联盟启动了第二期“数字信息计划”，该计划主要包括以专业的信息科学与信息技术方法实现科研数据的搜集、存储和开放共享，确保用于科研目的的科研数据不受访问限制，实现数字出版物的永久保存等。

2011 年，德国在汉诺威工业博览会上首次提出了“工业 4.0”概念，2013 年，德国联邦教研部与联邦经济和技术部正式将“工业 4.0”战略纳入了“高技术战略 2020”。德国认为，工业革命可以分为四个阶段，第三次工业革命引入了电子与信息技术，在此基础上，如果德国可以广泛地将物联网和服务网应用于制造领域，在智能工厂中实现数字和物质两个系统的无缝融合，德国就可以在第四次工业革命的道路上占领先机，巩固德国的竞争地位。德国“工业 4.0”战略打出“确保德国制造业的未来”的口号，以期将信息化与工业化紧密结合起来。还于 2014 年 8 月 20 日通过了《2014—2017 年数字议程》，提出在变革中推动“网络普及”“网络安全”“数字经济发展”三个重要进程，希望以此打造具有国际竞争力的“数字强国”。

在该政策带领下，2021 年 6 月，德国数据挖掘和 RPA 创业公司 Celonis 获得 10 亿美元 D 轮融资，估值达 110 亿美元，为其两年前估值的四倍多。据外媒 TechCrunch 报道，总部位于德国柏林的初创公司 LiveEO 专注于为交通、能源企业处理卫星影像原始数据的数据分析，于 2022 年宣布获得 1900 万欧元的新一轮融资，由 MMC Ventures 领投。除了风险投资，该公司还获得了来自 European Commission 和 Investitionsbank Berlin 两家公共机构的资金支持。

（四）中国

全球大数据技术发展的浪潮引起我国政府部门、商业企业和学术界的高度关注，政府也将大数据发展提升到国家战略的高度。2011 年 12 月，工信部发布的《物联网“十二五”发展规划》提出了信息处理技术，确认了其为 4 项关键技术创新工程之一的战略地位，其中包括了数据存储、数据挖掘、图像视频智能分析等，这些构成大数据的坚实基础。2012 年 4 月，政府推出《软件和信息技术服务业“十二五”发展规划》，积极发展数据编辑、整理、分析、挖掘等数据加工处理服务，可见政府高度重视大数据的应用，将其与国家发展规划联系在一起。2015 年 6 月 24 日，国务院办公厅发布了《国务院办公厅关于运用大数据加强对市场主体服务和监管的若干意见》，将大数据技术应用于市场主体的服务和监管，推进简政放权和政府职能转变，提高政府治理能力。

2012 年 7 月，阿里巴巴集团率先设立了“首席数据官”一职以挖掘大数据的商业价值，负责全面推进“数据分享平台”战略，并推出大型的数据分享平台“聚石塔”，为淘宝、天猫平台上的电商和电商服务商等提供数据云服务。其后，马云在 2012 年网商大会上发表演讲时称，自 2013 年 1 月 1 日起，阿里巴巴将转型重塑数据、金融和平台三大业务，因此其成为第一家提出利用数据进行企业数据化运营的企业。

国内学术界也紧跟国际前沿，广泛开展大数据技术的研究和开发。2012 年中国计算机学会（CCF）发起并组织了 CCF 大数据专家委员会，还特别成立了一个“大数据技术发展战略报告”撰写组，并于 2013 年、2014 年相继发布了《中国大数据技术与产业发展白皮书》。2013 年 10 月，“第十七次全国统计科学讨论会”开幕，其主题就是大数据背景下的统计。2013 年以来，国家自然科学基金、国家重点基础研究发展计划、国家高技术研究发展计划等重大研究计划都已经把大数据研究列为重大研究课题。2014 年 2 月在北京召开了以“科研大数据与数据科学”为主题的“科学数据大会”，探讨了大数据时代下数据的分析和应用，以及科研数据带来的挑战和机遇。2014 年 3 月，国家社会科学基金也将“大数据国家战略研究”列为国家社科重大研究项目指南。清华大学信息学院、清华信息科学与技术国家实验室也成立了清华大学数据科学院，并于 2014 年 12 月 22 日举办了“大数据论坛——数据科学与技术”，对大数据发展战略和各大数据专项进行了探讨。2015 年 5 月 26 日，“2015 贵阳国际大数据产业博览会暨全球大数据时代贵阳峰会”（简称数博会）在贵阳开幕。2017 年，数博会正式升格为国家级博览会，成为探讨大数据行业发展现状和趋势的世界级平台。

与此同时，“大数据”也逐步走进公众的视野。2013 年 4 月 14 日和 21 日，央视著名节目“对话”分别邀请了美国大数据存储技术公司 LSI 总裁阿比和《大数据时代——生活、工作与思维的大变革》的作者维克托·迈尔·舍恩伯格做了两期大数据专题谈话节目——《谁在引爆大数据》与《谁在掘金大数据》。官方媒体对大数据的关注和宣传充分体现了大数据技术已经成为国家和社会普遍关注的焦点。

2015 年 8 月 31 日，国务院发布《促进大数据发展行动纲要》，提出要系统部署大数据发展工作，重点推进大数据在多个领域的应用，利用大数据等新技术打造服务贸易新型网络平台。同时，要强化数据安全保障，提高管理水平，促进大数据产业的健康发展。

2016 年 4 月，为加快实施国家大数据战略，促进区域性大数据基础设施的整合和数据资源的汇聚应用，发挥示范带动作用，国家发展改革委、工业和信息化部、中央网信办函复贵州省人民政府，同意贵州省建设国家大数据（贵州）综合试验区。10 月 8 日，三部门发函批复，同意在京津冀等七个区域推进国家大数据综合试验区建设，包括两个跨区域类综合试验区（京津冀、珠江三角洲），四个区域示范类综合试验区（上海市、河南省、重庆市、沈阳市）及一个大数据基础设施统筹发展类综合试验区（内蒙古自治区）。大数据战略已经上升为国家战略的高度。同年 12 月，为了落实国务院《促进大数据发展行动纲要》，按照《中华人民共和国国民经济和社会发展第十三个五年规划纲要》的总体部署，国家工业和信息化部发布《大数据产业发展规划（2016—2020 年）》促进国家大数据产业持续健

康发展，进一步推动实施国家大数据战略。

随着 5G、云计算、人工智能等新一代信息技术的迅速发展，信息技术与传统产业加速融合，数字经济蓬勃发展。2017 年，党的十九大报告提出“推动大数据与实体经济深度融合”来加强数字基础设施建设，通过智能化、协同化的新生产方式对实体经济进行改造升级，全面提高实体经济的质量、效益和竞争力，打造数字经济形态下的实体经济，体现了经济发展的方向，进而推动经济体系优化升级，并在随后的几年，推动大数据融入人工智能、数字经济、数字治理等政策体系。

2020 年 4 月，《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》发布，引导各类要素协同向先进生产力集聚，推动经济发展质量、效率和动力的变革，明确了数据要素市场化配置上升为国家政策。2021 年 3 月 31 日，北京国际大数据交易有限公司成立，这是国内首家基于“数据可用不可见，用途可控可计量”新型交易范式的数据交易所。同年 11 月 25 日，上海数据交易所揭牌成立并达成了首单交易。全国各地将陆续设立数据交易机构作为促进数据要素流通的主要抓手。

2021 年 11 月，为了深入落实《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》对“打造数字经济新优势”的决策部署，和对大数据产业发展提出的新要求。国家工信部制定出台《“十四五”大数据产业发展规划》，提出实现数字产业化和产业数字化的有机统一，不断完善大数据体系建设，并以此作为未来五年大数据产业发展工作的行动纲领。释放数据要素价值和保障数据安全是产业发展的重点。

2023 年 2 月 27 日，中共中央、国务院印发了《数字中国建设整体布局规划》，强调要促进数字经济和实体经济的深度融合，以数字化驱动生产生活方式和治理方式变革。要全面赋能经济社会发展，做强做优做大数字经济，推动数字技术和实体经济深度融合。

近年，我国网民规模和互联网普及率稳定上升，至 2021 年互联网普及率为 73%，比上一年增长 2.6 个百分点。随着信息技术的创新，互联网的普及，数据量会不断扩大。据 IDC 统计，至 2021 年全球所产生的数据量已达到近 70ZB（1ZB = 1024 × 1024 × 1024TB），预计 2025 年将达到 175ZB。届时，大数据将在行业变革中承担更重要的角色。

第三节　大数据的概念界定

究竟何为大数据？“大数据”一词可以从字面上理解为“巨大的数据量”。Manyika 等认为，大数据是指数据的集合，其大小已经超出了现有典型数据库获取、存储、管理和分析数据的能力。达到什么程度的数据才可以叫作大数据？目前尚未形成一个普适性的定义。一般认为，大数据的量级应该是太字节，即 2 的 40 次方。当数据规模非常巨大，且达到某种程度时，会使数据呈现出某些有价值的特性，而由于数据体量较大，这些特性无法通过传统的数据处理技术进行归纳分析，需要新的技术进行挖掘与分析。因此，大数据不仅是指规模巨大的数据，而且是一种分析处理庞大数据的技术。涂子沛在其《大数据》一书中指出，“大数据”是指一般的软件工具难以捕捉、管理和分析的大容量数据，以太字节为单

位。“大数据”之大，不仅在于容量之大，更深层次的意义在于：因为人类分析和使用的数据量呈爆炸式增长，通过对海量大数据的交换、整合、挖掘和分析，可以发现新的知识，创造新的价值，由此带来“大知识”“大科技”“大利润”和“大发展”。

本节将从理论、技术、实践三个层面具体论述大数据的概念，如图 1-2 所示。

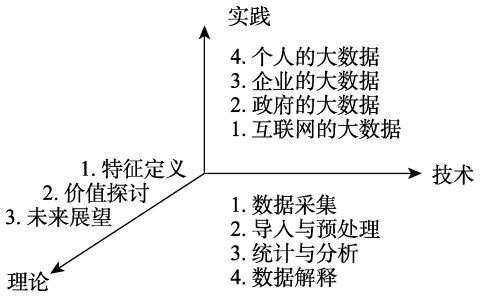

图 1-2　大数据概念的层次

一、理论层面

目前尚未有权威机构对大数据的概念进行统一界定，因此存在多个版本的定义。

- **John Rauser 亚马逊网络服务（AWS）大数据科学家**

大数据就是任何超过了一台计算机处理能力的庞大数据量。

- **麦肯锡**

大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合。

- **维基百科**

巨量资料，或称大数据，是指所涉及的数据量大到无法通过目前主流软件工具，在合理时间内达到撷取、处理并整理成为促进企业经营更积极的资讯。

- **研究机构 Gartner**

“大数据”是使用高效的信息处理方式以具备更强的决策力、洞察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”是指无法使用传统流程或工具处理或分析的信息。它定义了迫使用户采用非传统处理方法处理的超出正常处理范围及大小的数据集，其价值在于提高数据使用者的最终决策力，如图 1-3 所示。

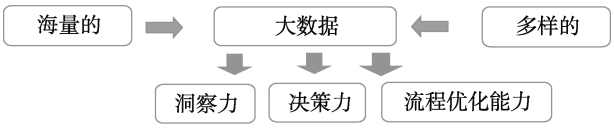

图 1-3　大数据的定义

- **互联网数据中心（IDC）**

大数据是为更经济地从高频率的、大容量的、不同结构和类型的数据中获取价值而设计的新一代架构和技术。

- **互联网周刊**

“大数据”的概念远远超过了海量数据及处理数据的技术，或是类似的“4 个 V”的简单理解，而是涵盖了人们只有基于大规模数据才能够做的事情，这些在小规模数据的基础上是根本无法实现的。也就是说，大数据让我们以一种前所未有的方式，通过对大规模数据进行分析，获取有巨大价值的产品及服务，或深刻的洞见，最后形成变革之力。

- **《大数据时代的历史机遇——产业变革与数据科学》**

“大数据”是在多样的或者大量数据中迅速获取信息的能力。前面几个定义都是从大数据本身出发，我们的定义更关心大数据的功用，它能帮助大家干什么。在这个定义中，重心是“能力”。大数据的核心能力是发现规律和预测未来。

上述定义基本上都是基于大数据内涵本身，但在现实中，更重要的是大数据的价值与应用。因此下文将从大数据的定义、特征出发，了解各行各业对大数据的整体描绘和定性分析，挖掘大数据的独特价值，洞悉大数据的未来发展趋势，并从数据安全的角度重新审视数据的合理有效使用问题。

（一）定义

作为数据，大数据具备三种特点：一是广泛存在性，即绝大多数产品与行为均可产生数据，是否记录主要取决于技术能力与成本考量；二是非独占性，即数据可被多次使用，尤其是公开的数据可以被其他人所使用；三是多认知性，即根据使用者的不同，同样的数据会产生不同的理解和使用方式。

而“大”为之带来的特点则是体量巨大，处理速度较快、数据类型多样、商业价值高和在线化。2001 年 2 月，梅塔集团分析师道格·莱尼发表了《3D 数据管理：控制数据容量、处理速度及数据种类》的研究报告，对大数据提出“3D 数据管理”的看法，即数据成长将朝 3 个方向发展，分别为数据即时处理的速度（velocity）、数据格式的多样化（variety）与数据量的规模（volume），被归纳为“3V 特征”。之后，随着资讯科技的进步，数据量的复杂程度越来越高，“3V”已经不足以形容新时代的大数据，因此，2012 年，莱尼提出调整现有的 3V 分析框架，此外，高科技公司 IBM、国际调查机构 Gartner、互联网数据中心（IDC）等纷纷对大数据提出新的论述，在原本的速度、多样化与规模三个特征上，增加价值性（value）和在线的（online）等特色。

大数据的五个特征联系紧密、协同交替，如图 1-4 所示。

第一，数据体量巨大。一般数据库的大小在 TB 级别，而大数据的起始计量单位在 PB（1PB = 1024TB）级别，有的甚至跃升至 EB、ZB 级别，包括采集、存储和计算的量都非常大。百度资料表明，其新首页导航每天需要提供的数据超过 1.5PB，这些数据如果打印出来，将超过 5 千亿张 A4 纸。有资料证实，到目前为止，人类生产的所有印刷材料的数

据量仅为 200PB。

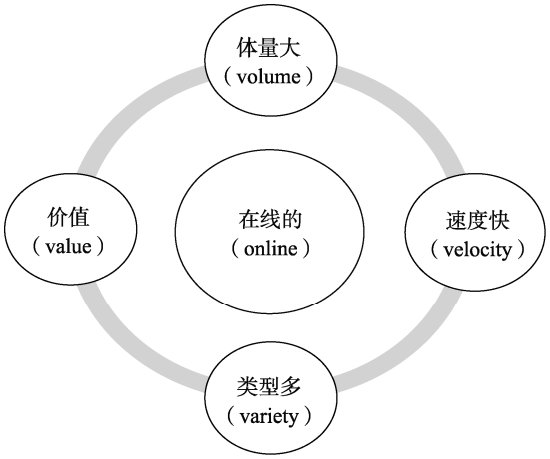

图 1-4　大数据的特征

第二，数据类型繁多，有结构化、半结构化及非结构化数据。具体表现为图片、地理位置信息、网络日志、视频、音频等，个性化数据占绝对多数。多类型的数据对数据的处理能力的要求更高，已冲破了之前所限定的结构化数据的范畴。

第三，处理速度快，在数据量非常庞大的情况下，也能够做到数据的实时处理与分析，这与传统的数据挖掘技术有着本质的不同。数据体量的增大也就对数据的处理速度、时效性提出了更高的要求，例如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。而大数据技术正好能满足这一需求，这也是其区别于传统数据挖掘的显著特征。

第四，价值密度低。随着互联网及物联网的广泛应用，信息感知无处不在，大量信息的价值密度很低，即尽管数据量大、类型多、处理速度快，但真正有价值的数据却很少。以视频为例，一小时的视频，在持续不间断的监控过程中，有价值的数据可能只有一两秒。如何结合业务逻辑并通过强大的机器算法来挖掘数据价值，是大数据时代最需要解决的问题。

第五，数据是在线的，即是随时能调用和计算的，这是大数据区别于传统数据最大的特征。在互联网高速发展的背景下，数据资源不仅仅是体量大，更重要的是表现出在线这一显著特征。数据只有在线，即数据在与产品用户或者客户产生连接的时候才有意义。例如，用户在使用某互联网应用时，其行为能够及时地传给数据使用方，数据使用方通过数据分析或者数据挖掘进行加工，对该应用的推送内容进行优化，把用户最想看到的内容推送给用户，以提升用户的使用体验。

此外，业界还有人总结出大数据的其他特征，如数据准确性高，随着社交数据、商业交易与应用数据等新型数据源的兴起，企业越来越需要有效的信息来确保其真实性及安全性；存活性低是指特定情况下的大数据具有很强的时效性。

与传统数据服务相比，大数据服务拥有企业内部、外部及市场环境等不同来源的海量数据，通过传感器采集、互联网抓取等方式获取。快速发展的分布式计算及多样的数据分析模型使海量数据处理成为可能（图 1-5）。

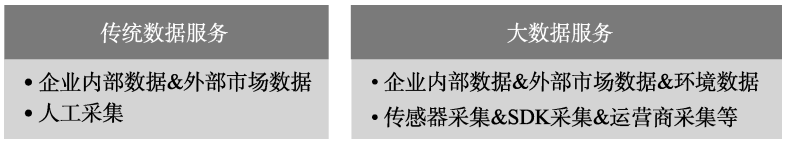

图 1-5　传统数据服务与大数据服务的不同

（二）特征与内涵

随着大数据时代的来临，大数据技术开始广泛应用于越来越多的领域，但只有了解大数据的价值，了解大数据究竟会如何改变生活，才能更好地利用大数据。因此，需要结合时代与社会背景来具体分析大数据，理解它如何在时代变革中发挥作用。

第一，技术变革。大数据的处理与分析正成为新兴信息技术与应用融合的结点，并持续推动信息产业高速增长。移动互联网、物联网、社交网络、电子商务等是新一代信息技术的应用形态，这些应用会不断地产生即时数据，成为大数据的重要来源。云计算技术则为这些海量、多样化的大数据提供存储和运算平台，并通过分析优化，将结果反馈到应用中，使其创造出巨大的经济和社会价值。大数据价值的实现呼唤新技术、新产品、新服务、新业态的产生。这在硬件与集成设备领域表现为对芯片、存储性能提出更高的要求，并催生一体化数据存储处理服务器、内存计算等市场；在软件与服务领域表现为，引发了数据快速处理分析技术、数据挖掘技术和软件产品的发展。

第二，行业变革。大数据日益成为提高企业核心竞争力的关键因素，不同行业的企业决策正在由“业务驱动”转向“数据驱动”。对大数据的分析可以帮助企业为消费者提供更加快速和个性化的服务；可以为商家制定精准的营销策略提供决策支持；在公共事业领域，大数据在促进经济发展、维护社会稳定等方面起重要作用。各行各业将在大数据技术的指导下，重新定义行业的未来，这将引发全行业的变革。

第三，思维变革。在大数据时代，科学研究方法将发生重大改变。抽样调查不再是社会科学研究中普遍采取的方法，而可以通过实时监测研究对象在互联网上产生的海量行为数据，进行挖掘分析，揭示出规律性的东西，提出研究结论和对策。采集、存储、分析数据能力的提高使大数据时代下我们可以收集全体数据而非随机样本。当我们掌握了海量数据时，精确性就不那么重要了，因为我们足以掌握事情的发展趋势。同时，我们不再关注数据之间的因果关系，而是仅仅从数据中发现相关关系，让数据自己“发声”。该相关关系分析法能够更快、更准确地处理数据之间的关系，而且不易受偏见的影响，提高了分析决策的效率。探求数据价值取决于把握数据的人，关键是人的数据思维，与其说是大数据创造了价值，不如说是大数据思维触发了新的价值增长。

从哲学意义上来说，大数据的价值来自“大成智慧”。每个数据来源都有一定的片面性和局限性，只有整合各类原始数据，才能体现事物的全貌。事物的本质与规律隐藏在各种原始数据之中。不同的数据能对同一个问题提供不同角度的互补信息，可以更深入地理解相关问题。因此，尽量汇集多种来源的数据是大数据分析的关键。

数据科学是数学（统计、代数、拓扑等）、计算机科学、基础科学和各种应用科学融合的结果。大数据能不能出智慧，关键在于对多种数据源的集成和融合。电气与电子工程师协会（Institute of Electrical and Electronics Engineers，IEEE）在 2014 年的计算机技术发展趋势预测报告中重点强调“无缝智慧”。发展大数据的目标就是要获得协同融合的“无缝智慧”，单靠一种数据源会导致片面性。数据的开放共享是决定大数据成败的重要前提。大数据研究和应用要改变过去各部门和各学科相互分割、独立发展的传统思路，更注重强调不同部门、不同学科的协作。

（三）未来展望

大数据的未来应用前景是非常光明的。虽然目前无法准确预测大数据最终会将人类社会带往哪种形态，但只要发展的脚步还在继续，因大数据而产生的变革浪潮将会波及这个星球的每一个角落。

未来大数据应用中一个难以绕开的问题就是用户隐私问题。以脸书为例，2018 年 9 月的一次攻击使得 5000 万用户的账户面临威胁，这是该公司历史上最大的一次泄露事件。随后，2019 年 12 月又发生了一次数据泄露，超过 2.67 亿用户的信息被黑客论坛的在线数据库获取。其中包括用户的 ID、全名和电话号码。目前为止，中国还没有出台专门的法律法规来定义用户隐私，必须利用其他相关法律法规来解释有关问题。但随着民众个人隐私保护意识的日益增强，在进行大数据分析时，必须遵循合法合规地获取、分析及应用数据的原则。数字时代在放大了信息分享带来的好处的同时，也增加了隐私风险。数字经济的特征是把越来越多维度的、碎片化的、实时的小数据转化为“大数据”，在此基础上提供各类线上服务，让消费者和商家都得到好处。但因为数据的广泛使用，在数据周期的每个阶段，从数据收集到存储、分析、使用，再到数据清除，都存在隐私泄露和数据安全的风险。为此，有些科技公司还会创设出专门的“蓝军”单位，叫作“网站可靠性工程师”，其任务是不断寻找和利用漏洞，定期“攻击”数据和隐私管理系统。这些攻击的目标范围包括数据安全、算法性能、云计算和中台软件等。

对于大数据安全分析而言，最关键的不在于大数据本身，而在于对这些数据的分析方法。大数据安全分析可以使用大数据分析通用的技术与方法，但是当具体应用到网络安全领域的时候，还须考虑安全数据自身独有的特点及安全分析的最终目标。只有这样，大数据安全分析的应用才更有价值。例如，在进行异常行为分析，或者恶意代码分析和 APT 攻击分析的时候，分析模型才是最重要的。其次才是考虑如何利用大数据分析技术（如并行计算、实时计算、分布式计算）来实现这个分析模型。

二、数据层面

大数据技术是大数据价值实现的手段和保障，下文将从数据采集、导入与预处理、统计分析和数据解释四个步骤来具体论述大数据分析过程。

（一）数据采集

数据采集是指利用多个数据库接收各种客户端（Web、App 或者传感器等形式）的数据，并且用户可以通过这些数据库进行简单的查询和处理工作。比如，电商会使用传统的关系型数据库 MySQL 和 Oracle 等存储每一笔事务数据，除此之外，Redis 和 MongoDB 等 NoSQL 数据库也常用于数据采集。数据采集是大数据处理流程的基础，目前常用的采集手段有条形码技术、射频识别技术（RFID）等。在大数据的采集过程中，面临的一个主要挑战是并发数高。如亚马逊、淘宝等网站同时访问与操作的用户数以万计，它们并发的访问量在峰值时达到上百万次，需要在采集端部署大量数据库才能支撑。因此，需要深入思考和设计如何在这些数据库之间进行负载均衡。

（二）数据导入与预处理

数据导入与预处理的主要任务是对采集到的数据进行适当的清洗、去噪、抽取和集成。一般而言，通过在采集端部署大量数据库能够采集到海量数据，但是通过各种渠道获取的数据类型非常复杂，给后续的数据分析造成了困难。要想对这些海量数据进行有效的分析，还应将这些来自前端的数据导入一个集中的大型分布式数据库或分布式存储集群，经过数据处理环节后，数据结构变得单一而且易于处理。除此之外，有必要使用聚类分析或者关联分析等方法对数据进行去噪及清洗，以保证数据的质量与可靠性。导入与预处理过程的主要问题是导入的数据量大，每秒钟的导入量常常会达到百兆、千兆，甚至更高级别的数据。

（三）统计分析

统计分析是大数据处理流程中最为关键的部分，也是发现数据价值的主要环节。由于大数据具有多样性特点，仅采用传统的数据挖掘、机器学习、智能计算等数据分析方法已无法满足大数据时代对算法提出的快速高效等要求。因此，需要利用新技术对大数据进行有效的处理分析。其中主要使用分布式数据库，或是分布式计算集群等工具对存储大规模数据进行普通的分类汇总及简单分析，从而满足大部分的基本分析需求。有些即时需求会用到 EMC 的 GreenPlum、Oracle 的 Exadata，以及基于 MySQL 的列式存储 Infobright 等，而 Hadoop 则被用来处理一些基于半结构化或批处理的数据。统计分析的主要挑战是关联的数据量大，其对系统资源，尤其是 I/O 会有极大的占用，亟待提高基础设备的性能。

大数据统计分析具体可以概括为以下四个基本方面。

一是可视化分析，这是用户最基本的要求。因为可视化分析可以直接呈现大数据的特

点，并且可以非常容易地为读者所接受，使得数据分析解读如同看图说话一样简明。

二是数据挖掘算法，这是大数据分析理论的核心部分。基于不同数据类型与格式需要，多种不同的算法才能更科学地展现出数据本身具备的特点，才能深入数据内部，挖掘出数据应有的价值。与此同时，基于大部分数据具有时效性的特征，数据挖掘算法对于迅速处理数据而言至关重要，否则，大数据的价值就会难以衡量。

三是预测性分析，这是大数据分析最核心的应用之一。该种分析从海量数据中勘探出某些特征，在此基础上建立科学的模型，并将新数据导入模型以预测未来可能的结果。

四是语义引擎。大数据时代下的数据类型更加多样化，而非结构化、半结构化数据的出现带来了挑战，需要新的技术加以解决。而“语义引擎”就能够从“文档”中智能提取信息。例如，从用户的搜索关键词、标签关键词或其他输入语义中分析并判断用户的需求，能实现更好的用户体验和精准营销，提高数据分析的效率。

（四）数据解析

解释与演示大数据的分析结果是数据解释的主要任务。不合适的数据显示结果会困扰和误导用户。在大数据时代，基于文本形式及屏幕输出的传统方式已不再适用，因此有必要通过数据可视化、人机交互等新型技术将分析结果生动形象地展示给用户，帮助用户更加清晰地了解整个数据的处理流程和最终结果。

三、实践层面

大数据的价值最终体现在实际运用中。下文将分别从互联网的大数据、政府的大数据、企业的大数据和个人的大数据四个方面描绘大数据时代的美好蓝图。

（一）互联网的大数据

据 IDC 统计，至 2021 年全球所产生的数据量已近 70ZB。互联网是大数据发展的前沿阵地，随着 Web2.0 时代的发展，人们似乎都习惯了将自己的生活通过网络进行数据化，这加速了大数据时代的来临。新型数字经济发展模式 Web3.0 则有望深刻影响下一代互联网形态。

互联网领域大数据应用的典型代表可以简要归纳为如下几点。

（1）用户行为数据，主要通过在手机移动端、智能穿戴设备、智能家居、社交网站等客户端采集此类数据，进行用户的行为习惯与喜好分析，实现内容推荐、精准广告投放、产品优化等目的。微信在其朋友圈逐步投放广告，也是其利用用户行为数据进行精准营销的实践之一。

（2）用户消费数据，主要通过电商平台、导购网站上的交易数据、浏览记录实现对产品的精准营销，以及对用户的信用记录分析，从而实现更精准地开展促销活动，评估用户的信用等级，并协助其理财等功能。阿里巴巴集团凭借旗下的淘宝、天猫等购物平台收集

了大量的用户交易数据和信用数据，能够对用户的消费习惯做出预测，在合适的时点进行大规模的促销，“双 11”购物节的成功就是很好的例子。同时，蚂蚁金融还推出了信用评估体系——芝麻信用分，并在此基础上开发了消费贷款产品——花呗，为其涉足互联网金融领域奠定基础。

（3）用户地理位置数据，主要通过移动端对用户的地理位置进行定位，从而实现线上到线下（O2O）推广、商家推荐、交友推荐等，以线上的营销带动线下的消费。大众点评、美团等团购平台就是利用这种数据类型实现营销。

（4）互联网金融数据，主要是指 P2P、小额贷款、支付等交易记录及信用记录，从而更精准地进行金融产品的营销、对金融产品及服务进行定价、提高风险控制的水平。

（5）用户社交等 UGC 数据，即用户通过互联网平台向其他用户分享自己原创的内容。UGC 不是某一种具体业务，而是用户使用互联网的新方式，即由原来的以下载为主转变为下载和上传两者并重。YouTube 等网站是 UGC 的成功案例，社区网络、图片分享、视频分享等都是 UGC 的主要应用形式。收集这些数据可以用于趋势分析、流行元素分析、受欢迎程度分析、舆论监控分析、社会问题分析等，并从里面挖掘出政治、社会、文化、商业、健康等领域的信息，甚至可以用于预测未来。

（二）政府的大数据

我国政府部门掌握构成社会基础的原始数据，如信用数据、气象数据、环保数据、金融数据、电力数据、教育数据、煤气数据、道路交通数据、自来水数据、医疗数据、安全刑事案件数据、住房数据、海关数据、出入境数据、旅游数据等。这些数据在每个政府部门里看起来是单一的、静态的。但是如果将这些数据关联起来，并对这些数据进行有效的关联分析和统一管理，那么这些数据必将创造出无法估量的价值。大数据拥有变革产业、变革社会的力量，在我国产业结构升级、城市规划、政治改革的进程中必然发挥重要的作用，这使得它成为国家战略的重要组成部分。

具体以智慧城市建设为例。现代化城市计划走向智能和智慧，如智能电网、智慧交通、智慧医疗、智慧环保、智慧城市等，而这些目标的实现都需要紧紧依托大数据，可以说，大数据是智慧的核心能源。据统计，截至 2021 年，我国智慧城市试点数量累计已达 753 个，智慧城市规模为 18.7 万亿元，同比增长 25.5%。“十四五”时期，我国加快推进以人为核心的新型城镇化，至 2027 年中国智慧城市市场规模预计将达 75 万亿元。大数据将为建设智慧城市涉及的多个领域提供决策帮助。对于城市规划，对城市地理、气象等自然信息及社会、经济、文化、人口等人文社会信息的挖掘可以为城市规划提供建议、协助决策，提高城市管理服务的科学性及前瞻性。对于交通管理，通过对道路交通信息的实时挖掘，可以有效缓解交通拥堵的情况，并且可以快速对突发状况作出响应，为城市交通的正常运行提供科学的决策依据。对于舆情监控，通过网络相关关键词的搜索和语义智能分析，可以加强舆情分析的及时性及全面性，全面把握舆情，应对网络突发公共事件，打击违法犯罪等恶劣行为，多角度提高公共服务能力。最后，在安防和防灾方面，利用大数据挖掘能够及

时发现自然或者人为灾害、恐怖袭击事件，可提高应急处理能力和安全防范能力。

（三）企业的大数据

企业决策者需要借助充足的数据来做出科学决策。在未来，大数据就像一个巨大的杠杆，能够从局部撬动企业整体，提升公司的影响力，带来竞争差异、增加利润、愉悦买家、奖赏忠诚用户，将潜在客户转化为客户、增加企业对顾客的吸引力、开拓用户群并创造市场。以下三类传统企业最需要大数据服务：一是对大量的消费者提供产品或服务的企业，大数据能够帮助它们实现精准营销，从而降低成本、提高利润、提升竞争力；二是做“小而美”模式的中长尾企业，借助大数据分析能够对目标市场及客户做出更准确的分析与评价，协助它们实现服务转型与升级；三是在互联网浪潮的冲击下必须转型的传统企业，这类企业必须抓住大数据这一机遇，大胆革新、适时转型，否则必将被互联网企业所淘汰。

在未来，数据有可能逐渐成为企业的一种资产，并逐渐形成数据产业，向传统企业的供应链模式发展，最终形成“数据供应链”模式。在这种情况下会出现以下两个较为明显的现象：一是外部数据的重要性日益超过内部数据，因为在互联网时代下，单一企业的内部数据与整个互联网数据相比犹如沧海一粟，企业外部的海量数据将发挥更重要的作用；二是如果一个企业能够提供数据供应、数据整合与加工、数据应用等多个环节的服务，那么这样的公司会有较为明显的综合竞争优势。在这样的时代发展趋势下，一直做企业服务的行业巨头优势将不复存在，不得不接受新兴互联网企业的挑战，开启新一轮的激烈竞争。以 IBM 为例，IBM 执行总裁罗睿兰认为，数据将成为一切行业中决定胜负的根本因素，最终数据将成为人类至关重要的自然资源。IBM 积极地提出了“大数据平台”架构，该平台的四大核心能力包括 Hadoop 系统、流计算、数据仓库和信息整合与治理，更多地专注于因大数据分析软件而带来的金融业务增长点。

（四）个人的大数据

顾名思义，个人的大数据就是与个人相关联的各种有价值的数据信息的总和。这些数据集被有效采集后，经本人授权后提供给第三方进行处理和使用，并获得第三方提供的数据服务。以个人为中心的大数据具有以下特征。一是数据仅保存在个人中心，只有经过本人授权，第三方机构才能够使用，并且有一定的使用期限，必须接受监管，用后即焚。二是采集个人数据应该明确分类，除了国家立法明确要求接受监控的数据外，其他类型的数据都由用户自己决定是否被采集。三是数据的使用将只能由用户进行授权，数据中心可帮助监控个人数据的整个生命周期。

在此，对个人大数据时代的光明前景进行大胆展望。未来，每个用户都可以在互联网上注册个人的数据中心，以存储个人的大数据信息。其中，有一部分个人数据是无须个人授权即可提供给国家相关部门进行实时监控的。比如，罪案预防监控中心可以实时监控本地区每个人的情绪和心理状态，以预防自杀和犯罪的发生。除此之外，用户可决定其他个

人数据哪些可以被采集，并通过可穿戴设备或植入芯片等感知技术采集捕获个人的大数据，如牙齿监控数据、心率数据、体温数据、运动数据、视力数据、记忆能力、饮食数据、购物数据、地理位置信息、社会关系数据等。用户可以将其中的牙齿监测数据授权给牙科诊所使用，由他们监控和使用这些数据，进而为用户制订有效的牙齿防治和维护计划；也可以将个人的运动数据授权提供给某运动健身机构，由他们监测自己的身体运动机能，并有针对性地制订和调整个人的运动计划；还可以将个人的消费数据授权给金融理财机构，由它们帮你制订合理的理财计划并对收益进行预测。但是，个人数据中心的产生必然伴随着个人数据隐私被泄露的隐患，所以，未来在推进个人数据中心建设的进程中，需要解决的问题是如何通过有效的数据监管措施来保障数据的安全与合理利用。

第四节　大数据的趋势变革

一、大数据变革的主要领域

大数据带来的趋势转变将逐渐改变我们生活的方方面面，而其中有几个领域将率先面临颠覆。

（一）数据安全

随着大数据的广泛应用，数据安全日益引起关注。近几年数据安全事件频发，随着全球各国逐渐采用更先进的数据安全技术并制定更完善的数据保护法律，数据安全监管趋严将是未来的一大趋势。

在数据安全方面，与商业机密的保护问题相比，人们更注重的问题是如何守护个人隐私。数据安全意识提升的背后是人们对数据公开化及其风险的担忧。如果数据风险无法被有效管控，人们就无法在真正意义上信任数据，而这将在很大程度上阻碍日益发展的大数据产业。个人的隐私、公司机密乃至国家和国家之间的数据保护，将会是未来亟待完善的部分。而当数据成为商业重要且关键的资产时，像“首席数据隐私官”这样的职业也就应运而生了，数据安全与数据隐私保护成为数据应用中不可或缺的一部分。

（二）分析的简化与外包

数据分析工作的简化与外包预示着未来大数据将会向产业链分工的方向发展。随着数据信息的迅速膨胀和大数据应用的逐步落地，越来越少有企业可以独立完成从原始资料采集、加工、分析乃至落地应用的完整程序。未来数据的不同处理阶段都有机会发展出专门的技术公司协助企业完成大数据应用前的整备工作。大数据产业链上的每一个环节都有可能产生新的问题与创新，大数据产业革新的每一阶段都有可能激荡出新的问题与机会。新的问题不断地聚焦起来，对应的数据源也趋于集中，这时，一个新的产业链机会也就随之产生。中间层的服务与创新在大数据产业的发展过程中将扮演至关重要的角色。

（三）政府的数据态度

从大数据产业链的整体来看，政府拥有最多的数据。因为政府锁定了大部分公共服务领域的关键数据源，它是公共数据开放的大资源，也是一把驱动大数据的“金钥匙”。政府的数据涵盖金融、医疗、能源、食品、交通、治安、环境等多个方面，且这些数据是相对集中且十分关键的。政府数据的开放将是产业创新的催化剂，对于整体数据产业的发展也至关重要。2022 年 6 月，《国务院关于加强数字政府建设的指导意见》明确提出有序推动公共数据资源开发利用，提升各行业各领域运用公共数据资源推动经济社会发展的能力。而各个行业也可以顺应政府数据政策的脚步，开始尝试进入大数据驱动乃至大数据变革的第三阶段。

（四）多屏时代

过去几年，手机极大地颠覆了我们的生活，但随着大数据的发展，可以预见，未来有两个新的屏即将改变我们的生活：一个是 Smart TV，即家里的电视屏，搜集你看过节目的相关数据并且会向你推荐你爱看的节目，形成了自然数据闭环；另一个是物联网汽车，将来所有汽车的内部都会像特斯拉（Tesla）一样，中央显示屏控制每个部分、记录汽车行驶中的所有数据，信息的流动由此产生。可以根据时间分配和载具分配两个层次来思考这个问题：通常情况下，人在不同的时间会因为当时具体的环境状态对不同设备的依赖程度不同。在家时对 Smart TV 的依赖一般会比手机高，而离开家前往下一个目的地时，则更需要手机，如果是自己开车，车载导航或车载屏幕将会成为主要的关注对象。因此，未来互动的过程中我们应当更加关注如何采集到有价值的数据，并进一步对使用者的日常生活做出优化的回馈。

（五）数据行业化

互联网影响比较大的行业必然更容易数据化，因而大部分大数据应用的落地点都与特定行业相关，已经崭露头角的有金融、医疗、电商等行业。下一步的大数据应用应该会在不同的领域各自发展，并不存在适用于每一个领域的通用的解决方案，而零售、医疗、教育、金融等行业都将因“互联网 +”的带动而发展。很多小公司在起步时产生了很多小数据，这是从 0 到 1 的过程，然后整合碎片化的数据，最后积累大量数据。这三个进程的时间点加上不同的应用，铸就了行业大数据。未来，大数据将从过去的浅层联结转变为深层联结。从大数据由浅而深的演变中可以观察到：从前习惯以行业为出发点，思考网络（数据）可以帮我们做什么，到了互联网和大数据时代，则转变为以网络（数据）为出发点切入思考，再把原行业的思维放进来碰撞，看看可以激荡出什么样的创新思维。Uber、Airbnb 都颠覆了以往行业运用网络的概念，但只有这种思考与创新的方式才能将跨行业的东西提升出来。

二、大数据的产业变革

（一）数据的挖掘和存储：对应云技术的运用和升级

云计算是大数据存储和分析的重要基础设施，正是云计算的发展迅速推动了大数据产业的发展，原因有如下三点：一是云计算按需付费和资源共享的特性降低了企业产生和使用大数据的门槛；二是低门槛的特性又推动了大量中小企业使用云计算，从而提高企业信息化程度，加速了数据的产生；三是云计算的低使用成本和高计算能力提升了企业的大数据处理能力。

云计算促进大数据产业发展的方式如图 1-6 所示。

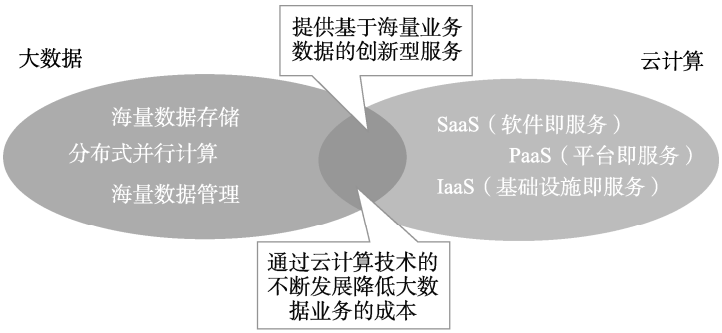

图 1-6　云计算大大促进大数据产业的发展

（二）数据的整理和分析：对应算法和 AI 的运用和升级

大数据算法及人工智能的迭代升级提升了数据分析能力。深度学习算法突破性地以更接近人脑的方式利用大量数据训练机器，通过训练使其自主掌握规律，并且结果将会随着数据量的不断增长而更加准确可靠。

（三）大数据技术的运用场景：对应政府、企业的开放，体量越大的传统行业，蕴含越大的大数据空间

《中国大数据发展调查报告（2018 年）》统计显示，过去制约企业大数据发展的首要因素是政策限制（如隐私保护），其次是数据资源和数据人才的短缺，随后是技术能力不足，最后是诸如应用模式不清晰等其他原因。政策方面，由于政策、法规的滞后性，数据跨部门、跨企业、跨行业甚至跨领域流动的需求被牢牢禁锢，这也使得政策限制成为摆在我国企业大数据应用面前的最大障碍。数据资源增长迅速，但是如何通过技术手段获取高质量的数据是企业面临的重要问题。

但随着政府对大数据战略的日益重视，政府和企业的数据开放流动正在开始。

政府拥有最丰富和优质的大数据资源，各个政府部门掌握着社会第一手的原始数据，

如房地产、医疗、教育、金融、交通、旅游、气象、电力、海关、司法、人口等各类经济和社会运营的基础数据。过去政府数据更新频率低且未向社会开放，因此社会并未充分利用这一块数据的价值，目前政府已经意识到数据的重要性并出台政策逐步开放大数据供社会利用。若政府大数据获得有效利用，将产生巨大的价值。

相应的，企业的经营中也将产生大量数据，包括用户信息、用户行为、产品运行数据等。根据《中国大数据发展调查报告（2018 年）》显示，企业对大数据资源最大的需求来自企业信息公开，继而是政府所拥有的大数据信息，对市政管理、教育科研及交通服务类公共数据资源的需求也有所增强。过去，互联网数据已经进行了大量的应用，未来，随着企业销售、供应、经营的互联网化，企业与客户、供应商、中间商的互动互联网化将产生大量的数据。

（四）大数据变革下的三大关键行业机会

从大数据对传统行业的颠覆看，主要的机会在于中间层，尤其是在金融、医疗、零售这三个行业。

1. 金融与保险

用一个词概括金融在大数据时代的机会点，那就是“微（micro）”。过去很多的创新都受到技术和数据能力的限制。未来数据的采集、加工和应用都将实现个人化的价值，将会激发很多金融商业模式。数据动态的意义首先体现在金融保险体系里的客户监测。过往的个人信用评估结果都是滞后的，往往无法有效地反映最新的个人信用风险，导致银行或是保险公司不能提供最符合顾客需求和利益的服务；其次体现在将解决服务合理性的问题，也就是当客户使用了服务时，才向客户收取费用。

以汽车保险为例。过往我们对于汽车保险的保费设定是根据客户的驾驶肇事记录来调整保费费率，也就是说汽车保险的价格标准建立在投保车主的驾驶行为基础之上。但肇事记录通常具有滞后性，在大数据时代可以利用更便捷的数据采集系统收集更动态的数据来预测危险驾驶的风险。或许，未来的汽车都会像某些新能源汽车一样，通过车内安装的传感器，记录驾驶员如何踩油门（例如，习惯性的紧急刹车就是一种危险驾驶的信号）、换道时是否打方向灯、是否频繁地按喇叭等行为。这些资讯都可以推断出驾驶员是否拥有安全的驾驶习惯。如果再把驾驶员的行车路线数据与政府公布的危险肇事路段的数据做对比，就可以知道这辆车每天上下班的路线属于怎样的安全等级。综合以上两类数据，即便没有肇事记录，保险公司也能根据这些数据来动态调整风险评比，并随时调整保费的费率。同理，如果车险是为了确定用车人在驾驶期间的承受风险，那么通过车辆的传感器可以清楚了解这辆车有多少时间停在车库，又有多少时间处于被使用的状态，保费的计费也可以根据车辆实际承受风险的时间来对客户收费，这也就实现了前面所提到的动态计价。

2. 医疗

医疗领域当前所面临的最大问题是数据未能整合，不同医院之间未能实现数据信息的互通。比如，在 A 医院拿不到之前在 B 医院开的病历。第二大问题是中国人口老龄化严重，导致医疗费用负担沉重，因病返贫十分常见。只有降低医疗成本，减少滥用资源和药物，

才能根本性地减少政府负担，将资源分配给真正需要的人。美国福特公司的 30 万名员工每年享有 30 亿美元的医疗保险预算，但这笔钱过去只有一个人在管。1997 年福特在这方面第一次引入数据应用，分析之后发现有人竟然 150 岁还在领医疗保险，有人一年领两次怀孕补助等不合常理的情况。这些都是无谓的资源浪费，但如果不通过数据，可能永远也不会被发现。医疗是一个连续性的行为，一个人从健康、亚健康乃至于疾病的阶段都不是突发的，背后都有遗传或是生活饮食习惯的脉络可循。很多疾病的危险信号常常是因为信息未能互通而被忽略，像美国就曾经通过传染病传播数据预估要生产多少疫苗，以及各区疫苗使用状况，大大地提升疾病防治效果。

3. 零售

对于零售业，最重要的就是如何运用数据把供应与零散的需求做匹配。买家希望得到的是：我要什么？最快帮我找到我要的东西。给我最好的价格，用对我最方便的方式付款，在刚好的时间送达。供应方则希望知道：怎样才能满足消费者？怎样用最小的库存，最快的方法，最合理的利润率来服务顾客？供应链的处理怎样可以变得更好？怎样减少成本浪费？以数据驱动为基础的线上零售发展已经十几年了，但线下将会出现什么情况？当 POI 逐渐成熟，拿着手机，处处都能发挥大数据的连接能力，时刻都是机会点。人们懒得再特意下载一个应用程序购物，线上再发达，某些时候它也不是最方便的渠道。要想改善零售用户的体验，只有开展全渠道的服务。对零售业来说，最好就是线上线下都能覆盖到。只有线上的应用程序是不够的，最好连线下的渠道都能拿到，不然很容易就被别人“弯道超车”。线上线下一体化的新零售方式逐渐成为一种刚需。

“推荐”将是另一个爆点服务。现代人能在同一时间享受多种服务，浏览多种产品信息。但数十万款与你有关的商品摊在你的眼皮底下，你怎么选择？人主观上都希望自己可以选择，但面临太多选择时，选择本身反而变成了一种负担。所以未来的“推荐”应该是在“优选”与“逛”之间。大数据让手机变成个人消费助理，不断跟着你走，也不断领着你走，在商家和消费者之间成为一个媒介。最后，零售业的根本是“生产”问题。数据将成为产品创新和改良的依据，从设计到生产、包装、销售、售后的过程中观察并不断优化，最终能够帮助商家生产出符合顾客需求的商品。大数据时代对零售业来说不是一个单纯的转型问题，而是一个攸关存亡的生死问题。

三、大数据变革趋势

（线上阅读）

本章小结

大数据概念的提出离不开特定的时代背景，技术进步、产业升级及社会历史的积淀都促使了“大数据”概念的诞生。本章首先从大数据概念的历史演变过程出发，讲述了大数据的主要发展阶段，以及国外主要发达国家和国内大数据的发展现状。其次从理论角度讲述大数据体量大、速度快、类型多、价值强、在线化等特征，并指出了大数据将带来变革时代的力量及未来可能面临的潜在挑战；从技术层面剖析大数据实现价值所经历的四个步骤，即数据采集、导入与预处理、统计分析和数据解释；从实践层面探讨互联网、政府、企业和个人的大数据价值体现，以便更深入地理解大数据的概念。最后阐述了大数据带来的产业变化趋势，行业颠覆及三大关键行业的变革，也揭示了数据市场的未来发展趋势和变革。

理解大数据的概念是实现数据价值的前提条件。只有深入理解大数据的概念，才能合理运用大数据技术，使其成为创新变革的动力和推动时代进步的力量。

即测即练

自学自测 扫描此码

简答题

1. 大数据技术是为了应对产业界的哪些挑战而产生的？
2. “大数据”除了指数据量大还包含哪些含义？
3. 在“3D 数据管理”及其后的优化观点中，能总结出大数据的哪些特征？
4. 大数据统计分析具体可以概括为哪些方面。
5. 过去制约企业大数据发展的最主要因素是什么？为什么？