中国计算机学会文集

中国大数据技术与产业发展报告（2014）

中国计算机学会大数据专家委员会
中关村大数据产业联盟 主编

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

中国大数据技术与产业发展报告（2014）/ 中国计算机学会大数据专家委员会，中关村大数据产业联盟主编 . —北京：机械工业出版社，2015.4
（中国计算机学会文集）

ISBN 978-7-111-49963-3

I. 中… II. ①中… ②中… III. 信息技术 – 产业发展 – 研究报告 – 中国 – 2014 IV. F49

中国版本图书馆 CIP 数据核字（2015）第 073943 号

本书是由中国计算机学会大数据专家委员会和中关村大数据产业联盟组织编写的，目的在于梳理大数据应用现状及发展趋势，为政府制定推动大数据产业发展的政策提供建议；同时，探讨大数据研究面临的科学问题和技术挑战，为研究机构和研究人员提供参考指南。书中详细阐述了大数据背景与动态、大数据典型应用、大数据技术进展、大数据 IT 产业链和生态环境，以及大数据发展趋势与建议。本书的编写集中了各个领域众多专家的知识和智慧，在一定程度上反映了我国大数据学术界和产业界的共识。

本书可作为广大计算机科学技术人员了解当前大数据发展动态的渠道，适合本领域决策人员和科研人员参考，并可作为高等院校计算机相关专业教师和学生的参考书。

中国大数据技术与产业发展报告（2014）

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：曲　熠　李　艺　陈佳媛　秦　健　　责任校对：董纪丽
印　　刷：中国电影出版社印刷厂　　版　　次：2015 年 4 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：17.25
书　　号：ISBN 978-7-111-49963-3　　定　　价：75.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

编　委　会

前　言

近年来，大数据浪潮以排山倒海之势席卷全球，既提供巨大的机遇，也带来一系列的挑战。为了推动大数据科学技术和产业的良性发展，中国计算机学会于 2012 年 10 月成立了“大数据专家委员会”，其宗旨是探讨大数据的科学与技术问题，推动大数据学科方向的建设与发展；构建面向大数据产学研用的学术交流、技术合作与数据共享平台，并为相关政府部门提供战略性的意见与建议。在中国计算机学会大数据专家委员会和中关村大数据产业联盟的精心组织下，众多专家历时大半年时间撰写完成本书。

中国计算机学会大数据专家委员会的 100 多位专家来自高校、科研院所、企业和政府部门，从事的专业涵盖计算机系统、通信、数据库和数据挖掘、大数据应用等各个不同的领域，本书的编写集中了来自几十家单位各领域专家的知识和智慧，在一定程度上反映了我国大数据学术界和产业界的共识。组织撰写本书的目的在于，梳理大数据应用现状及发展趋势，为政府制定推动大数据产业发展的政策提供建议；同时，探讨大数据研究面临的科学问题和技术挑战，为科研机构和科研人员提供参考指南。

本书内容分为五章：第 1 章介绍大数据背景与动态；第 2 章阐述互联网、金融、电信等 10 个重要行业的大数据应用现状和发展趋势；第 3 章阐述大数据技术的进展；第 4 章讨论大数据 IT 产业链与生态环境；第 5 章分析大数据发展趋势并提出相关建议。附录还提供了世界开源组织及国内大数据产业园的介绍。

本书的重点章节是第 2 章和第 3 章。第 2 章介绍大数据典型应用。我国的大数据应用刚刚开始，有些应用的数据规模可能还不够大，采用的方法也许不够新，但新兴产业是“用”出来的，只有广泛应用才能发现技术差距和需要突破的技术壁垒。发现典型大数据应用案例和宣传推广应用大数据技术的经验是本书的主要动机，今后我们会更加关注应用案例。

第 3 章分析大数据技术体系的现状。专家委员中多数是科研工作者，熟悉本领域中科学技术研究的进展，擅长探讨技术发展趋势，以及分析科学研究和技术开发中面临的问题与挑战，本书的主要价值就体现在对大数据技术的分析方面。第 3 章的每一节都由活跃在科研第一线的专家撰写，每人只写自己最熟悉的领域，许多见解都来源于实际工

作的体会。为了反映专家们的群体倾向，专家委每年做一次大数据技术发展趋势的年度预测，通过投票方式将最受关注的科学、技术、产业、应用、政策等相关变化趋势挑选出来。这部分内容反映在 5.2 节中，希望对读者有所启迪。

在其他章节中，企业界和政府部门的专家也表达了一些观点，如第 4 章提出的大数据产业链全景图、国内外大数据产业发展呈现的六大趋势、大数据产业发展的主要瓶颈等，都有独到的观点。第 4 章把大数据人才资源问题独立出来专门分析，是因为这是一个十分重要而紧迫的大问题，需要各方面高度重视。

电子计算机奠基人冯 • 诺依曼指出：“在每一门科学中，当通过研究那些与终极目标相比颇为朴实的问题，发展出一些可以不断加以推广的方法时，这门学科就得到了巨大的进展。”发展大数据科学与技术需要从“颇为朴实”的问题做起，要特别重视从应用中发现科学问题。检验一切技术的唯一标准是应用，技术有限，应用无限。我们一定要坚持“应用为先”的发展战略，坚持应用牵引的技术路线，不能完全走技术驱动的道路。大数据的炒作高峰期已过，现在是关注大数据技术“落地”的时候，希望本书能对大数据技术的落地有所贡献。

数据科学的形成还需要一个培育过程，现在还不能过分强调与统计等传统技术完全不同的革命性技术。统计学家们花了 200 多年的时间，已发现数据处理过程中的许多陷阱，大数据出现后这些陷阱并不会自动填平，大数据中仍然有大量的小数据问题。目前解决大数据问题要逻辑演绎和归纳相结合、白盒与黑盒技术相结合、大数据方法与小数据方法相结合。本书收集的大数据应用案例和阐述的大数据技术，没有与小数据划出明确的边界，随着大数据技术的成熟，这种边界会慢慢清晰起来。

限于时间和篇幅，本书只选择了部分发展较好的典型应用领域进行介绍，还有很多领域的大数据应用情况没有纳入本书。在后续工作中，大数据专家委将继续不断完善和丰富本书的内容，对于特色行业或应用领域会进行更为详细的调研，出版有针对性的面向行业应用的书籍。本书是专家委第二次组织撰写，虽反复修改了十余次，但书中肯定还存在一些错误，撰写组织工作也有很多不当之处，希望产业界和学术界的专家学者与广大读者提出批评和建议，共同推动中国大数据技术与产业的发展。

李国杰

2014 年 12 月 1 日

撰 写 说 明

本书由中国计算机学会大数据专家委员会主任李国杰院士担任主编，中国计算机学会大数据专家委员会秘书长程学旗和中关村大数据产业联盟秘书长赵国栋担任副主编，各章节的统稿人作为编委会成员。在李国杰院士的带领下，编委会多次召集各领域专家研讨内容框架，并联合了二十多家单位参与本书的撰写工作。其中，第 1 章由赵国栋、程学旗、杨东日、胡然、刘姝祎等撰写；第 2 章由陈新河、施水才、王维负责整理，互联网大数据由沈烁、田野、袁博等撰写，金融大数据由闻学臣、陈继东、林述民等撰写，电信大数据由何鸿凌、孙少陵、徐萌等撰写，电力大数据由邓春宇、张宇航等撰写，交通大数据由杨东日、刘姝祎、刘超、鲍侠等撰写，健康医疗大数据由张彦春、熊锦华、马建刚、徐红燕等撰写，政府大数据由施水才、贺兆辉、晋家骧等撰写，农业大数据由姜春铃、谢润梅等撰写，地理信息大数据由张林、张平、李先怡、陈艳武等撰写，新媒体大数据由王永滨、赵子忠、冯爽、郑维东、祝建华等撰写；第 3 章由吴甘沙、何利文、杜小勇、袁晓如、尹绪森、钟翔、连城、周虎成、石勇、陈继东、王健宗、陆嘉恒、董兆安、祝建华、罗圣美、张丹、徐红波、沈烁、田野、李航、白小勇、刘睿民等撰写；第 4 章由杨东日、潘柱廷、金波、刘姝祎、胡然、周涛、黄道丽、何治乐等撰写；第 5 章由李建中、靳小龙、石勇、潘柱廷、周涛、陈懿冰等撰写；附录由王维、查礼、刘伟等撰写。最后，程学旗、靳小龙、王元卓、杨婧负责材料组织和统稿等工作。黄宜华、王治平、彭智勇、邓波、阮彤、周霞、周志华等大数据专家委员会委员积极参与了本书的撰写，不仅提供了素材，还参与了本书的修改工作。由于本书经过多次修改，因此对参与专家的统计可能还有疏漏，在此深表歉意。最后对所有参与本书编写的专家表示感谢。

目　　录

第 1 章　大数据背景与动态

1.1　大数据的宏观价值与背景

“大数据”的内涵远远超越物联网、云计算等信息技术的概念。物联网本质上是器物层面的技术，从大数据的视角而言，是采集数据的终端。云计算本质上是 IT 服务交付手段的变革，并由此引发一系列技术基础架构的更新。物联网和云计算都是信息技术发展到一定阶段的自然延伸，依然属于信息技术范畴。而大数据则是数据积累到一定规模后引发的质变。大数据超越信息技术，使人们重新界定国家竞争的主战场，重新审视政府治理水平，重新认识科学研究的新范式，重新审视产业变迁的驱动因素，重新理解投资的决策依据，重新思考公司的战略和组织结构。

1.1.1　国家——保障数据安全，促进数据开放

2012 年 3 月份，奥巴马发布了美国版的“大数据研究和发展计划”。通过该计划可以看出，国家层面大数据技术领域的竞争事关一国的安全和未来。国家数字主权体现为对数据的占有和控制，是继边防、海防、空防之后的大国博弈的空间。大数据必须上升为国家意志，落实为国家战略。欧盟、日本、新加坡等组织和国家已经开始纷纷行动。

2013 年，美国人斯诺登给世人揭开了“数据战争”的冰山一角，美国的“棱镜计划”事实上把所有国家、个人都纳在 NSA（美国国家安全局）的监控之下。参与棱镜计划的公司包括谷歌、雅虎、Facebook、微软、苹果、思科、Oracle、IBM 等科技巨头。由此可以看到，在大数据时代，IT 产业的强大已经成为直接决定一个大国是否可能成为强国的最为关键的因素。没有数据安全，就不会有国家安全；没有强大的 IT 产业，就不会成为一流国家。

保护国家层面的数据安全是以数据开放为基础的。开放是一种态度，更是一种能力。一些重大基础数据开放，可以构成社会的数据基础，按照大数据定律之一“数据之和的价值远远大于数据的价值之和”来推断，来自不同领域的数据聚合在一起，开放给社会，将会产生类似核聚变一样的价值发现效应。

现在，电子商务、社交网络、基础通信所产生的数据以及国家相关部委的数据，具备聚合的效应和产生核聚变价值的基础。国家统计局联合百度、阿里巴巴已经做了一些探索性的尝试，这是非常好的开端。与此同时，“数据割据、拥数自重”的现象也是普遍存在的。譬如气象观测数据，这类数据对于研究大气变化、气候演变、农业指导等具备非常重要的科学意义。但目前来看，诸如此类的数据应用范围还有很大提升空间。再如住建部的购房数据，这类数据对于防止腐败、研究经济走势、人口迁移，甚至制定国家决策都具有至关重要的作用。这类数据如果开放给社会各界，一定程度上会繁荣多学科、跨领域交叉研究，由此可能会推动中国在各个方面的进步。

开放的数据是基础，促使信息产业繁荣，才能诞生真正的数据驱动的企业，企业反过来在数据领域的技术进步，才是确保国家数据安全的长治久安之策。很难想象，如果没有谷歌、微软、Facebook 这样的公司，单凭美国政府一己之力，难以实施如此庞大的“棱镜”计划。所以制定国家大数据战略，需要重新思考传统的所谓的“国家机密”和国家安全的关系，应当把消除部门数据割据，建立公开、透明、共享的数据公共平台作为长期的战略目标。

1.1.2 政府——转变理念，集成信息，抓住机遇

面对海量、动态、多样的大数据，传统的思维方式和行为方式将面临巨大挑战，尤其是在公共服务领域。有效集成信息资源的能力将会为政府治理理念和治理模式的转变提供强大的技术支撑。

当前，越来越多的国家开始从战略层面认识大数据，在政府治理领域融入大数据思维和技术，推进大数据的发展。英国 2006 年启动“数据权”运动，韩国 2011 年提出打造“首尔开放数据广场”，美国 2012 年启动“大数据研究和发展计划”，联合国 2012 年推出“数据脉动”计划，日本 2013 年正式公布以大数据为核心的新 IT 国家战略。在此背景下，我国政府也顺应时代发展趋势，契合推进国家治理能力现代化的时代要求，推动大数据发展，政府、企业和科研院所正在进行多方位布局，充分利用大数据提升国家治理能力。对于政府治理而言，大数据时代在带来机遇的同时也充满挑战。

大数据为政府治理能力的提升带来了发展机遇。首先是为推动政府治理理念和模式的变化带来机遇。在政府治理领域，通过让海量、动态、多样的数据有效集成为有价值的信息资源，推动政府转变治理理念和治理模式，进而加快治理体系和治理能力现代化。其次是为推动政府治理决策精细化和科学化带来机遇。在大数据时代，互联网数据

的价值随着海量积累而产生质变，能够对经济社会运行规律进行直观呈现，从而降低政府治理偏差概率，提高政府治理的精细化和科学化。再次是为推动政府治理提高效率和节约成本带来机遇。利用大数据可以使政府治理所依据的数据资料更加全面，不同部门和机构之间的协调更加顺畅，进而有效提高工作效率，节约治理成本。

大数据对提升政府治理能力的重要性不言而喻，但在实际工作中具体运用大数据却任重而道远。现阶段，大数据在政府治理领域还未得到足够重视。我国政府部门目前几乎没有使用大数据技术，很多政府部门并未对大数据提升业务能力予以足够重视，大数据资源管理的思维尚未建立。大数据在政府治理中的技术运用尚在探索。随着我国信息化技术应用不断扩展，国家及企业层面产生了巨量大数据，但总体集成、掌握、整合、分析这些数据需要成熟的技术投入，如何利用大数据进行精细分析仍处于摸索阶段。此外，大数据本身的管理还需要综合完善。如何管理大数据，我国各部门还缺乏统一标准。异质数据来源、架构、管理体系目前还不能有效整合，在一定程度上降低了数据的使用效率[㊀]。

1.1.3　学术——科学地研究数据，用数据来研究科学

学术界在大数据时代有了更为广阔的舞台。某种程度而言，近几年计算机领域的发展是谷歌、亚马逊等一线的互联网公司所推动的。虽然学术界在算法方面具备无可替代的优势，但在算法工程应用领域，却因为缺乏实践场景而裹足不前。

在大数据时代，许多学科表面上研究的方向大不相同，但从数据的视角看，其实是相通的。例如自然语言处理和生物大分子模型中都用到“隐式马氏过程”和“动态规划方法”。其最根本原因是它们处理的都是一维的随机信号。再如用于图像处理的算法和用于压缩感知的算法也有着许多共同之处。

数据视角为许多学科的发展带来了新的研究思路。以自然语言的机器翻译研究为例：最初科学家们试图为计算机建立一系列的语法规则，按照语法、词义来翻译成另外一门语言。该思路非常直观，因为人们就是如此理解学习语言的，但在实践中困难重重。基于语法规则的翻译器几乎没有商用过。而当科学家们改弦易张，计算每一个词、每一句话的“合理概率”时，复杂的机器翻译就简化成了文字的概率计算。通俗来讲就是：“如果大多数人都这么说，我们就认为是对的！”

从宏观尺度研究的天体信息学、社会行为学，微观尺度上分析人类的基因组，到追踪物理学家们梦寐以求的“上帝粒子”，这种思想在越来越多的领域得到应用。

㊀　引自：《大数据时代政府治理能力的提升》 吴建树

随着社会的数字化程度逐步加深，越来越多的学科在数据层面趋于一致。可以采用相似的思想来进行统一的研究，而这恰恰是数学家的特长。因此数据科学在数学和实际应用之间建立起了一个直接的桥梁。而这些实际应用正是来自于像信息服务等现代产业中最为活跃的一部分。对数学来说，这是一个千载难逢的机会。

通过建立大数据共享实验平台和国家级的大数据研究实验室，搭建产业界和学术界的桥梁，为学术界优秀的算法提供演练的舞台，为产业界困扰的难题提供破解的机会，从而间接推动数据科学领域学科建设与人才培养的工作。

1.1.4 产业——产业需要变革，行业需要互融互通

产业需要变革，行业需要互通互融。所谓“大数据 +”，就是将大数据思维嫁接到不同的产业中，推动大数据在各行各业落地。

大数据不仅仅只关系到 IT 行业，众多行业龙头公司都已经意识到了大数据新思维的巨大冲击。给企业家们带来冲击的并不是大数据本身，而是一些新兴公司不可思议的跨界能力。行业之间的界限变得越来越模糊，这些新兴公司采用新的技术、新的模式，大规模采集数据，迅速形成预判，并迅速扩张到相关企业行业。譬如乐视网已经涉及电视销售、电影拍摄；小米公司除手机销售外，也开始涉及电视销售；百度、360 等企业也都开始做各种硬件，如百度影棒、360 随身 Wi-Fi 等。

互联网金融行业发展快速，该行业对传统金融行业造成的冲击之大，仿佛一夜之间就成了传统金融业的公敌，阿里巴巴旗下的余额宝产品仅仅用了 5 个多月的时间，累积申购金额就超过了 1000 亿元人民币。事实上，目前互联网金融还是在发展的初级阶段，仅仅是把线上渠道对接了线下金融资源，但就是这种“对接”行为，已经引发了行业性的“地震”。下一步将是线上渠道向智慧方向演进，这个阶段大数据才真正派上用场。借此回顾阿里集团的战略排序：平台、金融、数据。数据是在金融之后的第三个发力点。

类似案例将在各行业轮番上演。信息化程度越高的行业，受大数据冲击的可能性越高，被颠覆的可能性越大。

因此，从大数据的视角来看，任何产业中，数据资产都将成为最核心的竞争力。

传统产业，或者说各行各业都将面临在大数据和移动互联网时代如何彻底转型和再造的问题。产业整合，将在大数据时代出现全新的整合逻辑和实现契机。各行各业都可能在大数据和移动互联网时代重现生机、焕发青春。当然，与此对应的是，如果企业和行业不能跟上这个时代步伐，它们可能就将永久地留在过去，退出未来的舞台。

1.1.5　公司——平台化竞争，特色应用化生存

未来，各行业只有两类公司得以生存，一类是平台，一类就是有特色的应用。这种“星空格局”将呈现众星拱月的景象。平台的竞争可能更加残酷，一个行业甚至只可能存在唯一一个压倒性的平台；应用的竞争同样残酷，产业成熟周期将缩短到 1 年，决胜期短至 2 个月。

在“星空格局”之下，公司的竞争力更多地体现在“平台 + 特种部队”的模式。以“星空格局”作为产业演化的最终形态，以特种部队作为业务竞争的基本单元，那么公司的战略、组织、文化等方面需要彻底的重组。

例如国内的某些公司，组织层级被高度压缩为两级：员工、合伙人。每个合伙人管理一方面的事务，譬如营销、采购、制造等。但是合伙人直接管理许多员工（团队）的模式完全颠覆了一般管理学上定义的，管理跨度不要超过 7 人的界限。而另外有些公司，直接向公司最高管理层汇报的团队有多达数十个。在组织高度扁平化的公司里，企业文化必然有其独到的地方，关键词无外乎包括：“专注、极致、口碑、快速、用户体验、全面体察等”。

传统公司的确需要重新审视自己的战略，重构组织，再育文化。否则，胜利的天平总是向这些类似泛互联网范式的公司倾斜。

1.1.6　投资——大数据将提供价值分析新视角

在对上市公司进行投资价值评估时，由于各家公司的成立及上市时间不同，运营结构、体制机制也各具特色，如果单从某些指标着手，评价结果难免有失偏颇，导致投资价值评估不具备典型性，无法实现投资价值评估的目的，也无法向投资者正确反馈上市公司投资价值高低的信息。

其中，“高科技”行业尤其难以评估。以软件公司为例，虽然不同的软件公司所服务的行业不同、采用的技术不同、产业成熟周期也不同，但是产业的成熟周期都有越来越短的趋势，这使得一旦发现机会，他们就会迅速成长为产业的新标杆。谷歌公司仅仅用了 15 年的时间就跨入了千亿美元市值公司的行列；小米公司成立三年，估值已达到 100 亿美元。如此高速地成长，在传统行业几乎是不可能发生的事情。

长期研究分析发现，以 TMT（电信、媒体、科技）为代表的高技术含量企业，虽然发展迅速，机会众多，但也随时可能出现潜在的风险，基于此，许多基金经理因看不懂、看不准而从不涉足该类企业。这的确是一个两难的问题。基金经理不了解，资金就

会投入的少，客观上对行业发展不利。用所谓的 WACC 方法评估 TMT 公司价值，在当前来讲已经不那么现实了。

虽然大数据的起源要归功于互联网与电子商务，但大数据最大的应用前景却在传统产业。一是因为几乎所有传统产业都在互联网化，二是因为传统产业仍然占据了国家 GDP 的绝大部分份额。随着数据逐渐成为企业的一种资产，对大量消费者提供产品或服务的企业、做“小而美”模式的中长尾企业、互联网压力之下急需转型的传统企业，这三类企业的投资价值与市场发展前景愈发显得难以看透。

针对这种情况，大数据将提供分析公司价值的新视角，帮助投资人发现公司的潜在价值。所谓公司的价值，与其拥有的数据资产的规模和活性成正比，与其解释、运用数据的能力成正比。

这里强调了数据资产的两个属性：规模和活性。数据资产评估模型从五个维度来评估数据资产的商业价值，规模和活性仅仅是其中的两个[㊀]。

利用数据资产评估商业价值的这种思想获得了越来越多的投资人认可。大数据已经成为基金经理切入企业发展前景与价值评估的绝好视角和新型工具。

产业界和资本市场沟通的“纽带”在大数据时代将显得越来越重要。需要让越来越多的投资人理解大数据，洞察行业发展趋势，帮助产业界更好地开拓产融结合的路子。

1.2 国内外大数据发展动态

1.2.1 国外大数据发展动态

1. 国际战略动态

纵观世界各国的大数据策略，存在三个共同点：一是推动大数据全产业链的应用；二是数据开放与信息安全并重；三是政府与社会力量共同推动大数据应用。

美国。美国从 2009 年开始全面开放了 40 万联邦政府原始数据集。data.gov（美国政府数据库）宣布采用新“开源政府平台”管理数据，代码向各国开发者开放。奥巴马政府将“大数据战略”上升为最高国策，认为大数据是“未来的新石油”，将对数据的占有和控制作为陆权、海权、空权之外的一种国家核心能力。首批共有 6 个联邦部门宣布投资 2 亿美元，共同提高收集、储存、保留、管理、分析和共享海量数据所需核心技术的先进性，并形成合力；加强对信息技术研发投入以推动超级计算和互联网的发展。2013

㊀ 引自：《大数据时代的历史机遇》第三章　赵国栋

年美国发布了《数据开放政策》行政命令，要求公开教育、健康等七大关键领域数据，并对各政府机构数据开放时间提出了明确要求。

已有美国大学专门开设了研究大数据技术的课程，培养下一代的“数据科学家”，一些美国公司也在向大学提供教育研究资助，并赞助与大数据有关的比赛，扩大大数据技术开发和应用所需人才的供给，提高美国的科学发展、环境与生物医药研究、教育和国家安全的能力；美国国家卫生研究院开展的免费开放国际千人基因组计划，它将创建的人类遗传变异研究数据集供研究人员自由访问和使用；美国国家科学基金会和美国国家卫生研究院对大数据进行联合招标，改进核心科学与技术手段，提高从各种大型数据集中提取重要信息并对其进行有效管理、分析和可视化的能力；美国国防部则计划每年投资 2.5 亿美元左右，在各个军事部门开展一系列研究计划，以创新方式使用海量数据，通过感知、认知和决策支持的结合，加强大数据决策能力；美国能源部则将斥资 2500 万美元建立可扩展数据管理与可视化研究所（SDAV），帮助科学家对数据进行有效管理，促进其生物和环境研究计划、美国核数据计划等的研究成果。

此外，美国纽约州能源研究和发展管理局运用一系列大数据技术来评估气候变化对纽约州的影响，并为农业、公共卫生、能源和交通运输等领域提供应对气候变化的策略。这一应用也被引入美国疾病控制中心，它正与美国其他 10 个州和城市一起开展“阅读州和城市计划”，共同研究和应对气候变化，而大数据技术是其中一个非常重要的组成部分。

英国。2011 年 11 月，英国政府发布了对公开数据进行研究的战略决策，英国内阁部长弗朗西斯 · 莫德说，其实英国政府早有意带头建立“英国数据银行”，政府想算清楚究竟这个国家或政府创造了什么；英国不只是要成为世界首个完全公布政府数据的国家，还应该成为一个国际榜样，去探索那些公开数据在商业创新和刺激经济增长方面的潜力。

2013 年 1 月，英国商业、创新和技能部宣布，将注资 6 亿英镑发展八类高新技术，大数据独揽其中的 1.89 亿英镑，超过总额三成以上。2013 年 8 月 12 日，英国政府发布《英国农业技术战略》。该战略指出，英国今后对农业技术的投资将集中在大数据上，目标是将英国的农业科技商业化。

2013 年英国首个综合运用大数据技术的医药卫生科研中心在牛津大学成立，这个研究中心总投资达 9000 万英镑，可容纳 600 名科研人员。中心通过搜集、存储和分析大量医疗信息，确定新药物的研发方向，减少药物开发成本，并为发现新的治疗手段提供

线索。同时，以英国为首的欧洲核子中心（CERN）将在匈牙利科学院魏格纳物理学研究中心建设一座超宽带数据中心。建成后，魏格纳数据中心将成为连接 CERN 且具有欧洲最大传输能力的数据处理中心，未来该设施在处理大型强子对撞机（LHC）的数据以及实验方面发挥重要作用。2013 年 10 月英国政府发布新的政府数字化战略，旨在使政府服务实现“默认数字化”，即“数字服务简单方便，任何可以使用的用户都会选择数字化服务，而不能使用的用户也不排除在外”。2014 年，英国宣布建立图灵大数据研究院，确保英国未来大数据发展在经济和社会中处于领导地位。2015 年，英国政府承诺将开放有关交通运输、天气和健康方面的核心公共数据库，并将投资 1000 万英镑建立世界上首个“开放数据研究所”。

日本。日本面临着由于长期经济低迷导致国际地位下降、人口老龄化以及日益增大的社会保险费用和社会基础设施老化等诸多问题。为了扭转这一现状，日本政府决定通过大力发展 IT 产业，特别是大数据及开发数据和云计算，以发展开放公共数据和大数据为核心的日本新 IT 国家战略，把日本建设成为一个具有“世界最高水准的广泛运用信息产业技术的社会”，并且将其发展成就扩展到国际范围内。

加拿大。随着大数据在全球范围内继续火热，加拿大的大数据产业也在慢慢升温。例如，在科研领域，加拿大政府科学、技术与创新委员会已要求科研组织就与加拿大经济发展和社会福利密切相关的问题，为加拿大政府提出基于证据的科技建议。2007 年，加拿大开始实施数字信息战略。2011 年 5 月加拿大广播电视和电信委员会（CRTC）就发布了新的“国家宽带计划”，该计划显示，到 2015 年加拿大全体国民将享有 5Mbps 的宽带接入速度。2012 年 9 月 IBM 正式启动在加拿大国内兴建的智能数据中心，该中心全称为 IBM 加拿大领导数据中心（IBM Canada Leadership Data Centre）。

法国。虽然法国在数学和统计学领域具有独一无二的优势，但法国的大数据产业发展情况远不如美国、英国等国家发展的火热。但近年来，法国在智慧城市建设方面却投入了大量精力，包括法国电信、施耐德集团和达索集团等诸多法国知名企业都在旗下设立了专门从事智慧城市设计和研发的工作室或实验室，在政府的引导下积极投身智慧城市建设。2013 年 2 月，法国政府发布《数字化路线图》，列出五项将会大力支持的战略性高新技术，其中一项就是大数据。2013 年 3 月，法国国家教育部推出了四项数字化服务。2013 年 4 月，法国经济、财政和工业部宣布，将投入 1150 万欧元用于支持七个未来投资项目，目的在于“通过发展创新性解决方案并将其用于实践来促进法国在大数据领域的发展。”

德国。2013 年，德国制定“数字德国 2015 的 ICT 战略”，在能源、交通、保健、教育、休闲、旅游和管理等传统行业采用现代 ICT 技术实现智能网络化。德国 IT 行业协会 BITKOM 于 2014 年初发表报告称，大数据业务在德国发展迅速，到 2016 年有望达到 136 亿欧元。在此背景下，德国经济和能源部为更好地开发德国大数据的未来市场，支持大数据相关技术的研发创新，启动了“智慧数据 – 来自数据的创新”项目。德国“智慧数据”项目将创新型服务作为研发重点，将中小企业也纳入行动范围，鼓励其开发和使用有吸引力且安全的大数据，力求为中小型企业开发出创新型的系统解决方案。同时，因为严谨的民族习惯，德国在数据保护方面做得非常出色，德国政府还将完善大数据相关的法律法规和基础框架，并选取工业、交通、能源和医疗等领域的典型项目进行孵化和推广，以克服大数据在制度、技术和法律方面的障碍。现在，相关法律对互联网等领域中个人数据的使用都做出了明确规定，还提出设立专职信息保护人员的建议，较好地维护了德国社会的信息安全。

澳大利亚。澳大利亚政府关于数据制定了一系列政策措施，其中包括澳洲公共服务大数据战略等，2013 年更是出台了开放公共部门信息原则（Open PSI），全文提出八条开放数据策略以及实施的难易程度，并提出了需开放数据的首要领域。

印度。印度联邦内阁批准了国家数据共享和开放政策。在数据开放方面，印度效仿美国政府的做法，制定了一站式政府数据门户网站（data.gov.in），把政府收集的所有非涉密数据集中起来，包括全国的人口、经济和社会信息。

2. 国际业界动态

互联网、金融、电信、医疗、政府等是大数据运营的重点领域。而大多数领域的大数据发展应用仍处在初级阶段，在大数据应用的实践过程中也遇到了数据资产不明、应用需求不定、平台建设、技术路线、安全隐私问题等方面的挑战，但是，大数据应用在各领域还是做出了一些有益的探索，并取得了一定的成绩。

在电信行业，一些发达国家电信运营商一方面提升服务质量，改善内部管理。包括客户维系、精准营销和网络运营与管理，这三点的代表企业分别为法国电信、英国 O2、NTT DoCoMo 和沃达丰。法国电信开展针对用户消费的大数据分析评估，借助大数据改善服务水平，提升用户体验。英国 O2 在英国推出了免费 Wi-Fi 服务，以积累更多的用户，从而收集到更多的用户数据，用在精准的媒体广告和营销服务方面。NTT DoCoMo 通过制作精细化表格，收集用户详细信息，大大加强了 CRM 系统和知识库，准确定位目标客户，提高了业务办理的成功性。沃达丰爱尔兰公司的 Tellabs“洞察力分析”服务

是将通信网络中的大数据转化为可利用的情报。

另一方面确立商业模式，创造外部收益。包括直接出售数据获取收益和与第三方公司合作项目给运营商创造盈利，代表企业有 AT&T、西班牙电信、DynamicInsights、Verizon、德国电信和沃达丰。AT&T 将与用户相关的数据出售给政府和企业以获利。西班牙电信成立了动态洞察部门 DynamicInsights 开展大数据业务，为客户提供数据分析打包服务。DynamicInsights 与市场研究机构 Gfk 进行合作，在英国、巴西推出了首款产品名为智慧足迹（SmartSteps）。Verizon 成立了精准营销部门 PrecisionMarketingDivisio，提供了精准营销洞察、精准营销、移动商务等服务，包括联合第三方机构对其用户群进行大数据分析，再将有价值的信息提供给政府或企业获取额外价值，数据业务的盈利在其整个业务中占比非常高。德国电信和沃达丰主要尝试通过开放 API 向数据挖掘公司等合作方提供部分用户匿名地理位置数据，以掌握人群出行规律，有效地与一些 LBS 应用服务对接。

在连锁零售业中，英国最大的连锁超市特易购（TESCO）已经开始运用大数据技术采集并分析其客户行为信息数据集。特易购首先在大数据系统内给每个顾客确定一个编号，然后通过顾客的刷卡消费、填写调查问卷、打客服电话等行为采集他们的相关数据，再用计算机系统建立特定模型，对每个顾客的海量数据进行分析，得出特定顾客的消费习惯、近期可能的消费需求等结论，以此来制定有针对性的促销计划并调整商品价格。这种“有的放矢”的营销和定价模式为特易购提供了更加高效的盈利方法。

在交通运输方面，美国 Inrix 公司和新泽西州运输部之间的合作伙伴关系。Inrix 公司通过汽车和移动电话 GPS 装置上的信号和数据，采集主干道上的车速数据，然后实时向新泽西州运输部警示任意主干道上的路况险情，同时向司机的车载 GPS 装置或移动电话发送警示来提醒司机注意路况险情。

在农业方面，美国气候公司（The Climate Corporation）是一家天气保险公司，他们制作保单来弥补联邦农作物保险和因气候造成的农民损失之间的差额。该公司通过庞大的传感器网络分析和预测 2000 万美国农田的气温、降水、土壤湿度和产量。在知晓高温天的天数以及土壤湿度数据后，建立模型来帮助其预判农民需要的天气保险金额以及公司需要支付的保费。

在外包领域，大数据技术也已成为信息技术行业的“下一个大事件”。目前，一些外包行业巨头也开始进军大数据市场，试想瓜分这一块大蛋糕。印度全国软件与服务企业协会预计，印度大数据行业规模在 3 年内将达到 12 亿美元，是目前规模的 6 倍，同

时还是全球大数据行业平均增长速度的两倍。

在信息安全行业，FireEye 和 Splunk 这类国际企业在大数据安全方面发展迅速，他们在大数据安全方面的技术也值得国内企业借鉴。专做 DLP 产品的 Websense 公司，他们基于数据流的分析技术十分有利于大数据的分析、挖掘。

1.2.2　我国大数据发展动态

1. 我国政府促进大数据发展的措施

2012 年 8 月国务院制定了促进信息消费扩大内需的文件，推动商业企业加快信息基础设施演进升级，增强信息产品供给能力，形成行业联盟，制定行业标准，构建大数据产业链，促进创新链与产业链有效嫁接。同时，构建大数据研究平台，整合创新资源，实施“专项计划”，突破关键技术。工业和信息化部为鼓励和推进大数据产业发展也制定了三大措施。一是在已通过促进信息消费扩大内需的意见、软件和信息技术服务业“十二五”规划等政策规划中，对大数据发展进行了部署。二是推动全国信息技术标准化技术委员会开展了大数据标准化的需求分析、标准体系框架研究及相关标准研制工作，并向相关国际标准化组织提交了大数据研究提案。三是利用项目资金等手段进行了前沿部署，支持关键技术产品的研发和产业化。

各省市也相继部署了大数据战略。广东率先启动大数据战略推动政府转型，上海在 2013 年启动大数据研发三年行动计划，2014 年 5 月 15 日，上海市推动各级政府部门将数据对外开放，并鼓励社会对其进行加工和运用。根据上海市经信委印发的《2014 年度上海市政府数据资源向社会开放工作计划》，将 190 项数据内容作为 2014 年重点开放领域，涵盖 28 个市级部门，涉及公共安全、公共服务、交通服务、教育科技、产业发展、金融服务、能源环境、健康卫生、文化娱乐等 11 个领域。其中市场监管类数据和交通数据资源的开放将成为重点，这些与市民息息相关的信息查询届时将完全开放。上海市经信委方面也公布，上海将参照图书资源的管理模式，力争通过 3 年时间，完成各政府部门的信息系统所承载的信息资源分类、目录编制和注册工作，实现全市政府数据资源目录的集中存储和统一管理，基本摸清政府数据资源的“家底”，从而解决数据资源“有哪些、在哪里、谁负责”等关键问题。今后每一年，上海都会发布年度开放清单，力争用 3~5 年，最终形成数据开放的“负面清单”——只要不涉及国家机密、商业秘密和个人隐私的数据基本全部开放。这意味着企业运用大数据在上海“掘金”的时代来临，企业投资和上海民生相关的产业（如交通运输、餐饮等）可以不再“盲人摸象”。

2013 年 2 月 25 日，国家食品药品监督管理局与百度在北京联合举行“安全用药，搜索护航”战略合作签约仪式。国家药监局的三大药品数据库，总计 20 余万条权威药品信息全面入驻百度。北京也正在积极探索政府如何公布大数据供社会利用，2014 年 7 月 23 日，百度推出名为“北京健康云”的智能医疗平台产品，这是北京市政府支持推动下的一个民生医疗项目。另外，2014 年 5 月 27 日，中国气象局公共气象服务中心与阿里云达成战略合作，共同搭建“中国气象专业服务云”，面向有气象数据需求的企业提供专业化的云计算服务。2014 年 10 月 15 日，“云上贵州”系统平台正式开通运行，这是贵州省政府与阿里牵头的企业合建的云计算基础设施，应用在交通等领域。目前，中国政府对互联网、高科技和大数据产业空前重视，并且明确表态要开放大数据。2015 年 1 月 13 日，阿里健康宣布将药品监管网的基础设施从甲骨文数据库迁移到阿里云平台，阿里将利用大数据技术帮助解决假药问题。

为推动数字福建（长乐）产业园、中国国际信息技术（福建）产业园加快建设成为全省大数据产业重点园区和“数字福建”建设的重要承载基地，福建省政府将从完善园区发展规划、引进培育产业龙头、推动资源汇聚开发、建设大数据创新平台、加强人才引进培养、做好园区用地保障、确保园区用电需求、强化园区网络支撑、实施财税优惠政策、提高安全保障能力等 10 个方面给予支持。在引进培育产业龙头方面，福建省将通过市场开放、资源开发、技术采购、服务外包等方式，大力引进大企业、大行业、大平台，吸引国家部委、央企、电信运营商、金融机构、知名 IT 企业、互联网公司等入园建设新一代信息技术设计、研发、生产基地，发展一批行业性大数据云平台；吸引大数据产业链各个业务环节龙头企业入园，培育一批细分领域全国性领先的大数据服务提供商。

2015 年，中国政府对互联网、高科技和大数据产业更加重视，并且明确表态要开放大数据。2015 年全国人民代表大会和中国人民政治协商会议提出要高度重视大数据采集规则制定、大数据立法和信息保护等方面的工作。

2. 我国大数据企业发展情况

中国移动提出了大数据时代全新的移动互联网战略，即：构筑“智能管道”、搭建“开放平台”、打造“特色业务”与提供“友好界面”，这体现了中国移动在移动互联网时代全面开启之际的全新战略定位。中国移动在 2014 年成立了苏州研发中心，计划构建 3000 ~ 4000 人的研发团队和运营团队，宗旨就是要进一步完善云计算和大数据产品体系，尽快形成国际一流的云计算和大数据服务能力。中国移动构建了大云产业联盟，与

技术提供商、集成商、高等院校、政府机构等超过 50 家单位在核心模块合作、授权技术服务、应用开发技术攻关等产业不同层面开展了合作。

百度、阿里、腾讯、360 等互联网企业依靠自身的数据优势，均已将大数据作为公司的重要战略。大数据正在从理论走向实践，从专业领域走向全民应用的阶段。

百度在大数据方面让人印象深刻的有百度迁徙这样的公益项目，应用在民生和新闻等领域。最新动态是，百度网盟利用基于大数据的 CTR（广告内容匹配）数据使站长的平均收入提升 70%。

阿里则对外宣称已经拥有 100PB 数据并以令人欣喜的速度增长，马云最新的内部邮件将阿里战略阐述为“云端＋大数据”，阿里要进入数据时代。2014 年 10 月 14 日，阿里巴巴集团宣布无线开放战略，启动百川计划。该计划将全面分享阿里无线资源，为移动开发者提供技术、数据、商业等全链条基础设施服务。百川计划是通过阿里无线开发的重要平台产生，并将从技术、数据和电商能力上向移动开发者提供基础服务。其中，云技术层面将提供架构搭建、数据存储、安全防护等服务，并可有专业人员对 APP 进行一对一的开发、维护、技术支持；大数据层面则将联合移动应用统计分析平台友盟，帮助开发者完善数据精准挖掘分析及完善个性化推送体系。

2014 年 10 月 10 日，360 举办首届数字世界大会，并发布三个产品，帮助广告商利用大数据做更有效的营销。360 宣称推出的实效平台、聚效平台和来店通等三款产品就是把集合了数十亿用户信息的数据，免费分享给广告主。

京东在无线领域也正进行着深层次地探索。首先是充分利用移动设备去拓展人性化功能，例如多模式交互，通过京东 APP，用户可以通过移动设备的摄像头拍照、扫码进行购物；另外，京东 APP 还在积极通过大数据技术挖掘用户需求，提供更精准的服务。借助微信能够带来巨大流量的优势，举行京东微信购物的众筹活动，一个月参与的人数就达到 40 万人次。

3. 我国推动大数据产业园区发展

大数据是继云计算、物联网、移动互联网之后的又一个具有国家战略意义的新兴产业。现今，大数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素，据预测，到 2020 年中国数据产业市场将达到 2 万亿以上规模。但任何一个技术概念都需要“落地”，先进技术只有与产业集合，并切实推进经济的发展，才能成为真正的生产力。只有实现真正的产业落地，才能真正推动经济的发展，中国大数据产业才能挤掉泡沫，驶入健康发展的快车道。

产业园区作为产业集群的重要载体和组成部分，其经济效应已引起越来越多的人关注。产业园区能够有效地创造聚集力，通过共享资源、克服外部负效应，带动关联产业的发展，从而有效地推动产业集群的形成。大数据产业园作为大数据产业的聚集区或大数据技术的产业化项目孵化区，是大数据企业走向产业化道路的集中区域。大数据产业园区可以通过自身的规模、品牌、资源等价值为区域经济发展和企业资本扩张起到巨大的推动作用。主要体现在以下几个方面。

提升企业效益。大数据产业园的建立会迅速聚集大数据企业发展所需要的多种资源，可以吸引众多互补型企业、产业链上下游企业等，为企业提供一个良好的发展空间，是企业腾飞的重要平台。

提升地区品牌。大数据产业园的建立必将带来大量高科技企业的入驻，这必将带动地区经济快速发展，为区域经济建设提供高效助推器。另外，随着国家对大数据等新兴技术产业的重视，建立大数据产业园的地区将领先于国家的发展规划之前，提升本地区的知名度，并且可以借此吸引更多高新技术企业的投资。

创造社会价值。由于建设规模较大，涉及投资建设金额巨大，大数据产业园区建成后，在年产值、税收等方面贡献巨大，并可直接解决部分当地失业人员的就业问题。除此之外，园区的生产生活配套设施，如住宿、餐饮、商业区等，不仅可以满足园区内工作人员的个人需求，还可以为地区和其他服务型企业带来巨大的经济利益。大数据产业园在为自己创造经济效益的同时，也获得了社会效益的大丰收。

总之，通过建立大数据产业园，能够更有效地组织和使用大数据，人类也将得到更多的机会发挥科学技术对社会发展的巨大推动作用。

4. 国内外大数据发展对比分析

我国大数据产业起步较晚，同时由于互联网技术也有所滞后，使得我国的大数据发展较领先国家还尚有一段距离。但是，我国又有得天独厚的优势——庞大的用户群，每日有庞大的数据量不断生成，同时，受惠用户量也极为众多。从政策环境上看，我国尚未出台完善的信息数据相关的法律法规，对隐私方面的问题没有明确可执行的规章制度。在政府数据开放方面也亟需进一步加强。

1.2.3 大数据相关社区

2014 年已经迎来了大数据发展的黄金成长期，大数据以“降低信息不对称和提高决

策有效性”为目标，几乎能应用于包括金融、互联网、医学、生物、教育在内的所有行业。然而随着数据量不断剧增，各个国家、各组织被迫寻找创新性的方法来管理、控制并分析利用这些巨量的数据。但是，数据孤岛现象却制约着大数据的发展，例如各个组织之间的数据封闭，甚至同一个组织各个部门之间的数据也不开放。因此数据开放和共享成为大数据的核心和关键。为了推动数据开放和共享，世界各国都在不断寻找新的突破口。

1. 国际大数据相关社区

美国在 2009 年设立“ data.gov”门户网站，将政府和主要州、市等持有的公共数据对外公开。2014 年 1 月，该网站全面改版，目前已经收录了 90154 个数据集、349 个应用程序和 140 个移动应用，参与部门达到 175 个。同时，美国还有 40 个州、44 个县市建立了单独的数据门户。美国的数据开放格式多达 46 种，其中应用最广的格式是 HTML、ZIP 和 XML，数据集分别有 20775 个、12517 个和 11992 个。

英国在 2010 年开设数据开放门户网站（data.gov.uk），政府规定各部委需公开重要公共数据，政府财政支出、犯罪高发区域、市政工程施工等与市民生活密切相关的信息需优先公开，共开放了 13670 个公开的数据集以及 4170 个非公开的数据集。此外，伦敦、曼彻斯特以及索尔福德市议会等 16 个地方和部门还建立了独立的开放数据门户。

2011 年欧盟设立公共数据网站，提出了“欧盟开放数据战略”。

新加坡的数据开放网站（ data.gov.sg）实现了全国范围内的数据整合，是世界上发展最为完善的公共信息资源开放共享网站之一，目前已经汇集了来自 68 个政府部门和机构的 8600 多个数据集。这些数据可以被企业和个人访问，帮助他们发现机遇并提供服务。

印度政府建立了一站式官方数据门户网站 (data.gov.in)，使其成为政府收集的所有数据的总网站。该网站可以把政府收集的所有非涉密数据集中在一起，包括全国的人口、经济和社会信息，共开放了 5811 个数据集，共有 58 个部门和 4 个州参与，开放了 24 个应用程序，在 5811 个数据集中，以 XLS 格式开放的有 1793 个，以 ZIP 格式开放的 4 个，以 CSV 格式开放的 2087 个，以 HTML 格式开放的有 30 个，以 XML 格式开放的有 1897 个。

2011 年 9 月 20 日，开放政府合作伙伴（ Open Government Partnership，OGP）成立，该组织由巴西、印度尼西亚、墨西哥、挪威、菲律宾、南非、英国、美国八个国家联合签署《开放数据声明》。截至 2014 年 2 月 10 日，全球已有 63 个国家加入开放政府合作

伙伴。

2013 年 6 月，八国集团首脑在北爱尔兰峰会上签署《开放数据宪章》，法国、美国、英国、德国、日本、意大利、加拿大和俄罗斯承诺，最迟在 2015 年末按照宪章和技术附件要求进一步向公众开放可机读的政府数据。

除政府部门外，民间团体也在积极推进公共数据的灵活利用。2012 年 7 月日本非盈利性机构 OKF（Open Knowledge Foundation）诞生，旨在促进公共数据的有效利用，其开发的开源软件 CKAN 对于构筑地方政府的公共数据门户网站起到重要支撑作用。

2. 中国大数据相关社区

中国政府同样也在做着开放、共享数据的努力，建立了类似英美政府的国家数据公开网站——国家数据网（data.stats.gov.cn）。我国各地也开始了公共信息资源的开放共享，北京、上海等地数据开放网站相继投入运行。

北京市政务数据资源网通过原始数据（WPS、CSV）下载、带有地理坐标信息的空间数据（SHAPE）下载和在线调用 API 三种形式提供数据开放共享服务。开发者、分析研究人员、普通用户通过互联网登录该网站可查询、下载数据以及在线调用 API 开发服务。该网站已经整合了北京市 35 个政府部门的信息，为社会提供了土地使用、教育、旅游、交通、文化、医疗等 269 类、多达 36 万条的原始数据资源。

上海市政府数据服务网（datashanghai.gov.cn）作为外界获取数据的统一入口，自 2012 年起对外试用。上海市政府召开推进政府数据资源向社会开放工作会议，要求在包括市公安局、市工商局、市统计局、市商务委员会等 9 家试点政府部门公开信息资源的基础上，所有政府部门在 2014 年内通过“上海政府数据服务网”向公众提供数据产品浏览、查询和下载等服务，具体涵盖公共安全、公共服务、交通服务、教育科技、产业发展、金融服务、能源环境、健康卫生、文化娱乐等 11 个领域。

中国香港均依托原有网站，通过进一步增强网站服务功能或者增设网站栏目的办法，实现政府数据的开放。

中国社会团体也纷纷自发成立数据共享中心，以期借助大数据技术寻求新的商机。其中有以推动数据开放为目的的数据开放平台；有为推动大数据技术进步而成立的大数据技术社区；也有为推动大数据的商业化应用而成立的大数据平台。

（1）数据开放社区

百度数据开放平台。这是以百度框计算先进技术理念为基础，连接站长优质数据和百度搜索结果的开放平台。站长通过结构化优质数据，获得百度搜索结果页“即搜即得”

的搜索展现。

全国可信网站数据库开放平台。这是由中网依托“可信网站”验证数据库为基础搭建的全国最大的可信网站数据库开放平台。该平台收录对象先由网站自行申请或公众推荐（含合作伙伴等），然后由中网按照“可信网站”验证审核规则或流程对商家或网站的身份信息完成交互审核后予以收录。该平台将面向浏览器、搜索引擎、电子商务平台、微博等诸多领域实现免费开放，帮助各大互联网平台快速、高效地实现平台上各类商家、网站的真实身份验证，形成无所不在的商家、网站身份验证查询途径，方便网民按照上网习惯快速便捷地甄别网站真伪，识破网络钓鱼骗局。

百度开放研究社区[㊀]。提供了 100G 的百度真实的数据集以及由 250 台服务器组成的开放研究云平台，并利用网络社区没有地域限制、自由、灵活、快捷等特点，为科研人员提供一个互动交流的平台，让技术沟通和海量数据分析变得简洁而高效。

开放爬虫系统。主要面向高校和科研单位的大数据研究团队，提供个性化定制的互联网数据获取服务。该系统底层采用众包模式进行数据获取，数据采集具有广泛性、时效性和精准性的特点。开放爬虫系统是中科院高能所开放大数据平台的一部分，开放大数据平台收集、整合与管理互联网、物联网、科研等多个领域的数据，着力解决大数据应用开发中海量数据来源、数据传输、基础能力支撑等问题，实现多领域数据交叉融合，提供一站式数据服务，激活数据价值，降低应用门槛。

（2）大数据技术社区

炼数成金[㊁]。成立于 2011 年，是一个关于数据仓库、数据挖掘、商业智能等技术和业务讨论的数据分析专业社区网站。该社区通过将大数据各应用领域的业务专家、数据分析专家、IT 专家以及这些领域的从业人员汇聚起来，为大家提供大数据技术及应用讨论、培训的平台。

ITPUB 社区。属于 IT168 旗下网站，是国内知名的数据库社区，目前注册会员数量已达 170 多万，该论坛以探讨数据库管理技术和信息化技术以及各种开发技术为主，希望汇聚社区的力量，吸引更多的人才推动大数据的发展。

1.2.4　我国大数据行业协会

IT 技术的迅速发展使爆炸增长的数据成了各行各业共同面对的问题。为了有效应对

㊀ http://openresearch.baidu.com/?locale=zh_CN

㊁ http://www.dataguru.cn/

大数据引起的挑战，同时充分利用大数据带来的机遇，2012 年 5 月，以“网络数据科学与工程——一门新兴的交叉学科”为主题的第 424 次香山科学会议在北京香山饭店成功召开，会议建议在中国计算机学会下成立大数据专家委员会。该委员会于 2012 年 10 月正式成立。此后，业界对大数据日趋关注，中国大数据行业协会纷纷成立，中国通信协会、中国电子协会等也成立了大数据专家委员会，各界人士共同推动了我国大数据技术及产业的发展。

中国计算机学会大数据专家委员会。成立于 2012 年 10 月，由中科院计算所李国杰院士担任主任。其宗旨包括三个方面：探讨大数据的核心科学与技术问题，推动大数据学科方向的建设与发展；构建面向大数据产学研用的学术交流、技术合作与数据共享平台；对相关政府部门提供大数据研究与应用的战略性意见与建议。大数据专家委员会下设有五个工作组，分别负责专家委员会的会议组织（学术会议、技术会议）、学术交流、产学研用合作、开源社区与大数据共享联盟以及大数据发展战略工作组。

中国通信学会大数据专家委员会。成立于 2012 年 10 月，由中国通信学会牵头组建，是我国首个专门研究大数据应用和发展的学术咨询组织。其主要任务是组织大数据发展重点问题研讨会并提出有关建议；开展大数据相关理论、方法、实践课题的研究；为企业的大数据研发提供咨询服务；促进产业间的资源共享与合作。专家委员会主任委员由中国工程院士、中南大学校长张尧学院士担任，来自政府部门、学术界、研究机构和企业的知名专家学者担任委员。

中关村大数据产业联盟。成立于 2012 年 12 月 13 日，由中关村管委会直接领导。联盟的宗旨是把握云计算、大数据与产业革新浪潮带来的战略机遇，聚合厂商、用户、投资机构、院校与研究机构、政府部门的力量，通过研讨交流、数据共享、联合开发、推广应用、产业标准制定与推行、联合人才培养、业务与投资合作、促进政策支持等工作，推进实现数据开发共享，并形成相关技术与产业的突破性创新，产业的跨越式发展，推动培育世界领先的大数据技术、产品、产业和市场。

中国国际经贸大数据研究院（中国大数据与智慧城市研究院）。成立于 2013 年 3 月 22 日，隶属于国家工业和信息化部中国电子商务协会，是国家工商总局和国家发改委信息中心共建共管单位。主要任务是开展大数据应用与智慧城市建设专项课题研究、组织开展大数据专业技术人才培训、承担科技产业大数据应用开发任务、提供智慧城市解决方案。

“数据中心联盟”（Data Center Alliance）。成立于 2014 年 1 月 16 日，是在工信部

通信发展司的指导下，由电信研究院联合国内三家基础电信运营商、十余家主要互联网企业、国内主要硬件制造企业以及若干科研单位和组织等单位共同发起组建的。目前联盟共有会员单位 71 家，其中全权会员单位 30 家，高级会员单位 4 家，普通会员单位 37 家。联盟的主要工作是为了推动数据中心、云计算和大数据的技术和业务研究，掌握国内国际数据中心行业动态，为政府和企业制定发展战略提供依据；建设数据中心和云计算的信息服务平台，提供行业、市场等公共信息服务等。

中国电子学会大数据专家委员会。成立于 2014 年 4 月，凝聚了 112 位多个学科的专家。这样的交流平台和机制为业界提供了技术支持，可以有力地推动国内大数据的发展，为我国信息产业和服务的发展提供强大的动力。

中国信息协会大数据分会。成立于 2014 年 8 月 27 日，是中国信息协会为了落实党的十八大关于促进“新四化”同步发展的有关精神而组建。其宗旨在于联合国内政产学研各界，共同把握全球大数据发展的有利时机和战略资源，促进我国大数据产业的发展。

中国法律大数据联盟。成立于 2014 年 11 月 18 日，将联合包括大数据、云计算、移动互联网等各领域的专家资源和学术优势，为法律大数据的深入研究和应用提供前沿的理论支撑，更好地发挥法律大数据和法律云在国家治理、国土安全和网络空间安全以及在经济和社会发展中的重要作用。联盟成立中国法律大数据研究中心，编制《中国法律大数据蓝皮书》，组织法律大数据学术研讨会，构建法律大数据与云服务平台，对法律大数据与云技术进行深度分析、挖掘，探究法律大数据在法治国家、法治政府、法治社会一体化建设中的应用。

第 2 章　大数据典型应用

大数据作为一种赋能性技术，如同电一样，作用于经济社会的各个层面。

任何一种技术的应用都要经历从简到繁、由浅入深的过程，大数据的应用仍然遵从这一发展路径。数十亿用户在互联网上的处处留痕、时时留迹，使得其在网络空间的活动画像日益丰满，从而其需求很容易被准确洞察，精准营销仍为大数据最具产业规模的领域。

围绕大数据精准营销产业链，互联网、金融、电信、新媒体等领域的大数据技术产品创新此起彼伏，应用广度不断拓宽，应用深度不断加强。同时，电网、交通、医卫、地信、政府、农业领域的大数据应用也明显提速。

2.1　互联网大数据

2.1.1　互联网大数据应用现状

随着互联网普及率的不断提升以及移动互联网的快速发展，互联网应用的发展趋势也在不断发生转变，发展重心从“广泛”转向“深入”，对大众生活的改变从点到面，对网民生活的全方位渗透程度进一步增加。互联网应用的深入发展产生了海量的大数据，大数据是互联网的重要资源，也是互联网商业模式的核心价值所在，因此，大数据理论和技术在互联网应用中起到至关重要的作用。互联网应用的多样性导致其涉及的大数据内容呈现不同的特点，针对不同需求研究和采用适宜的大数据技术能够获得更好的互联网应用和服务，从而提升用户体验，带动互联网的整体发展。

根据中国互联网络信息中心 2014 年 7 月发布的最新一期《中国互联网络发展状况统计报告》，互联网应用主要分为四类：商务交易类应用、信息获取类应用、交流沟通类应用、网络娱乐类应用。各应用领域分别包括不同的应用场景，其中绝大多数互联网应用涉及大数据相关技术。特定互联网应用具有其固有特点，例如，增长率较高的与支付相关的商务交易类应用，要求大数据技术具有更强的针对性。又如，移动互联网的快速发展使数据本身发生了变化，也使大数据技术的应用面临新的机遇和挑战。通过深入分析

互联网应用的特点，不断改进和完善大数据技术，使其与互联网应用更加紧密地结合，能够让数据本身为互联网应用带来更高的附加价值。

以下就大数据技术与互联网应用相结合的典型场景进行深入讨论。

2.1.2　大数据应用于互联网商务交易

截至 2014 年 6 月，我国网络购物用户规模达到 3.32 亿，网上支付是用户规模增长速度最快的商务类应用。网络购物用户规模的增长除了得益于商务部政策和新《消费者权益保护法》等对于电子商务市场的规范活动之外，很大程度上得益于电商平台服务的提升，尤其是企业基于大数据应用推出 C2B 定制化创新模式，更好地匹配了用户个性化需求，实现精准销售。

在互联网时代，由于用户群体庞大，当前电子商务平台必须面对海量的应用大数据。与此同时，用户提出了越来越强的信息过滤和个性化的需求。要想匹配用户个性化需求，实现精准销售，需要借助大数据技术在海量数据中提取精准信息，其中首要任务是充分分析数据特征。大数据时代，随着人们的生活全面向互联网和移动互联网转移，随之而来的是信息过载（information overload）问题，大量信息位于所谓的“长尾”区域。从电子商务的角度来看，由于货架成本极其低廉，因而商品总数很大。比如在淘宝网上，每天在线的货品超过 8 亿件。传统上的 2/8 原则（即 80% 的销售来自于 20% 的热门商品）受到了挑战。虽然绝大部分商品不热门，甚至得不到曝光的机会，但它们的数量极其庞大，因此这些长尾商品的总销售额将是一个不容忽视的数字。从用户需求的角度来说，这部分具有特异性的商品往往对应他们的个性化需求。

智能推荐系统是商务交易类互联网应用中典型的大数据应用实例，其目标是通过发掘用户行为，找到用户的个性化需求，帮助用户发现那些感兴趣但很难发现的信息。它以大数据为基础，应用个性化技术，帮助用户从海量信息中筛取所需信息。当用户需求明确的时候，他们会进行搜索；而当用户需求不明确的时候，则需要推荐。相比于广告系统直接提高收益的目标，推荐系统是在满足用户体验的基础上间接创造价值。推荐系统的应用场景主要有三个：第一个是根据物品推荐物品，典型的例子是在电商网站上推荐已购买过的商品；第二个场景是为用户推荐物品，最典型的是个性化邮件过滤和基于浏览历史的推荐；第三个场景是为用户推荐用户，像社交网站上的“好友推荐”和电商网站上的“跟您相似的顾客”等都属于此类。

大数据技术在智能推荐系统等相关商务交易类互联网应用中以不同的形式发挥作

用，按照其所基于的数据类型可以分为三大类，分别是基于内容的方法、基于协同过滤的方法和组合方法，特点如下。

基于内容的方法。处理对象包括用户和物品信息，对数据进行特征化表述，并收集用户是否选择过某个物品的数据，把推荐问题转化为分类问题。基于内容的方法简单且直接，适合处理冷启动问题。

基于协同过滤的方法。基于用户行为分析的推荐算法，用户通过协作，即不断地和网站互动，使推荐列表不断过滤掉自己不感兴趣的物品，从而越来越满足自己的需求。协同过滤类推荐方法的适用性较广，更倾向于推荐比较流行的物品，较难实现推荐的多样性。

组合方法。同时实现两个或者多个不同的方法并组合最终结果。

大数据技术在互联网商务交易领域应用十分广泛，在智能推荐系统等重要应用中起着至关重要的作用。但是必须指出的是，互联网商务交易中大数据处理仍然存在一定的挑战，主要体现在以下几个方面：

数据稀疏性问题。由于互联网应用中有效数据所占比例较低，因此在极端不均衡数据上会出现参数抖动严重等情况。

数据规模问题。在很多应用场景中，有千万级的用户和百万级别的物品，用户 / 物品关联矩阵达到百亿甚至十万亿的规模，对于大数据存储、处理和计算是很大的挑战。

冷启动问题。这一类问题没有固定的解决方案，需要针对不同的应用场景提出不同思路。

评估的多标准问题。数据评估标准有很多，包括精准性、覆盖率等定量指标和实效性、健壮性等定性指标，需要根据应用场景制定大数据处理原则，通过合适的技术构建系统并完成评估。

2.1.3 大数据应用于互联网信息获取

搜索引擎是最主要的互联网信息获取类应用，截至 2014 年 6 月，我国搜索引擎用户规模达 5.07 亿，使用率为 80.3%。2014 年上半年，搜索引擎创新技术的实际应用取得了一定进展，企业基于“语义搜索”与“知识图谱”技术，整合社交、视频、旅游、软件应用下载等多类信息，开发并上线新的搜索产品。搜索引擎在 PC 端及移动端均形成了以搜索产品为核心，集地图、娱乐、购物、社交、本地生活服务等应用为一体的搜索服务，提升了用户体验和使用黏性。

搜索引擎天生就是一个大数据系统，互联网产生了海量数据，如何从中找到需要的信息就是一个大数据的命题。同时，利用大数据理论和技术，通过对网民搜索内容、习惯、爱好、行为、关键词等的深入分析，可为网站的建设和搜索引擎技术的改进等提供依据。

搜索引擎的诞生从一定程度上满足了用户在海量互联网数据中查找信息的需求，但还存在很大的可优化空间。传统的搜索引擎根据查询词返回相关网页链接，还需要用户自己阅读大量网页内容才可获得所需信息，对于手机等移动终端用户而言很不方便。

为了克服传统搜索引擎的弊端，人们正尝试探索更高效、更人性化的搜索引擎技术，如直接搜索或知识图谱搜索。直接搜索是指提问或输入关键词后系统直接提供答案，而不是包含答案信息的链接或相关文档。系统自动完成答案抽取，帮助用户快速定位所需信息，从而可节省用户阅读大量网页或文档的时间。知识图谱则是展现搜索词相关知识，通过结构化的方式予以呈现，让用户可以快速、全面地了解相关信息，增强使用体验。为了达到这一目的，首先需要理解用户查询的问题，同时要了解问题所对应的答案。答案需要从海量垂直网站或者用户搜索词中挖掘，挖掘方法包括基于半监督学习的新词及专名挖掘、面向垂直领域定向抽取的三元组挖掘、基于用户行为日志数据的实体关联挖掘和基于搜索引擎的实体语义标签挖掘等。

抓取并索引的网页数量是衡量搜索引擎质量的重要因素之一。如今，百度、必应、谷歌等主流搜索引擎都要抓取数以千亿计的网页，同时索引数百亿的网页，以提供良好的搜索服务。为了处理如此巨量的数据，MapReduce、Hadoop 等大规模数据处理系统应运而生，利用这些系统，搜索引擎公司就能高效地计算网页的各项特征，为索引数十亿计的网页打下基础。此外，在线系统也已演化为高度并行的容错系统，以保证在有上百万用户同时使用的情况下，搜索引擎仍可在 1 秒内为绝大多数查询返回结果。

为了进一步提高搜索效果，搜索引擎越来越多地引入自然语言处理技术和知识库技术。和传统方法相比，这些技术更加复杂、计算量更高，因此要求大数据系统提供更高的计算能力。改造大数据系统的途径有很多，例如，进一步增加系统中节点的数量、充分利用 CPU 的多核能力、利用显卡的运算能力，甚至直接使用 FPGA 执行定制化的处理算法等。另一方面，进一步理解查询和用户意图 / 兴趣变得越来越重要。与理解网页不同的是，利用查询历史理解查询和用户意图 / 兴趣常常需要用到基于图和矩阵的算法，这些算法和 MapReduce 式的计算框架并不完全契合，因此需要研究新的计算框架，以便提供支持。

2.1.4 大数据应用于互联网交流沟通

随着国外社交网站 Facebook、Twitter、LinkedIn 等的发展以及国内的微博、微信等社交工具的不断壮大，基于社交网络的各种互联网交流沟通类应用不断演进和发展。单就我国互联网应用现状来看，截至 2014 年 6 月，即时通信网民规模达 5.64 亿，使用率仍高居互联网交流沟通类应用第一位。与此同时，社交网站使用率则持续下滑，前景不容乐观。以上两方面现象的根本原因都是数据流向决定用户黏性。社交网络之所以吸引人，是因为其用户产生了大量有价值的用户数据（User Generated Content，UGC），而且这些数据能和一个个活生生的人对应起来。因此，对社交网络上产生的各种用户数据进行分析，是社交网络分析中极其重要的一个方面。

无论是即时通信、社交网站、微博还是博客，都是网民交流的平台，每天产生大量的数据。通过对社交网络中的大数据进行分析，可以了解用户的思维习惯及其对社会的认知。对微博等社交网络信息空间的大数据进行挖掘，能够及时反映社会的动态与情绪，预警重大、突发和敏感事件（如流行疾病爆发、群体异常行为等），协助提高社会公共服务的应对能力，对维护国家安全和社会稳定具有重大意义。

社交网络中一个典型的大数据应用场景是舆情分析，即对热点事件在网络上的传播过程加以监测，了解人们的态度，从而在必要的时候加以干预和引导。在当前这个社交网络高度发达的时代，公众对于很多问题都有发言的欲望，加之社交网络上信息传播的快速性和不可控性，舆情分析对于政府、商业实体和公众人物都有着重要的意义。对于舆情分析的支持系统来说，处理逻辑往往相对简单，比如按照特定关键字对相应的内容进行过滤，其中最关键的技术要求就是对海量数据的实时处理，这就需要高性能的大规模并行处理数据库或者流数据处理系统的支持。另一方面，社交网络上的用户产生的大量内容往往存在着大量的隐含信息，对这些信息进行综合处理往往会得到非常有价值的信息，比如对热点事件的预测或预警。在社交网络时代，信息发布的门槛较低，信息传播速度快，各种真实和不真实的信息在社交网络上随时都可以形成爆炸式的传播。利用大数据技术收集社交网络上传播的数据内容，分析其背后隐含的意义，能够对特定的事件进行预测。

此外，影响力分析也是社交网络中需要运用大数据技术进行处理的关键问题，用户在社交网络中的行为（比如对微博的评论、转发等）也代表了影响力，例如新浪微博上的大 V 账号，其粉丝动辄上百万，发布的任何一条微博的转发数和评论数都极其巨大，换句话说，他们有着巨大的影响力。在利用大数据计算影响力的时候，需要将相应的数

据考虑进来，对社交网络节点的影响力进行量化就是影响力分析要解决的问题。影响力分析的一个扩展应用是用户搜索。社交网络将人们的互联网生活和虚拟生活融合在一起，人们在社交网络上形成了各种社区，产生了大量不同领域的内容。通过其所属的社区和其产生的内容，可以定义一个社交网络账号的属性，为其打上特定标签，为其他用户可以方便地通过关键字进行搜索。

对社交网络大数据的分析和挖掘仍需要解决以下两个重要问题：

第一由于微博、博客等数据源自多个服务提供商，具有不同的属性和不同的表达方式，是典型的多源异构数据，因此需要使用基于语义模型的表示方法，以实现对社交网络数据的有效表示和理解。

第二对社会事件、流行病等相关内容、信息和数据，实时性要求非常强，需要强有力的流式处理、增量处理等技术手段提供支撑，并且需要建立快速反馈机制。

2.1.5　大数据应用于移动互联网

移动互联网虽然发展较晚，但发展速度要远快于互联网。截至 2014 年 6 月，我国手机网民规模达 5.27 亿，网民中使用手机上网的人群占比进一步提升，由 2013 年的 81.0% 提升至 83.4%，首次超越传统 PC 网民规模。随着移动设备的功能越来越强大，移动互联网与传统互联网之间的差异愈发不容忽视。移动互联网具有互联网的传统特征，但也有其自身的固有特征，特别是和时间信息相对应的地理位置信息是其独有的数据特征。

移动互联网最大的特点是以用户为中心。一个移动设备对应一个用户，包含位置、联系人、浏览内容等所有信息。数据本身的价值在于更完整和更生动地描绘了用户的生活轨迹。互联网时代，内容的组织不以用户为中心，所以主要商业模式是广告，跟内容相关而不是跟用户相关。移动互联网则能够进行更好的数据整合，属于同一个用户的数据都可以关联在一起。以用户为中心，就能够方便地把属于一个用户的散乱在不同应用里的信息整合起来进行分析。

对于大数据分析来说，移动互联网的特殊性首先是能够锁定一个特定用户，其次是能够获取用户地理位置信息，再次是时空信息等多元化的数据种类。由于这三点，导致移动互联网上的数据数量比传统互联网更大，形式也比传统互联网更加丰富，从而有更高的价值。移动大数据有三个特点：

数据的核心节点是人。随着各种移动设备、物联网和云存储等技术的发展，人和物

的所有轨迹都可以被记录。与互联网不同的是，在移动互联网中的核心网络节点是人，而不再是网页。用户在移动终端上的所有行为是具有一定延续性的，这使得用户档案（profile）的建立成为可能。

用户地理位置等上下文信息。移动互联网上能获取的最重要的信息就是用户的地理位置，通过地理位置信息与服务数据的结合，能够更加精确地分析用户行为，实现用户描述精准化。

数据量更大，时空数据维度更高且更复杂。移动互联网需要统计很多互联网网页分析所没有的数据，例如设备型号、应用版本、推广渠道甚至位置信息，同时还有很多开发者自定义的事件。在移动互联网的数据中，文字以外的其他信息占到更加重要的比例。从数据的属性上来讲，移动互联网上的数据更加复杂，其中一个原因就是这些数据包含了大量时间和空间信息，需要把普通数据挖掘延伸到时空数据挖掘的领域。因为多了一个维度，时空数据挖掘的复杂度比一般的数据挖掘又深了一层。

移动应用是抢占用户移动终端桌面的主力军，是用户获取信息和休闲娱乐的主要方式，是用户上网的主要入口。对于移动应用来说，大数据技术的应用主要在于如何通过数据挖掘改善产品体验、实现差异化竞争和产生商业价值。移动应用分析的目标是理解用户如何与移动应用进行交互，其分析内容包括：

用户获取、活跃、留存、转换分析。统计新用户、活跃用户、一次性用户以及用户属性信息（如设备信息、地理位置、运营商）；统计用户的日 / 周 / 月留存率、应用使用时长、使用频率等，以通过分析提高用户忠诚度。

用户分群分析。对用户按人口属性、兴趣、地理位置、使用情况、留存情况、付费情况等分类。通过用户群划分，开发者可以根据自身需求设置出不同的用户群，这一过程实际上就是通过一定设定条件将用户筛选出来。

跨应用和跨平台分析。除了前面提到的单应用内数据统计分析外，通过跨应用大数据分析能够了解目前发布的所有应用的数据情况，让开发者了解自身应用在行业内所处的位置和现状。

移动互联网上的数据除了拥有特殊价值外，也给大数据分析带来了很多挑战。主要体现在以下几个方面：

数据质量。首先是数据采集的质量，在移动数据采集过程中，网络不稳定等状态导致移动端数据无法较好地上传，这使得移动大数据分析需要更复杂的数据补偿策略。其次是数据的噪音和稀疏性更强，移动端的应用数以十万计，且在每个应用中两个用户之

间的重叠非常少，很难根据特征通过分类模型对移动用户的人口属性进行分类预测。最后是移动互联网本身的各种作弊行为导致数据不准确，大量移动应用通过刷量来冲击移动互联网应用排行榜以获取投资人的青睐。大量移动互联网公司付费给水军来给自己的应用发好评，给竞争对手的应用发差评。这些数据所占比例过高，已经严重干扰了数据的准确性，大大降低了移动互联网数据的整体价值。

用户时空行为模式的挖掘和利用。移动互联网应用之间差异较大，用户对于应用的选择更是千差万别，深入挖掘用户的行为模式能更准确地抓住用户喜好，同时也是当前大数据技术在移动互联网中面临的重大挑战。

跨应用、跨平台多维数据的交叉利用与用户隐私。目前移动端的应用多处于信息孤岛的状态，并没有真正实现数据的互联互通。数据的整合分析无疑能够带来全新的应用环境和用户体验，如何在保护用户隐私的情况下进行用户全局数据互联互通和交叉利用将成为移动大数据分析的挑战。

2.1.6　互联网大数据发展趋势

从以上典型应用场景可以看出，互联网与大数据互相依托，互联网是产生大数据的最主要的平台。在互联网应用中，大数据源源不断地产生，通过分析、处理反作用于互联网应用。因此，大数据技术是互联网发展的动力，大数据技术使互联网应用更加贴合用户需求和网络发展方向，从而不断发展壮大。

大数据最主要的发展趋势是与移动互联网结合，针对移动互联网固有特征，改进自身技术以更加适应移动互联网应用需求，实现移动互联网和大数据的有机结合，并且渗透到人们日常生活的各个角落，真正达到以大数据和移动互联网影响和改变人们生活的目的。短期内可以预见的是，大数据技术将在移动电子商务、即时通信、社交平台、移动网络游戏等应用领域中迅速发展并寻求突破，成为相关领域的核心技术。

2.2　金融大数据

2.2.1　金融大数据应用现状

从数据角度看，金融无非是各种数据的排列组合。金融大数据是指高度依靠数据资源实现资金融通、支付和信息中介等业务的一种新兴金融。具体可以表现在两个层面：一是技术层面，强调金融业务的完全电子化，基于结构化和非结构化的海量数据进行业务运作；二是商业层面，强调运用大数据资产和大数据思维经营金融，基于数据进行决策。

金融大数据和互联网金融不能完全割裂来看，互联网金融强调运用互联网技术和互联网思维在运营层面和结构层面改造金融业，而金融大数据更强调运用大数据资产，大数据可以称为互联网金融的内核。同时，大数据与互联网都是信息技术的历史产物，二者存在着密切的关联性。

金融大数据相对于传统金融呈现出以下五方面的新特征：

新的技术手段。传统技术主要处理结构化的数据，在大数据时代，新的信息技术需要低成本且高效率地处理海量结构化和非结构化的数据。

新的服务功能。中国的多层次金融服务体系尚未建立，金融大数据能够有效地服务传统金融业的市场空白，满足中小企业和“屌丝”客户的金融需求。

新的金融业态。基于数据资产的各类新型金融业态层出不穷，如大数据投资、大数据保险、大数据征信等。

新的参与主体。伴随着政策制度红利的释放和信息技术创新的推进，各类“搅局者”涌入金融市场，凡是掌握数据和客户的企业都有可能介入金融服务。

新的竞争核心。营业网点是传统金融机构竞争的重要要素，金融大数据更强调数据资产的价值，数据将成为继土地、劳动力、资本之后的新型要素。

在中国，抛开金融电子化，互联网金融的概念最早是在 2012 年由中国投资有限责任公司副总经理谢平教授首次提出，他将既不同于商业银行间接融资也不同于资本市场直接融资的第三种金融融资模式称为互联网金融。2013 年互联网金融概念的热度开始急速攀升，6 月 17 日，阿里巴巴推出“余额宝”产品，该产品上线 6 天，用户数就突破 100 万，上线 18 天，累计用户数达到 251.56 万，累计转入资金达到 66.01 亿元，截至 2014 年 7 月，“余额宝”资金规模已突破 6000 亿元，客户数超过一亿户，创造了令金融界震惊的奇迹，随后不仅出现了“活期宝”和“现金宝”等众多“宝”类产品，同时也拉开了互联网金融的大幕，产业界、投资界、学术界乃至监管部门开始纷纷加大在互联网金融领域的布局，P2P、众筹、网络小额信贷、比特币等新兴的互联网金融模式层出不穷，一时间，互联网金融成为社会各界争论的焦点和抢夺的重点。

金融大数据与互联网金融相伴而生，2013 年、2014 年进入快速发展期。以 P2P 借贷为例，在 2006 年、2007 年前后 P2P 网贷就已进入中国，自 2011 年开始，中国 P2P 贷款呈现出爆发式增长的态势，大量的 P2P 贷款平台在市场上涌现。2011 年 P2P 网贷平台数量仅为 20 家，月成交额仅为 5 亿元。但到 2013 年底，网贷平台数量增加到 600 家，月交易额达到 110 亿元，截止 2014 年 6 月，P2P 网贷数量突破 1200 家，平均每天成立

一家平台。网贷投资人规模也从 2012 年的 5 万人增加到 2014 年 6 月份的 29 万人。

随着金融大数据的野蛮生长，相关监管力度有待提高。目前我国尚无一部专门的法律对个人信息数据特别是个人金融信息的收集、使用、披露等行为进行规范，主要通过宪法和相关法律法规对个人信息进行间接保护，但立法散乱且不成体系。因此需尽快明确监管职责与权限，对金融大数据进行合理规范和监管，保障其稳定、持续与健康的发展。

由于金融大数据的基础缺失，因此数据孤岛、数据壁垒现象明显。企业数据必须要流动才能产生价值，而静止数据只是看上去宝贵，实则敝帚自珍，最终将被时间淘汰。我国金融机构往往由于业务领域的不通抑或是政策法规的限制，导致不同机构掌握的数据不尽相同，而且数据的流通性较差。数据孤岛对于数据的量级、时效性、流通性产生了极大的影响。打破数据壁垒，实现数据的有效和及时流通是金融大数据发展的基础。

大数据是金融业的底层基础设施，大数据时代的到来，将对金融业的各个方面都产生广泛而深远的影响。本节将分别针对大数据对信贷、征信和投资的影响展开论述。

2.2.2　大数据信贷

目前，信贷服务多是在企业产生信贷需求以后再进行信贷审核，以企业当前的经营状况来进行信贷风险判断。但矛盾的是，很多企业产生信贷需求的时间点恰是企业经营的低点，由此造成企业所获得的信贷资金往往和其需求不匹配；而且单以一个时间点的状况对企业进行判断，无法掌握企业后续的经营走势，不利于风险防范。

正是基于这种思考，在阿里小贷数据运用中，如何透过数据判断小微企业的信用等级、未来经营走向和融资需求（包括其需求出现的时间节点和量）成为了重点。

在整个阿里小贷业务决策中，数据分析处在核心位置。其中，上百个基于电商数据建立的模型交叉生效，构成整个风险决策体系的核心。目前，阿里小贷的模型体系涵盖贷前、贷中、贷后管理、反欺诈、市场分析、信用体系和创新研究六大部分。在众多模型中，有一个特殊的模型，它像是一个水晶魔球，专门负责预测未来，叫作“水文模型”。

试想，在淘宝网这个电商生态系统里，每家店铺的销售数据就像一条条河道里的水文，会随着市场的热度、客户的喜好、卖家的运营以及季节周期性的波动而忽高忽低。不同的淘宝店铺尽管销售的产品千差万别，但可以根据行业、主营类目、星级进行归类划分，这好比不同的店铺出现在河道里的不同位置，可以纵向对比他们的水位高低。如果能提前预知店铺未来的交易变化，就可以提前预判其资金需求的节点、量级，而且能横向对比出店铺运营的好坏，尽早识别出店铺运营潜在的风险。

那么，如何预测店铺未来的经营状况呢？这涉及交易、流量、收藏、广告、评价、处罚、社交、运营等综合反映店铺经营情况的变量，需要将单一变量通过各种衍生获得一系列变量。将这些变量从原来的绝对值变量转变为相对值变量，剔除季节性对绝对值的影响，就能使不同规模、类型的卖家具有相同的可比性。对于一个卖家来说，在不同的月份，这些变量的表现就会如同水文一样呈现变动，称为“水文变量”。

在构建水文模型时，淘宝摸索出了解决两大世界级数据分析领域难题的方法，一是如何进行大浪淘金般的数据挖掘，二是海量数据如何在云计算平台进行加工运算处理。

目前，直接或间接反映店铺情况的基础指标就有 70 多种，涉及交易类、流量类、收藏类、广告类、评价类、处罚类、社交类、运营类等。淘宝按类目、星级分别统计一个“水文数据”库，且在库中记录过去的静态常数及其分布情况。有了总体的统计常量，就可以把每个店铺在各自所属的类目和星级做比对，进行各种衍生，如同比、环比增长以及在同行业中的排名和变化趋势，总计几百种衍生方式，从而得到为数众多的相对性指标（即自变量），同时对自变量进行标准化，使之脱离量纲，变得可比。目前水文库中存在的水文指标已超过三万个。

庞大的数据需要有足够的运算能力支撑。在构建水文模型的过程中，所有数据放在一起有数百 T 之大。传统的数据分析方法根本无法直接处理，只能靠抽样损失信息来处理。而阿里云平台提供了强大的云计算功能，对数亿条数据进行分析时，在传统数据挖掘平台上需要数天甚至数周才能完成运算，而在阿里云平台上却只要数小时就能获取结果。

正是依赖于这样的云计算和大数据，阿里小贷的纯信用贷款模式才能顺利运作，在完全无抵押、无担保的情况下为小微企业提供资金支持。

2.2.3 大数据征信

征信（credit checking/credit investigation）是指依法收集、整理、保存、加工自然人、法人及其他组织的信用信息，并对外提供信用报告、信用评估、信用信息咨询等服务，帮助客户判断、控制信用风险，进行信用管理的活动。

信息网络技术历来在个人征信行业的发展历程中扮演着异常重要的角色。征信行业的产品除了信用服务咨询外，主要组成部分是信用报告，所以信用信息数据的获得与分析处理对征信行业来说至关重要。近年来随着大数据技术的推广，可以将用户的信用交易记录同步转化为信用记录，一方面极大地节约了数据收集、传递和整合的时间，拓宽了数据的来源，降低了数据处理的成本；同时，凭借大数据技术的大存储量、快速查询

等功能，可以实现信息收集、查询与存储的自动化、智能化，为大规模的个人信用评分系统的建立创造了可能。

ZestFinance 是一家金融大数据分析服务的提供商，由前谷歌 CTO 和管理高层 Douglas Merrill 和美国第一资本投资集团 Capital One 公司前高管 Shawn Buddle 联合创立，总部位于美国加州洛杉矶。该公司主要利用机器学习模型和大数据分析为发薪日贷款行业提供客户信贷评分，旨在为那些个人信用一般或不满足传统银行贷款资格的个人重新进行信用评级，使其能够享受正常的金融服务，主要的合作伙伴有 Spot Loan 等公司。2014 年 2 月该公司推出了新一代大数据模型，据称可以提高 30% 的数据收集量。未来，通过整合发薪日贷款所积累的经验与技术，该公司的业务可以拓展到其他贷款领域，如信用卡、汽车贷款甚至包括房屋贷款等领域。

ZestFinance 形成信用报告的数据来源主要包括两部分，一是用户的社交数据，这需要在获得借贷人授权的前提下，由 Facebook、Twitter、Linked In 等第三方社交平台提供；二是用户的行为数据，这也是 Zestfinance 与行业内其他企业的不同之处。目前同行业里大多数公司极其注重社交数据对于客户信用评估的影响，而 Zestfinance 对社交数据并不看重，其收集的来源极其广泛，涉及众多看起来好像毫不相关的边缘数据和非结构化数据。通过社交数据与行为数据的整合与分析，ZestFinance 形成信用评估，并以报告的形式向金融机构出售（见图 2-1）。只有满足发薪日贷款业务的经营方的信用指标，借贷人方可取得贷款。

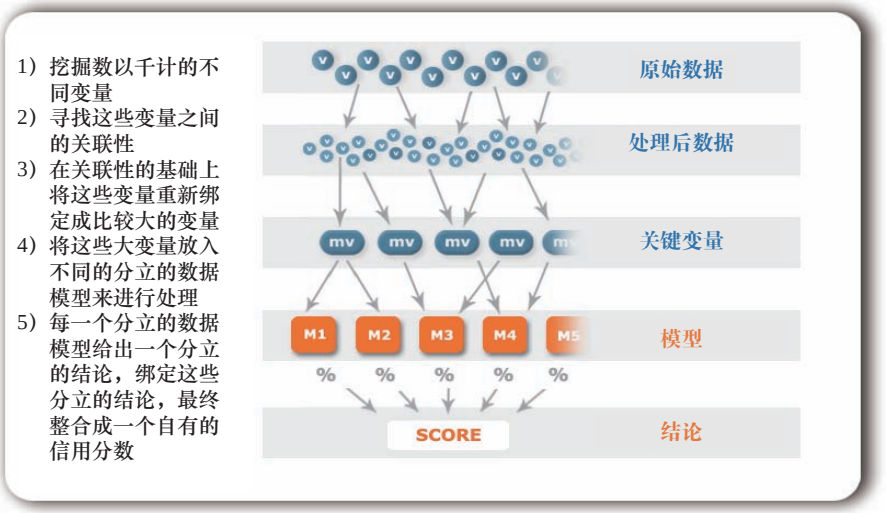


图 2-1　ZestFinance 分析模型

2.2.4 大数据投资

一直以来人们都希望在纷繁复杂的股票市场上找到规律，数以亿计的股民是二级市场的终端消费者，他们的市场判断与投资情绪对整个股市产生着不可忽视的影响。然而在人们的情绪信息不可见也无法收集的时代，人们无法通过这种方法对股市进行判断。直到社交网络的兴起，自媒体成为人们表达个人观点的重要平台。海量的交互信息在这里汇集并以可计算的数据形式储存，为云计算进行情感分析并有效地预测股市走向提供了可能。

2010 年美国印第安纳大学的 Johan Bollen 等人就对 Twitter 上的文章能否预测股票市场进行了研究[㊀]。他们制作了一个叫作 GPOMS 的情绪追踪工具，将 Twitter 上的文章分为冷静、警惕、确信、活力、友善和幸福六个心情类别。其中，如果将“冷静”情绪指数后移 3 天，其曲线与道琼斯工业平均指数能够高度契合，在图 2-2 中灰色阴影范围内这种匹配的趋势更加显著。该文章称，使用这种方法预测道琼斯工业指数的涨跌变化准确率可以达到 87.6%，预测值的平均误差百分比减少超过 6%。但是，仍然有一些事件对股市的影响是无法预测的，比如 10 月 13 日的股市意外高涨，就是由于美国联邦储备委员会突然启动的一项银行纾困计划造成的，而之前在 Twitter 上的民众没有意识到任何征兆。

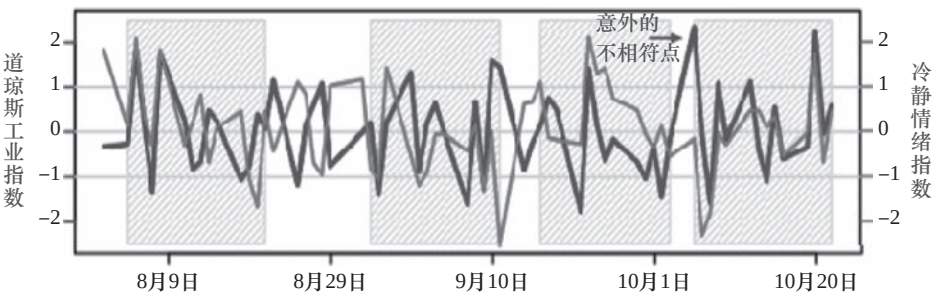


图 2-2 冷静情绪指数与道琼斯工业指数[㊁]

2012 年社交媒体监测平台 DataSift 也进行了关于情绪倾向对股市影响的测试。在 Facebook 首次公开募股当天，Datasift 收集了 58 665 位用户产生的 95 019 条关于

㊀ 资料来源：Bollen J, Mao HN, and Zeng XJ. (2010). Twitter mood predicts the stock market, Arxiv working paper, Indiana University。

㊁ 资料来源：Bollen J, Mao HN, and Zeng XJ. (2010). Twitter mood predicts the stock market, Arxiv working paper, Indiana University。

Facebook IPO 的博文，并对每条博文的情感倾向进行定义和分析。结论显示，Twitter 讨论中情感倾向的变化都会在 5 ~ 20 分钟内反映到 Facebook 的股价波动上（见图 2-3）。

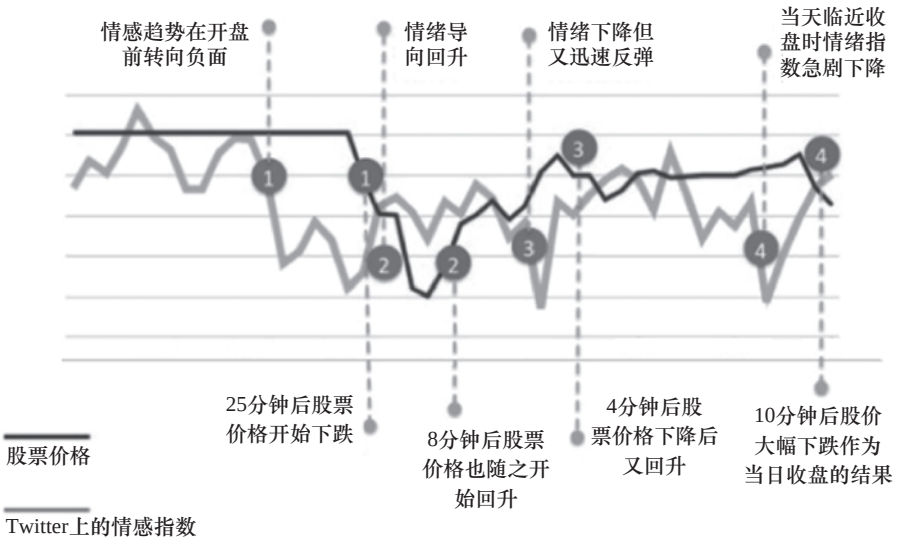


图 2-3　Facebook 发行首日的股价与 Twitter 上的情绪变化指数

2.2.5　金融大数据发展趋势

交易中介脱媒、服务中介弱化是金融大数据未来商业演变的终极形态。金融脱媒一词出现较早，又称为“金融非中介化”，是指资金融通绕开商业银行这一媒介体系，表现为企业组织借助股票、债券等金融工具实现直接融资，居民的金融资产从储蓄为主向证券资产转变。技术脱媒一词则诞生于互联网等信息技术兴起之后，是指在新兴技术的作用下金融体系呈现的“去中介化”的特征，具体表现为交易中介脱媒和服务中介弱化。

目前，我国的金融脱媒和技术脱媒正在并行推进，二者结合的化学效应更为凸显。我国的资金融通主要是通过交易中介和服务中介两类中介机构，交易中介主要包括银行、证券公司等机构，交易服务主要包括会计事务所、律师事务所、投资咨询公司等机构。交易中介提供两种融资方式：一是通过银行的间接融资方式，这也是中国当前主要的资金融通方式；二是通过证券公司进行股票或债券的直接融资方式（见图 2-4）。这两种资金融通方式对于促进经济增长和资源配置起到了非常重要的作用，但同时也产生了巨大的交易成本，直接体现为银行等金融机构的利润，据中国银监会的统计显示，2012 年中国商业银行的净利润近 1.24 万亿元，较 2011 年增长了近 18.96%，远高于 2012 年

GDP 增长速度 7.8%。

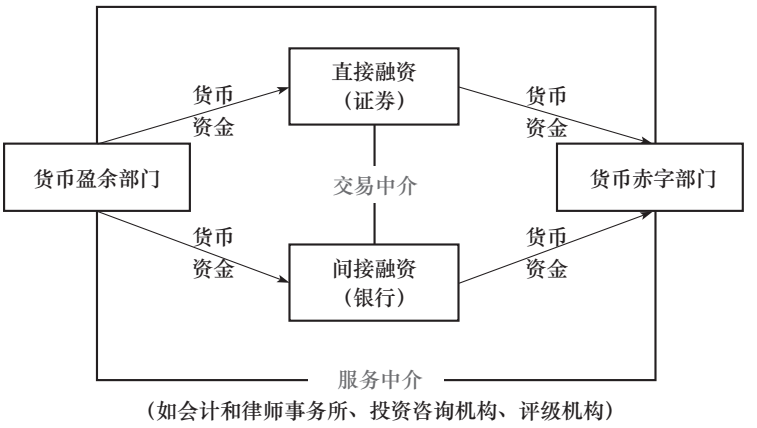


图 2-4 中国传统金融市场的资金融通方式

银行等金融中介的存在有两个主要前提：一是交易费用的存在，金融交易是跨地域和跨时间的，不确定性很大，旨在降低风险和不确定性的交易费用很高，金融中介通过专有技术可以实现规模经济；二是信息不对称的存在，导致逆向选择和道德风险，金融中介通过信息生产加工和账户监控等方式可以缓解这两个问题。

搜索引擎、社交网络、物联网、移动互联网、云计算、大数据等新兴信息技术改变了传统的信息产生、传播、加工利用的方式，打破了信息不对称，降低了信息获取和加工成本，这将加速交易中介的脱媒化进程。未来的金融模式将实现资金供求双方的自由匹配，且是双向互动社交化（见图 2-5）。但金融业不仅存在信息不对称，同时也存在知识不对称，金融产品具有风险性特征，因而个性化的解决方案咨询仍有市场。不过 IT 可以将人类知识结构化，且随着机器学习、IT 智能的发展，服务中介的部分功能也会逐渐被 IT 智能支持所取代。

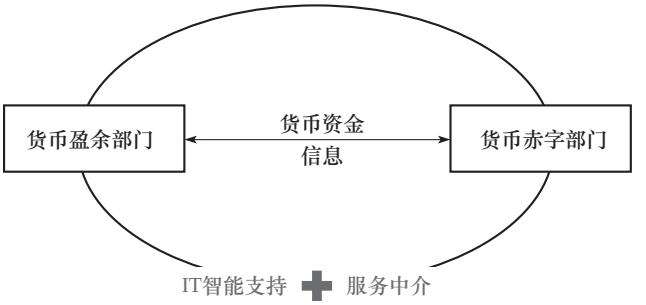


图 2-5 未来的金融运作模式

2.3　电信大数据

2.3.1　电信大数据应用现状

电信行业已经有上百年的历史，当前正面临着大数据带来的新机遇和新挑战。以中国移动为例，每日从事务性系统中产生的结构化数据就达到 8TB，汇聚的经过压缩后的上网日志数据达到 400TB，而最原始的信令量更是达到数 PB 到数十 PB。如果将这些数据视为成本的话，那么电信运营商将面临巨大的投资压力。但从另外一个角度上考虑，如果这些数据能像资产一样产生收益、增值保值，那么无异于拓展了一个全新的领域。

电信运营商为了提供更好的网络通信质量和更灵活的计费方式而建设了一系列的 IT 系统，比如网络管理系统、深度包分析（DPI）系统、信令分析系统、计费系统、客户关系管理系统、企业信息管理系统（MIS）。这些系统原本的目的是用于内部管理，但是其不经意累计下来的海量数据却可以用于其他领域，用于增强电信运营商本身的商业模式，或者让其他行业或企业的商业模式更具竞争力。这就是大数据典型的“数据外部化”特性，即数据的价值可能发挥在 IT 系统设计者所意想不到的地方，因此任何企业和 IT 系统都应当尽可能地留存数据。

此外，电信运营商在提供固化、移动通信、宽带等基础电信服务的基础上，也提供一系列增值业务，例如音乐、图书、下载、动漫、支付等。这些互联网和移动互联网的应用在为用户提供服务的同时也积累了大量的信息。

由于电信运营商本质上是提供信息沟通渠道和桥梁的，因此从一开始就非常重视数据在生产和经营环节中的指导作用，大数据的兴起更是为电信运营商拓展了一个全新的领域。电信与媒体市场调研公司 Informa Telecoms & Media 在 2013 年的调查结果[㊀]显示，全球 120 家运营商中约有 48% 的运营商正在实施大数据，其他没有实施大数据的运营商有 60% 准备在两年内实施，如图 2-6 所示。

此外，该调查还显示：大数据成本平均占运营商总 IT 预算的 10%，并且在未来五年内将升至 23% 左右，移动运营商更加积极地布局大数据，但也不会比全业务运营商投资更快。如图 2-7 所示。

㊀ Big Data: Building the next business platform for telecoms operators, Informa Telecoms & Media。

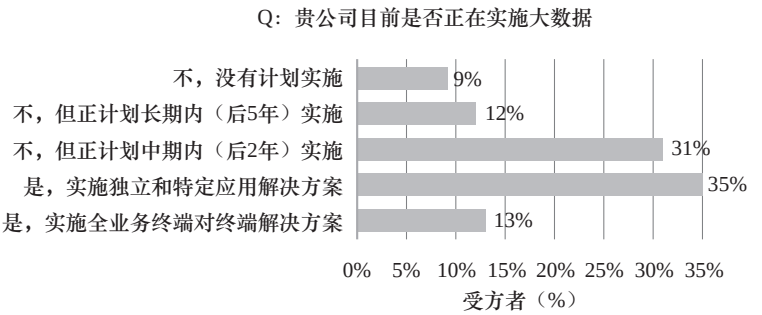


图 2-6 运营商大数据实施情况调查结果

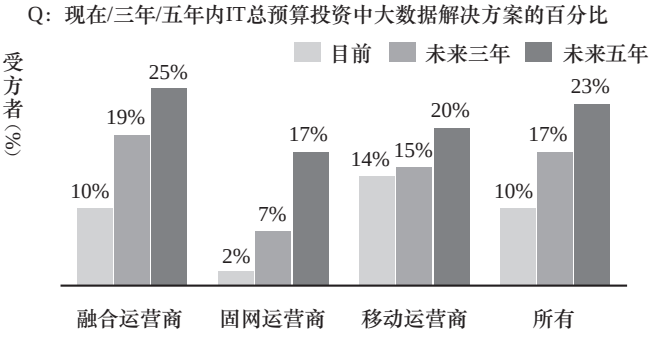


图 2-7 投资额调查结果

大数据在电信运营商中的应用可以分为对内和对外两个部分。对内利用大数据来增强自身的商业模式，包括利用大数据进行网络管理和优化、细分市场产品研发、精确营销、客户挽留、企业精细化管理等。对外方面，国内外电信运营商正在尝试将数据货币化，利用大数据发现新的商业模式，发挥数据的外部社会和经济价值。

以下按照电信运营商在大数据领域的应用分类展开讨论。

2.3.2 电信运营商的网络管理和优化

电信运营商所拥有的电信运营网络是其核心能力，如何利用大数据技术进行网络管理和优化成为研究热点[㊀]。在网络管理和优化方面，主要包括以下两大方面：

基础设施建设的优化。如利用大数据实现基站和热点的选址以及资源的分配，运营商可以通过分析话单和信令中用户的流量在时间周期和位置特征方面的分布，对 2G、3G 的高流量区域设计 4G 基站和 WLAN 热点；同时，运营商还可以对建立评估模型对

㊀ 引自：《大数据在电信行业的应用》，傅志华。

已有基站的效率和成本进行评估，发现基站建设的资源浪费问题，如某些地区为了完成基站建设指标将基站建设在人迹罕至的地方等。

网络运营管理及优化。在网络运营层面，运营商可以通过大数据分析网络的流量、流向变化趋势，及时调整资源配置，同时还可以分析网络日志，进行全网络优化，不断提升网络质量和网络利用率。

在这种场景里，主要利用大数据技术实时采集处理网络信令数据，监控网络状况，进而基于监控到的大数据，实现相关应用。国内国外运营商在此方向均有较多的应用，实际案例如下。

德国电信建立预测城市各区域无线资源占用情况的模型，根据预测结果，灵活地提前配置无线资源，如在白天给 CBD 地区多分配无线资源，在晚上则给酒吧地区多分配无线资源，使得无线网络的运行效率和利用率更高。

中国移动辽宁公司基于大数据挖掘技术进行感知掉话分析。掉话率是通信保持类的重要性能指标之一，也是公司重点考核的 KPI 指标之一，它反映了终端用户在正常接入信道后由于各种原因没能正常结束通话的概率，掉话率过高会直接影响用户感受。早期掉话的判定主要基于话务统计数据，然而一部分用户的真实体验并不能完全在该数字上体现出来，这会导致掉话指标很好，但用户的感知却很差。该公司通过对网络信令的数据分析，实施采集数据合并、字段筛选、数学建模等过程，通过随机森林的挖掘算法找出网络中存在的用户感知掉话，以解决网络中的隐性问题，提高用户满意度。以沈阳地市掉话率分析为例，通过大数据分析技术得到的掉话率在话统掉话率的左右。图 2-8 和图 2-9 为感知掉话热力图与用户投诉的对比情况，可以看出，感知掉话率与用户投诉情况较为一致。

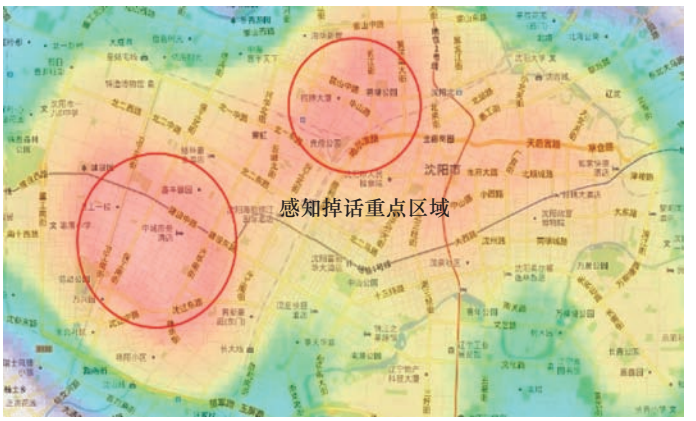


图 2-8　感知掉话热力图

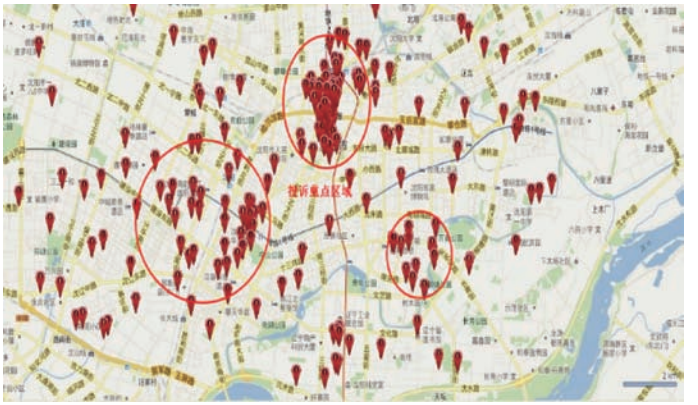


图 2-9 掉话投诉用户地理分布

法国电信通过分析发现某段网络上的掉话率持续过高，借助大数据手段诊断出通话中断产生的原因是网络负荷过重，并根据分析结果优化网络布局，为客户提供了更好的体验，获得了更多的客户以及业务增长。

2.3.3 电信运营商的精准营销

电信运营商的精准营销应该是应用大数据最为广泛的领域，此方向包括客户画像、关系链研究、精准营销等。

客户画像。运营商可以基于客户终端信息、位置信息、通话行为、手机上网行为轨迹等丰富的数据，为每个客户打上人口统计学特征、消费行为、上网行为和兴趣爱好标签，并借助数据挖掘技术（如分类、聚类、RFM 等）进行客户分群，完善客户的 360° 画像，帮助运营商深入了解客户行为偏好和需求特征。

社交圈研究。运营商可以通过分析客户通讯录、通话行为、网络社交行为以及客户资料等数据，开展交往圈分析。尤其是利用各种联系记录形成社交网络来丰富对用户的洞察，并进一步利用图挖掘的方法来发现各种圈子，发现圈子中的关键人员，以及识别家庭和政企客户，或者分析社交圈子以寻找营销机会。

精准营销和实时营销。运营商在客户画像的基础上对客户特征进行深入理解，建立客户与业务、资费套餐、终端类型、在用网络的精准匹配，并在推送渠道、推送时机、推送方式上满足客户的需求，实现精准营销。

【案例】 中国移动江苏公司流量套餐营销

移动语音通信市场逐渐趋于饱和，基于语音通信的短信、彩铃、彩信等增值业务

已经有了明显的下降趋势，移动互联网的发展、智能终端的普及、App 智能应用的层出不穷和终端性能的不断强化将数据业务流量提升到新的档次，流量成为最有价值的增长点。实时把握用户的流量需求，进行相应套餐的匹配推荐，将会带来互赢的效益。

中国移动江苏公司首先梳理当前流量套餐产品线，将融合套餐和纯流量套餐进行全量组合，去除冗余后，得到当前备选套餐组合。然后基于用户的流量使用量以及变化趋势，依据回归算法进行趋势分析以及波动分析，对其流量生命周期进行划分。最后对于处于不同生命周期的用户，依据其 ARPU、本地通话、长途通话、漫游通话、流量使用情况，对各备选套餐组合进行欧氏距离测算，选择最佳适配套餐组合。

2.3.4　电信运营商的数据变现

数据变现指通过企业自身拥有的大数据资产进行对外商业化，获取收益，主要包括精准广告投放以及基于大数据的监控决策等。

目前国内外运营商的数据变现都处于探索阶段，相对来说，国外运营商在这方面发展得更快一些。比如，西班牙电信成立了名为“动态洞察”的大数据业务部门，推出的首款产品名为“智慧足迹”，基于完全匿名和聚合的移动网络数据，可对某个时段、某个地点人流量的关键影响因素进行分析，并将洞察结果提供给政企客户。德国电信和 Vodafone 在利用大数据为自身业务服务之余，已向商业模式跨出了一步，主要尝试是通过开放 API，向数据挖掘公司等合作方提供部分用户匿名地理位置数据，以掌握人群出行规律，有效地与一些 LBS 应用服务对接。美国 Verizon 则成立了精准营销部门（precision marketing division），提供三方面的服务，首先是精准营销洞察（precision market insights），提供商业数据分析服务；其次是精准营销（precision marketing），提供广告投放支撑；最后是移动商务（mobile commerce），主要面向 Isis（Verizon、AT&T 和 T-Mobile 发起的移动支付系统）。

【案例 1】 中国移动江苏公司“智慧洞察”对外数据服务平台

中国移动江苏公司建设了“智慧洞察”（smart insight）对外数据服务平台。平台依托企业数据中心强大的处理能力与海量数据，基于完全匿名和聚合的移动数据，利用统计分析、数据挖掘等技术，向客户提供标准化数据产品、大数据分析报告、高效 Open API 服务，为社会、政府、企业以及家庭、个人客户提供经过分析挖掘而形成的价值产品与服务，实现数据价值提升与共享。

江苏公司大数据对外服务平台支持多元化服务支撑方式，包括标准化产品、大数据

分析报告和数据开放 API，而所有服务模式都基于企业数据中心的统一数据服务工具，其平台架构如图 2-10 所示。

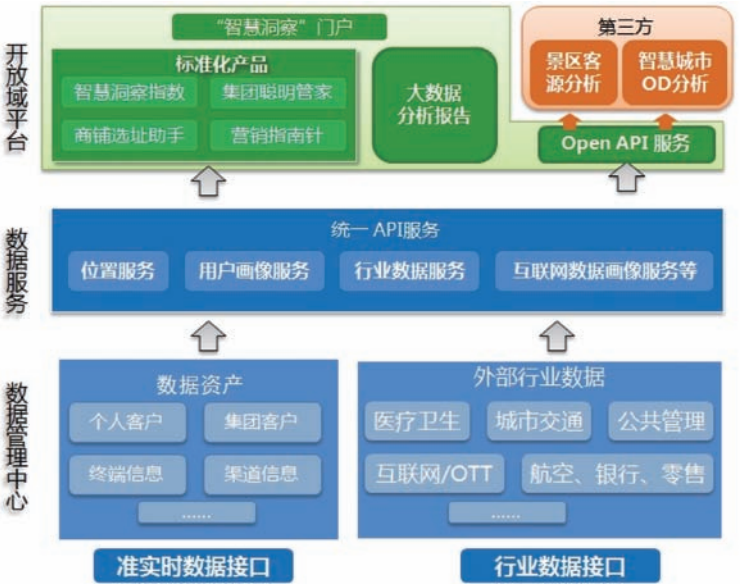


图 2-10 中国移动江苏公司“智慧洞察”平台架构

标准化产品。根据特定客户群或者具体行业的业务特征进行数据分析，并封装成独立产品，给客户提供方便快捷的产品和应用，助力客户提升核心竞争力。

大数据分析报告。对数据中心相关数据进行综合分析，挖掘市场的各类热点，形成高价值分析报告，为广大消费者提供消费向导，为客户的产品分析提供全新研究思路，为企业提供决策辅助依据。

数据开放 API。面向第三方开发者、合作伙伴等提供数据开放服务，对数据中心海量数据进行统计分析，将分析结果信息进行数据封装，满足第三方的大数据共享需求，助力行业客户实现大数据应用的快速开发。API 可用于支撑 ICT 为客户提供定制化方案，也可直接面向第三方提供服务。

“智慧洞察”依托大数据对外服务平台提供服务，而其背后则是企业级大数据平台强大的数据采集、处理和分析能力。例如，用户的位置信息可通过语音话单、GPRS 话单、A/Mc 口信令、Gb 口信令等方式获得，但每种方式都有所缺失，只有整合所有数据来源并进行分析和加工，才能对用户时空序列特征进行全面、统一的描述。

“智慧洞察”的具体应用案例如下：

客源分析标准产品。目前已应用于常州中华恐龙园、苏州金鸡湖、南京夫子庙、玄武湖、太湖等数十家景区。其主要功能点包括：

- 客源构成分析：分析人群构成，区分出真正的游客。
- 景区景点人员密度分析：通过基站分析各景点实时客流情况，便于疏导与管理。
- 流量监控与预警：提高景区管理职能、服务能力及安全保障能力。

央视在大数据类栏目引用此平台界面与数据报道假日期间景区的游客情况，如图 2-11 所示。

图 2-11　假日期间景区游客情况

数据开放 API。对外数据服务平台研究通过数据开放方式实现数据价值的输出，同时考虑用户信息安全、数据传输实时性等要求，形成以终端数据、位置数据、流量数据为重点业务的对外数据开放 API 服务。

目前位置类 API 已应用在南京旅委“南京智慧旅游数据运行监测中心”项目中，在 2014 年度江苏省智慧旅游推进会上，此项目被江苏省旅游局评为“江苏省智慧旅游优秀项目”。其中通过 API 提供大数据服务作为基础服务为旅委管理提供如下服务：

- 实时分析当日客源数据和瞬时客流量，为节假日景区舒适度指数提供数据。
- 对重点景区、特色景点客流情况进行实时监测，提高应急响应水平。
- 支撑数据监控中心，实现客源分布、景区舒适度指数等监控数据的实时呈现。

此外，该项目与无锡旅游委也已达成合作协议，将通过 API 聚合类数据提供深度服务。

【案例 2】 中国电信广告精准投放平台

根据艾瑞咨询对中国网络广告市场的分析：2013 年，市场规模达到 1 100 亿元，同比增长 46.1%，在目前的市场预期下，2015 年市场规模或达 2 000 亿元，未来三年互联网广告产业会进入 30% 年增长的水平，并在更长的一段时期里保持成熟发展。

从国际和国内的情况来看，在网络广告中的程序化投放广告，特别是 RTB 实时竞价的广告投放呈上升趋势（见图 2-12）。

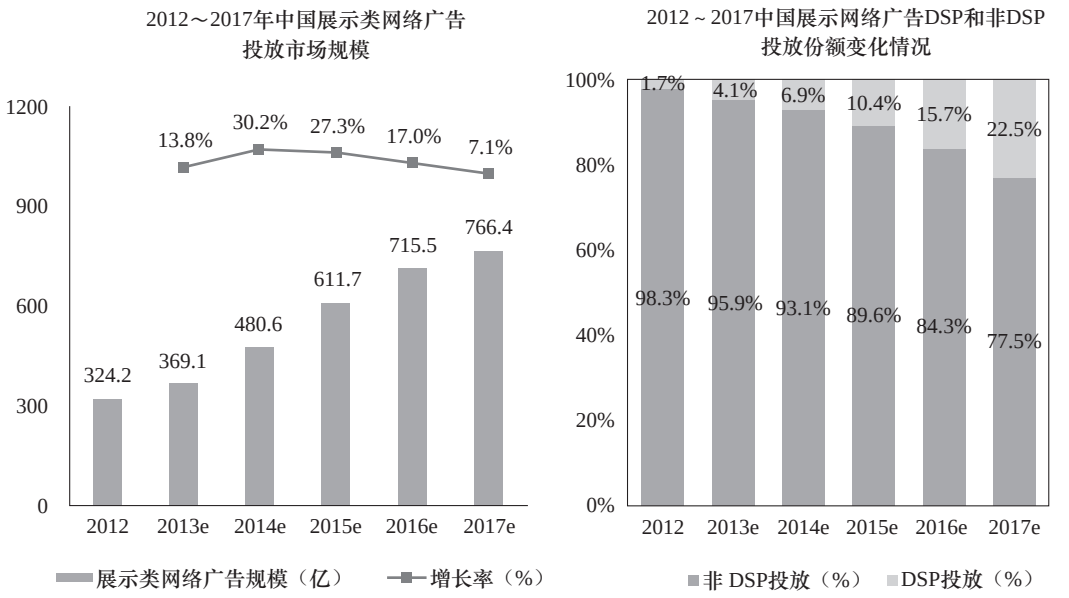


图 2-12　中国网络广告市场分析

程序化投放类广告中最具潜力的就是根据媒体、受众、广告内容三者实时最优匹配的 RTB（实时竞价）广告。它的原理大致如图 2-13 所示。

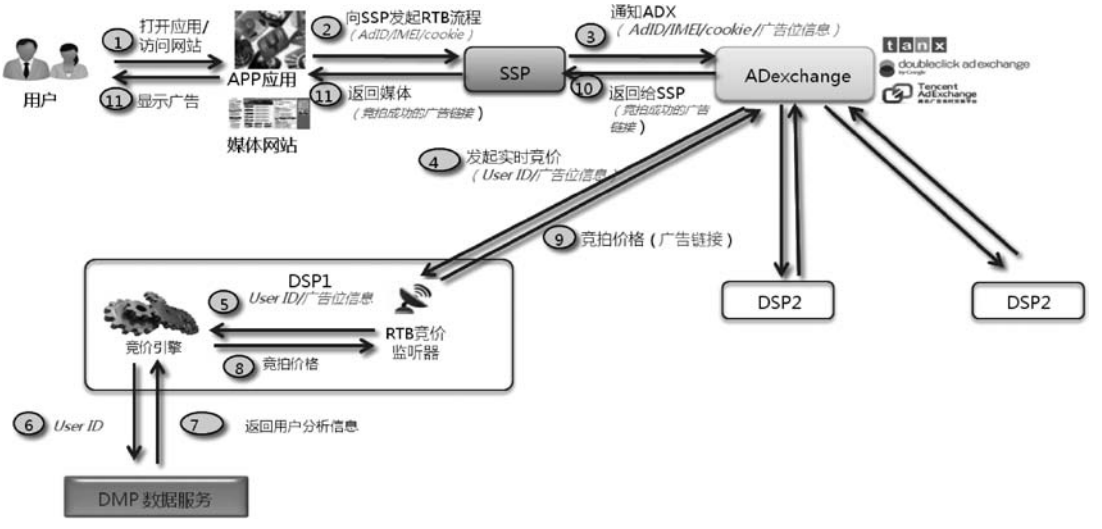


图 2-13　最优匹配原理

出现消费者广告位所在应用 / 页面时，Ad Exchange 就会实时通知 DSP，DSP 需要在 60 ~ 300ms 内决定要不要竞价这个广告。一旦竞价成功，广告就会展示出来。DSP 后端连接 DMP，为了让 DSP 在要求的时间内完成响应，以便参与竞价，DMP 的响应效率也需要相应匹配。当 DSP 竞价时，实时返回竞拍价格和广告链接，以支持广告展示，同时也可以同步返回监播地址，实时监控消费受众在广告位的行为，以进一步调整和优化广告策略。

因此，在 RTB 广告中，最关键的就是如何尽可能地识别受众和分析受众，并能打破时间、空间和终端的分割，以形成对受众全方位的洞察。原有互联网模式下基于 cookie 的链接和判断方式在移动互联网模式下存在难以跨屏、cookie 生命周期短、受众行为信息缺失等种种问题。而电信运营商对“管道”中的数据进行 DPI 解析，通过深度分析用户行为，可以识别用户的 cookie、IMEI、计费代码等信息，有效帮助 DSP 更加精确地投放广告。根据数据显示，基于运营商 DMP 进行精确投放的 CPC 类广告，点击率相比未区分受众的模式提高了一个数量级。

中国电信 2009 年试点精准广告业务，主要是通过分光系统 DPI 与外部媒体进行合作，形成精准广告推送业务，共有 11 个省开展了试点业务，并取得了比较好的经济效益。2013 年又启动了大数据 RTB 广告业务。

以上海电信举例（见图 2-14），其在电信机房搭建了一个上百节点的 Hadoop 集群，将来自城域网（其中涵盖固网宽带、企业宽带和手机宽带的流量数据）的 DPI 结果以及 AAA 认证服务器的认证记录进行关联、预处理、去隐私和压缩。然后将这些数据和数据处理的基础能力（MapReduce 和 Hbase）以多租户的模式提供给每家 DSP 供其生成自身出价所需的客户标签信息。然后这些 DSP 接入来自 ADX 的广告媒体流量，进行出价。

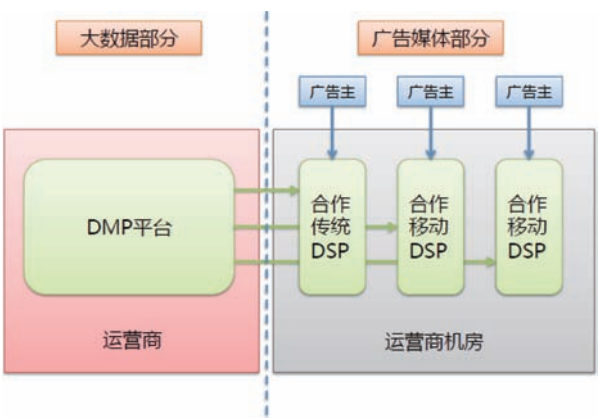


图 2-14　上海电信广告投放模式

目前上海电信的 DSP 有四家合作伙伴：集奥聚合、亚信、晶赞科技、MediaV。数据和技术服务费当年收入约为 1 500 万元，预计 2015 年翻倍。

基于受众的 RTB 广告模式适宜某些品牌和效果类广告的投放。以游戏类广告举例，

某游戏类广告投放时，无精准定向的 CPA 为 80 元，业内平均精准效果为 30 元，利用运营商大数据优化后，CPA 降低到 19.6 元。

【案例 3】 Wi2 免费 Wi-Fi 运营

在日本的火车站、公交站、咖啡馆、购物场所等地可以搜索到一个免费的 Wi-Fi 热点——Wi2。它不仅给日本本国的国民，更是给到日本旅游又希望省钱的外国国民带来了极大的方便。

Wi2 由日本运营商 KDDI 控股，在大东京都市圈部署了 2 万个左右的热点。Wi2 的盈利模式与国内电信运营商的模式不一样，是电信的互联网前向免费、后向收费的模式。通过为客户免费提供 Wi-Fi 访问的方式，积累了大量的客户数据，包括位置、终端和访问内容，然后将这些信息打包为数据产品提供给商家，同时也基于这些信息部署展示类或效果类精确广告，实现盈利。值得注意的是，Wi2 在用户免费使用 Wi-Fi 的协议中，明确了对数据的所有权和使用权，这是免费使用 Wi-Fi 的一个交换条件。

Wi2 提供了两款大数据产品：Ideal-insight 和 Wifig。

Ideal-insight 是提供给 Wi2 的热点附近的商家的一项数据产品。它将从 2 万个 Wi2 热点中获取用户时间序列、使用终端和网络访问内容，其中，Wi2 客户端（Wi2 connet）收集个人信息和 GPS 信息，结合政府公开数据中的天气数据和商家自己上传的 POS 收银数据进行联合分析，告诉商家顾客的来源和国别、顾客曾经光顾和停留过的地点，以及附近人流的趋势和特征，帮助商家更好地进行市场营销（见图 2-15）。同时，通过 Wi2 的客户端，商家也可以根据顾客特征为顾客推送优惠券信息。

图 2-15 顾客流向分析及商圈分析

Wifig 是另一款数据产品。它主要针对外国来日本旅游的用户。由于 Wi2 分析了用户的访问记录，因此可以从网页的 URL 或编码信息中知道用户来自哪个国家。由此分析不同国家游客的旅游、住宿和娱乐偏好，帮助旅游公司、巴士公司和纪念品商家进行

更好的规划。

更加直接的是，根据分析到的用户特征，Wi2 可以在客户端中推送，或者在 Wi-Fi 登录的首页放置交通、旅游、娱乐等推荐信息，这种效果类广告定向投放非常精准，为 Wi2 公司带来的收益也颇为可观。

值得一提的是，Wi2 是整合了国内原有的三四家 Wi-Fi 提供商来提供无线上网服务，后端分析和推荐功能完全部署在 AWS 的云计算之上，采用 EMR 进行日志的加工和汇总，采用 Mahout 进行推荐和预测。Wi2 无需建设大数据基础设施，便可快速灵活地开展业务，其核心是大数据算法和大数据商业思维。

国内三大电信运营商为了弥补原有网络容量的不足，也建立了大量的 Wi-Fi 进行补充，但是其前向收费的商业模式阻碍了用户的使用，没有用户使用也就没有数据的沉淀，使得后向变现变得几乎不可能。因此三大电信运营商都在反思耗资巨大的 Wi-Fi 网络到底应该如何发展。Wi2 成功进行的免费 Wi-Fi 商业运营是一个很好的方向。部分省的电信、移动和联通纷纷开展了类似地探索。

2.3.5　电信大数据发展趋势

总的来看，电信行业的大数据依然处于探索阶段，未来几年，无论是内部大数据应用还是外部大数据商业化都有很大的成长空间。但电信行业大数据最大的障碍是数据孤岛效应严重，由于国内运营商的区域化运营，电信企业的数据分别存储在各地区分公司，甚至在分公司内不同业务的数据都有可能没打通。而互联网和大数据则是没有边界的。对于三大运营商来讲，各家对于大数据的发展思路也各不相同，但总体来说均在加速推进。

中国移动从“移动通信专家”到“移动信息专家”的策略转变，就是为顺应移动互联网时代潮流而做出的改变。这一战略的发展基础就是中国移动针对大数据和云计算研究所获得的应用发展方向。中国移动在大云 2.5 平台上部署了分析型 PaaS 产品，利用 BC-Hadoop 构建大数据处理平台以及海量结构化数据存储系统（BC-Hugetable），同时建设了并行数据挖掘系统（BC-PDM&ETL）以及商务智能平台（BI-PAAS）等大数据应用平台，为将来在大数据应用和服务市场的发展做了充分准备。

中国联通对大数据的探索源自于 2010 年中国联通数据大集中策略的提出。2009 年，中国联通 3G 业务正式商用，提出“统一品牌、统一业务、统一包装、统一资费、统一终端政策、统一服务标准”的“六个统一”策略。这意味着中国联通要走一条数据大集中的路线。2012 年底，中国联通就已经成功将大数据和 Hadoop 技术引入移动通信用户

上网记录集中查询与分析支撑系统。当前，中国联通已经新增 100 亿元投资重庆大数据计划，显现了其发展大数据和转型自身业务的决心。

中国电信很早就已经意识到移动互联网时代的到来，并于 2005 年提出了战略转型的构想，主要目的就是为了应对移动互联网时代的挑战。而当前，中国电信已经提出了“智慧城市”发展战略，其中很重要的技术结合点就是物联网和大数据。基于以上战略，中国电信定位成为智能管道的主导者、综合平台的提供者、内容应用的参与者。而在经营方面，中国电信从“话务经营”向“流量经营”转型。结合大数据技术，中国电信也将深入 IDC 服务以及智慧城市建设，并发掘移动互联网与之结合的商机，重塑转型之路。

总之，运营商利用大数据来推动业务转型将是未来电信市场的一个重要方向。电信运营商如果能够通过技术的进步，不断释放其管道中庞大数据的潜在力量，那么必将成为未来移动互联网时代中最大的赢家。所以，电信业也势必将大部分投资转向大数据应用市场。根据赛迪顾问分析预测，中国电信业大数据应用市场将保持快速增长势头，增长水平高于大数据整体市场增速，到 2015 年，电信业大数据应用市场规模预计将达到 25.3 亿元。

2.4 电力大数据

2.4.1 电力大数据应用现状

智能电网的理念是通过获取更多关于我们如何用电、怎样用电的信息，来优化电的生产、分配以及消耗。在本质上，智能电网是大数据在电力上的应用。同样，智慧城市的互联设备也会依靠大数据来确保其工作的有效性[㊀]。

“十二五”期间，国家电网公司初步建成具有信息化、自动化、互动化特征的坚强智能电网。把智能电网中产生的所有数据组织和收集起来，每年产生的数据将以几何级数速度增长，大量的半结构化和非结构化信息有待进一步的管理和存储，需要有专业化的解决方案应对大数据挑战[㊁]。

智能电网的实时运营要求快速处理海量数据、实时采集数据、在线分析决策，传统的数据仓库平台无法支持这些新形势下的需求。电力大数据是大数据理念、技术和方法在电力行业的实践，是大数据应用的重点领域之一，涵盖了从数据采集、存储、处理、分析到最后提供预测、评估等数据价值挖掘服务的全过程。电力大数据将贯穿未来电力

㊀ 引自：《大数据时代》，维克托·迈尔 – 舍恩伯格。

㊁ 引自：《浅析电力企业如何应对大数据》，姚玮，江樱。

工业生产及管理等的各个环节，起到独特而巨大的作用。

目前，在智能电网大数据应用基础方面，已经完成了电力调度、电网运行状态监测、电力用户用电信息采集、电力营销、客户服务、用能服务、需求侧管理等系统和平台的建设，积累了大量的数据资源，业务数据从总量和种类上都已颇具规模，具备了良好的数据基础，并初步实现了企业级数据资源整合及共享利用。

在电力用户用电信息采集方面，建成智能电网用户用电信息采集系统，实现对电力系统范围内电力用户的“全覆盖、全采集、全费控”。目前仅在国家电网管理范围内累计安装智能电能表约 3 亿只，预计 2016 年智能电表的数量将成倍增长。系统数据累计达到 PB 级别，每年以百 T 的规模递增。随着采集覆盖率的增加，以及智能电表、采集终端等设备的海量增长，计量设备的应用呈现出数量大、类型多、智能化、运行环境复杂的特点，对数据采集和分析挖掘能力的要求也越来越高。对用电信息数据这样复杂的大规模数据集，用传统的数据处理方式存在一定困难，因此亟需应用大数据技术。

在电力客户服务方面，国家电网采取了范围内客户服务工作集中运营模式，建成了 95598 业务支持系统、95598 呼叫平台、95598 智能互动平台等基础支撑平台，统一管理投诉、建议、表扬、举报、意见、咨询、故障报修等几大类业务，每日产生万级客户电话语音数据、十万级工单记录数据以及与其他系统同步的业务数据，通过对客户服务数据的进一步深入挖掘，从电网管理提升和安全稳定运行等方面提供数据支撑。

在便捷居民缴费服务方面，智能电网一体化缴费管理平台整合现有的电网缴费系统，全面支撑电力公司、金融机构、非金融机构等多种渠道缴费，实现各类缴费渠道的统一规划、统一接入、统一管理。

智能电网电能服务管理平台向全社会提供规范、高效、专业的节能服务，促进国家节能减排目标的顺利实现，为推动国家经济发展方式转变做出了重要贡献。

目前，对于智能电网大数据应用已经开展了较多研究，并取得了一定的应用成果，主要集中在用户用电行为分析、线损多维度分析、计量装置在线监测与智能诊断、负荷特性与有序用电分析、经济趋势分析等众多方面。在用户用电行为分析方面，根据客户用电量、用电特性、业务办理、缴费信息、客户投诉记录等用电信息，研究用户的负荷特性及用电行为习惯，针对目标用户通过不同的服务渠道开展精准服务；在多维线损分析方面，通过配置线路及台区损耗模型，系统自动计算每日损耗，实现 10kV 输配电线路和台区线损的在线监测，及时发现线损异常，掌握现场供电情况，实现线损可视化图形展现分析；在计量装置在线监测与智能诊断方面，通过用电信息采集系统、营销系统

的数据共享和交互，利用大数据分析技术进行计量与采集装置故障诊断、设备整体工况分析，实现计量装置异常分析、采集设备故障分析、各类事件分析和用电异常分析等功能，为考核采集设备和计量设备厂家提供评价统计数据；在负荷特性与有序用电分析方面，分析了用户负荷曲线特征中的峰谷出现位置、峰谷持续时长、尖峰负荷出现时点、波峰出现的次数等特性，并分析了专变和专线的移峰填谷潜力；在经济趋势分析方面，根据不同地区、行业的历史电量走势，结合电量与负荷预测，分析不同产业的用电量在各地区的增长、占比及变化情况，了解产业结构、产业链、产业转行升级、产业区域迁移以及地方特色产业的发展情况，依据行业度电产值，分析电力弹性系数据、电力密集度、电力 GDP，结合用电走势、工厂企业产能利用率、业扩报装，分析电力生产力，参考 PMI 指数发布各行业景气指数。

2.4.2 利用电力负荷值实现智能电力现代化管理

近年来，国内电网负荷峰谷差不断增大，为满足尖峰负荷需要，发电侧、电网侧都投入了大量资金，造成社会资源浪费。同时，随着智能电表的推广及用电信息采集系统建设的全面提速，积累了海量的用户负荷曲线数据，为负荷特性分析奠定了扎实基础。电力负荷预测是电力系统规划的重要组成部分，是电力调度、用电、规划等部门的重要工作之一，也是电力系统经济运行的基础，对电力系统的规划和运行都极其重要，其实质是对电力市场需求的预测，提高负荷预测技术水平，有利于电力调度管理，有利于合理安排电网规划建设，有利于提高电力系统的经济效益和社会效益。如何通过大数据分析技术提升预测的及时性和准确性成为迫切需要解决的问题。

1. 数据现状与应用需求分析

用电信息采集系统经过多年的运行，已积累了各行业功率、电量、电压、电流类海量数据。例如，用电信息采集系统每天采集各用户 96 点的电量数据和 96 点的负荷数据，数据表中存有近 5 年的数据量，数据量达到数十亿级[㊀]。这些数据根据不同用户类型，结合业务场景和数据采集范围进行分类。

但是，目前各类用电海量数据的价值尚未得到深入的挖掘和利用，在负荷预测方面尚未发挥基础数据作用，需要依托大数据技术提升数据价值。另一方面，现阶段电力负荷预测主要存在以下提升点：预测算法较多基于经验法，预测的准确率有待提升，业扩、天气、气温变化等因素考虑较少，预测模型未实现系统化支撑，费时费力等等。

㊀ 引自：电力企业《基于用电信息采集的大数据应用研究》。

为了更加准确地计算预测用电负荷，需要利用多个业务系统的海量数据进行联合分析和数据挖掘。传统技术存在计算速度慢、计算周期长、扩展性差等缺点，且对数据的利用仍然停留在以统计报表为主的浅层应用层面，难以挖掘出新的业务模型和数据蕴含的深层次业务价值，迫切需要采用新的技术进行处理。现今以分布式存储、分布式计算、海量数据挖掘为代表的大数据技术得到了空前的发展并日趋成熟，为电力行业海量数据的深度应用提供了有效的技术手段。

在应用大数据技术进行电力负荷预测分析时，主要使用的是用电信息采集系统、生产管理系统、营销业务应用系统中的数据，这些数据可分为两类——营销数据和外部数据。营销数据包括用电负荷、用户档案、电网设备台账、电网网架结构，外部数据包括国民经济增长速度、产业结构调整、消费水平、工业与居民电气化程度、电价政策、气候 / 气温变化等。

2. 应用场景分析方法

电力负荷预测主要用于预测下一阶段的用电负荷情况，以便提前了解用户的未来用电需求量，为正常供电做好准备。可以根据负荷使用情况及行业分类用户群，针对不同用户群进行电力负荷预测，即根据历史用电负荷数据对用户群进行聚类，识别出用电负荷行为模式一致或相近的用户群组，根据其用电负荷的变动特征，为下一阶段的用电进行准备，提供供电保障。

负荷预测工作的关键在于收集大量的历史数据，并结合大数据处理技术——如 Hadoop、流计算以及内存计算等，采用有效的并行算法进行大量试验性研究，总结经验，不断修正模型和算法，以真正反映负荷变化规律，其基本过程包括海量数据导入及整理、海量负荷数据的预处理、建立负荷预测模型。

值得一提的是，外部数据可能包括半结构化和非结构化数据，需要应用大数据技术进行挖掘与分析，例如使用关联规则挖掘方法，提炼出影响电力负荷的各类因素，并对其进行影响力度、发生频率、频发时间等方面的统计与分析。最终目标是建立高效的负荷预测模型并促进其优化，提高预测结果的准确性。

电力负荷预测分析根据预测周期的长短可划分为短期电力负荷预测与中长期电力负荷预测。

进行电力负荷短期预测一般指提前几小时到几周的负荷预测。它对于机组协调、经济调度、水火电协调、负荷管理等都有重要作用。短期预测涉及数据量大、取样复杂，

导致样本数据之间的差异比较大。针对此类问题，可基于大数据分析模型，采用基于加权相似度和加权支持向量机的短期电力负荷预测算法。

中长期负荷预测一般指提前几周到几个月甚至几年的预测。它对于燃料供应计划、机组维修计划、能量交易、发电厂税额估计等是必需的，可帮助电网规划、增容和改建，并作为电力规划的重要依据，合理安排电源和电网的建设进度，提供宏观决策依据，使电力建设满足国民经济增长和人民生活水平提高的需要。与短期预测相比，长期预测根据历史用电情况，结合国家宏观政策影响等因素进行预测，相关因子相对较少，主要根据历史用电的趋势进行预测分析。

另外，还可以按省、市、区县等地域级别进行电力负荷预测分析以及进行重要客户电力负荷预测等。

电力负荷预测主要用来预测未来电力需求量、未来用电量、用电负荷曲线、负荷时间和空间分布等，故可以通过多元线性回归算法、神经网络算法、SVN 算法建立分析预测模型的方式进行预测分析。预测模型通过调整影响因子变量的方法分析各影响因子对电力负荷的影响，从而指导电网规划优化设计。

3. 应用场景成效

基于用电信息、电费信息及用户负荷等数据，利用分布式计算、数据挖掘分析等技术进行的电力负荷预测，是电力市场的重要组成部分，对电力系统控制、运行和计划都很重要。

短期负荷预测成效包括通过挖掘不同地区、不同时段和行业的用电特色，全面透析电网负荷情况；综合历史负荷情况、用户习惯、天气情况等，实现对未来一小时、三小时、一天乃至一周的负荷预测；预设阈值，当预测值到达预警范围，立刻通知相关人员，从而辅助优化供电调配，降低电网过载风险。

中长期负荷预测成效包括发掘各地区电网负荷随季节、政策的变化规律，并结合政府规划、国民经济增长速度、产业结构调整、消费水平、工业与居民电气化程度、电价政策、气候 / 气温变化等外部因素对未来一年到三年的电网负荷变化情况做出预测，为公司电网规划、设备检修、电能调配等提供决策支持。

2.4.3 利用用电信息数据指导用户合理优化用电

基于用电信息、电费信息及用户负荷等数据，通过大数据思维进行用户及行业能耗分析，利用分布式计算框架 MapReduce、Spark 框架、数据挖掘分析等技术，分析出各

地区、各行业的电能使用情况以及用户的用电行为习惯，构建用户用电影响模型，将相关分析结果向用户进行推送，对用户不合理的用电用能习惯提出建议，从而引导用户合理优化自身用电行为，提高能源利用率，降低用能费用，保障用户经济利益。同时，也有利于电网削峰填谷、平稳运行，提升电力企业供电服务质量。

1. 数据现状与需求分析

主要使用的数据涉及用电信息采集系统中所有电量数据，以及营销业务应用系统中的数据，包括用电量、用电负荷、客户信息等数据。

用户用电数据之间存在关联性和相似性，隐藏着用户的用电行为习惯，对这些用电数据进行基于大数据技术的分布式计算与数据挖掘分析，可以研究用户的生活习惯、用电习惯、电费政策和企业用户的用能、用能与合同的关系等。同时，由于用电数据受气象、节假日、业扩报装等外部因素影响，因此可以构建用户用电选择、环境、业扩报装等用电量影响因素模型，根据居民用户的用电行为习惯（峰谷电量使用情况、阶梯电量使用情况、缴费习惯等）和企业用户的用电耗能行为（用电量情况、用电负荷情况、电费及缴纳情况、合同容量与超容情况、能耗水平、产能利用水平、峰谷用能合理性），对不合理的用电和耗能行为进行合理性建议并推送。

2. 应用场景分析方法

（1）建立用电影响模型

行业结构的变化可由用电量变化反映出来，用电结构变化的背后，就是行业结构的优化。通过数据分析商业用电、工业用电、居民用电的用电量趋势对全社会用电量的影响程度，对温度、湿度、节假日、居民意愿、业扩报装等电量影响因素进行分析，根据单一用电影响因素以及多因素的重叠影响，通过非监督性学习聚类分析方法进行建模，并利用数据平滑算法进行模型修正，生成气象电量影响模型、节假日电量影响模型、业扩报装电量影响模型、行业结构电量影响模型等，达到研究用户负荷特性和用电行为习惯的目标。目前，省级各模型反映的电量影响关系达到上亿条。

（2）用户用电行为分析

主要分析目标用户的用电电量、电费、负荷特性、用电模型、产能利用率等数据。利用 MapReduce 并行计算框架对这些数据进行处理，然后结合模型和算法进行分析。全面展示用户用电行为是生成用户用电合理建议书的前提。

依据企业所属行业、负荷特性、电量等信息分析企业的用电行为，并进行分类。分

析企业的产能、产值及产能利用效率，通过企业类型、气象条件、生产计划等因素分析企业变压器的最大负荷变化情况，并对未来中长期的负荷增长进行预测，提前对变压器超载情况进行预警。另外，开展对企业用户供用电形势、用电情况的调查分析，可以进行企业用户营销服务策略研究，实现电力企业营销服务的创新，制定有效措施并积极行动，以高质量、全方位的服务满足企业用户的用电需要。

对于居民用户需要分析电量走势、峰谷电量、阶梯电量、季节电量、电表示数等数据。用户数据具有海量、高频、分散等特点，对这些数据进行挖掘并研究用户类型，分析各地区城网、农网的居民用户用电基数。分析居民的电器化率及电器使用情况等，可以帮助电力企业了解用户的个性化、差异化服务需求，从而进一步拓展服务的深度和广度，为未来电力需求侧响应政策的制定提供数据支撑。

（3）节假日对用电量的影响分析

分析节假日对用户用电量的影响，建立用户节假日电量模型。该模型主要根据用户用电量走势和节假日用电量的关系计算，分析用户用电量受节假日的影响率，通过展示某年所有节假日的电量影响模型，反映节假日对企业生产、居民生活的影响。

（4）用户能耗分析

针对企业用电用户而言，可以对各地区各行业的整体能耗情况、小时能耗情况以及峰谷能耗情况进行分析，并分析企业用户的能耗水平、产能利用水平、峰谷用能合理性以及企业的电量、功率因素等变化，显示出用户可能存在的异常耗能情况。

依据各地区居民用电基数及峰谷电使用情况，分析居民用电能耗水平和峰谷用能合理性，同时，通过分析居民的电量变化提醒用户可能存在的异常耗能情况。

（5）生成并推送合理用电建议书

对于企业用户，分析其用电行为并与企业用户所属行业用户的普遍行为进行比较，生成针对该企业用户的合理用电建议书。提出电量分析建议项、负荷分析建议项、产能利用率分析建议项、温度对负荷影响分析建议项等内容，提醒用户合理选择适合本行业的电价政策、用电习惯。

对于居民用户，通过居民各月份分时明细用电视图、居民用电能耗与用户的用电习惯对居民用户提出合理用电调整建议，同时也使得电力收费过程更透明。根据用户用电行为分析以及和同地区居民用电情况的比较分析生成针对该用户的合理用电建议书，包含峰谷电量、接替电量和季节电量等分析建议项。

依据用户的用电行为、负荷特性类别以及电价情况，制定目标用户的节能方案，并

将用户合理用电建议书的分析结果与节能方案通过短信、电话、微信和门户网站等渠道向用户进行信息推送。

3. 应用场景成效

基于用电信息、用户负荷等数据，利用分布式计算、数据挖掘分析等技术，研究用户的负荷特性及用电行为习惯，构建用户用电意愿、业扩报装、环境因素、行业用电结构变化与电量变化影响关系模型。对用户用电能耗进行分析，对不合理用电行为进行合理性用电建议与信息推送，可以引导用户调整电用行为习惯，提醒用户有可能存在的设备缺陷，还可以指导用户合理签订购售电合同、选择电价政策、合理安排生产活动、错峰用电，降低用户的生产、用能成本，保障用户经济利益，达到节能减耗和平稳优化电网运行的目的。

2.4.4　利用消费能耗数据进行节能减排

Pecan Street Inc. 是一个由大学、技术公司和公用事业提供商组成的非盈利性组织，他们协作在智能电网技术领域进行测试、试运行和商业化运营工作。对他们来说，“智能电网”一词定义了公用事业行业进行电网系统改造所做的工作，其中包括采用高级计量系统、家庭能源管理系统（HEMS）、传感器、太阳能光伏（PV）系统、智能家电、电动汽车以及互联网服务等，以帮助消费者提高能源效率。当前，Pecan Street 已经通过德克萨斯州奥斯丁市 Mueller 社区 200 多个家庭中的传感器系统，收集了近两年的能耗数据。

Pecan Street 的主要目的是推动在消费者能源管理领域发现新的产品、服务和经济机会。他们的研究将可以向人们提供管理和减少其能耗的知识和工具，使其家庭生活更舒适。此外，公用事业公司将能够利用此类数据更好地管理电网，并投资进行基础设施优化工作。

Pecan Street 研究的核心是一种终端设备到云的架构，能够捕获多个来源的数据，并进行存储以供分析和可视化之用。他们通过一些系统收集电力数据，这些系统能够每隔六秒测量一次流过 6 ~ 8 个电路的电量，还通过使用无线网关的公用事业量表收集燃气和水的数据。此外，社区中旧机场塔台上的无线电收集器可通过高可靠性公用事业网络，将 Landis+Gyr 高级智能量表中的数据传送到德克萨斯大学的超级计算机设施及其数据库中。

他们记录消费者的行为，如修改其家庭中的环境控制方式，或者调整其能源信息查看方式等。还计划收集来自高级恒温器、家庭自动化系统、家庭安保系统、运动探测器以及新能源技术（如太阳能板和电动汽车充电站）的数据。在此项目中，了解消费者如何对待他们收集的数据非常关键。因此，在研究中，有些参与者可以通过基于 Web 的门户或智能手机应用访问其数据。

在两年的时间里，他们捕捉了大量数据（总共有超过 80GB 的信息），预计在计划实施期间数据量将增加到 1TB。虽然 1TB 可能看起来不是很多，但作为单个数据点，这实际上数量巨大。每个数据点都代表一个唯一事件，因此对此类数据的处理极其复杂。大数据带来的挑战是整合这一来自多个分散来源的稳定非结构化数据流，并将其传输至德克萨斯大学，以进行分析和可视化处理工作。

他们正在第三代数据库架构上尝试找出存储和分析如此大量数据的方式。在第一个月过后，他们认识到 MySQL 方法无法应付复杂的查询，例如电冰箱在 24 小时内的总体使用情况如何，或是否有机会实施公用事业需求响应或调峰方案。因此，他们迁移到了其他数据库架构，现在正在评估 EMC Greenplum 大数据解决方案。Greenplum 提供了一种集成式大数据分析系统，该系统采用了大量并行处理（MPP）架构，没有专有硬件的复杂性和约束，所用的 Hadoop 产品可帮助他们利用针对结构化和非结构化数据的模块化解决方案处理和分析数据。

除了寻求合适的大数据分析方法，他们面临的最大挑战之一是收集数据的完整性。数据系统中的无效信道或者居民宽带连接中断会提供不可靠的值。他们通过生成确知完好数据的合格数据集来解决此问题，将这些数据标记为极高质量，并指引研究人员使用这些数据。

包括英特尔在内的参与 Pecan Street 组织的机构将使用智能电网项目作为其新产品概念的开发平台，在此基础之上进行开拓创新。利用大数据分析可以更好地了解人们的能源消费方式及其希望的能源管理方式。此外，他们可以向公用事业公司提供决策支持，帮助他们在电网改造领域进行最佳投资。

2.4.5 电力大数据发展趋势

随着我国智能电网建设速度不断加快，电力大数据的作用和优势也日益突出。面对海量的电力数据，综合电网营销数据、配网数据、调度数据等多环节业务的综合数据挖掘，其成果将对电力系统故障诊断、电力调度、负荷预测、安全性评估等几个方面做出

非常大的贡献。通过良好的大数据管理，可切实提高电力生产、营销及电网运维等方面的管理水平。

未来在实时采集用电侧用户电能信息的情况下，可针对特定区域、特定用电群体甚至特定电器开展实时用电数据、阶段用电数据、累计用电数据的价值挖掘，从而预测用户用电需求，预制合理的供电计划，支撑需求侧管理。通过大数据分析技术分析用户用电行为，可重塑用户用电习惯，推动节电减排。根据各区域用电数据分析，可整合配供电资源，确定最优用户业扩方案，支撑用电方案优化制定。分析用电用户对于电价的敏感度，可为分时电价业务开展做准备。同时，电力大数据相关技术的深化应用，还将对各级政府、事业单位及其他商业团体带来越来越大的价值，例如：对特定行业用户用电数据进行分析，可衡量该行业发展情况，支撑行业发展分析；对特定区域内用户用电数据进行分析，可衡量该地区经济情况，支撑地区经济发展分析；对不同用户的用电习惯进行分析，可衡量用户经济能力，从而支撑家电厂商、房地产商等进行业务推广或调整业务发展方向。

大数据时代为智能电网带来了新的发展机遇，开创了智能电网的新纪元，同时也提出了新的挑战。在电力行业内开展的电力大数据研究中，大多都是将现有 Hadoop、Storm、内存库等大数据技术单独应用，而对技术的融合创新研究尚未取得特别多的成果，这方面的研究仍然是一大难题。从长远来看，作为中国社会发展的晴雨表，电力数据因其与经济发展紧密而广泛的联系，将会对我国经济社会发展乃至人类社会进步形成更为强大的推动力。

2.5　交通大数据

2.5.1　交通大数据应用现状

近年来，大数据、云计算、物联网等技术给交通领域的技术发展、商业模式和交通理念都带来了重大变革。目前，大数据技术在交通运行管理优化、面向车辆和出行者的智能化服务，以及交通应急和安全保障等方面都有重大发展，未来基于大数据技术的智慧交通建设将会成为我国交通的发展方向。

我国交通信息系统普遍存在跨系统、跨地点、跨专业的问题，具体表现为系统众多、整合困难、信息离散、共享薄弱、信息孤岛、同步不畅、信息模式复杂、缺乏统一数据标准、受众局限、利用率偏低、交通服务产业链不健全等。与其他行业相比，交通

行业有着其独特的特点，如交通大数据的实时性、并发性、交通系统基础网络带宽存在限制等，导致传统大数据算法并不适用于交通行业。经过多年的发展，我国交通信息系统已经从控制阶段发展到集成的需求阶段，从单项应用阶段发展到在客观数据支持下综合化、智能化的应用阶段。

随着交通系统步入数字化时代，交通行业虽然面临着巨大的挑战，但也迎来了前所未有的机遇。在城市轨道交通信息化建设方面，前瞻产业研究院发布的《2013 ~ 2017年中国铁路信息化行业市场前瞻与投资战略规划分析报告》显示，2008 年我国城市轨道交通信息化系统市场规模仅为 31 亿元，增长率为 12.73% ；到 2010 年，我国城市轨道交通信息化系统市场规模已达 45 亿元；到 2012 年，市场规模已达 65 亿元，增长率为 23.19% ；2013 年我国轨道交通信息化系统市场规模已达 81 亿元，预计 2014 年将有可能达百亿规模。在铁路信息化建设方面，中国铁路信息化的投资规模近几年保持快速增长，特别是最近两年的增长率更达到 30% 以上。就投资规模来讲，据统计，2010 年全国铁路信息化投资 150 亿元，2011 年将近 200 亿元，预计 2015 年将超过 300 亿元，5 年复合增长率接近 15%。“十二五”期间铁路行业信息化投资总额将达到 1 400 亿元，预计“十二五”比“十一五”期间信息化投资增长超过 200%。如果考虑对于既有线路的改造，市场空间则更为庞大。在中国城市道路智能交通系统方面，预计 2009 ~ 2016 年总投资额将达到 1 077.58 亿元，其中视频监控系统投资额超过 400 亿元，年均投资额超 50 亿元。

面对中国交通系统的发展趋势，传统的劳动力密集型管理方式和运维方式已经不利于交通产业的发展。为了更好地实现交通运输科学发展，交通运输部提出，要全面深化改革，加快发展“四个交通”，并强调，综合交通是核心，智慧交通是关键，绿色交通是引领，平安交通是基础。“四个交通”相互关联、相辅相成，共同构成了推进交通运输现代化发展的有机体系，为下一阶段交通运输行业全面深化改革和转型发展指明了方向。“四个交通”中，智慧交通是关键。加快发展智慧交通，是推进交通运输管理创新的重要抓手，是提升交通运输服务水平的有效途径，也是推动交通运输转型发展的重要支撑。建设智慧交通，是城市发展的新范式和新战略，而大数据是交通领域都实现“智慧化”的关键性支撑技术。根据赛迪顾问预测，未来 3 年智慧交通大数据应用规模年均复合增长将达到 119.7%，到 2015 年，市场规模将达到 17.61 亿元（见图 2-16），远远高于其他领域，可见交通领域大数据应用有着广阔的市场前景。

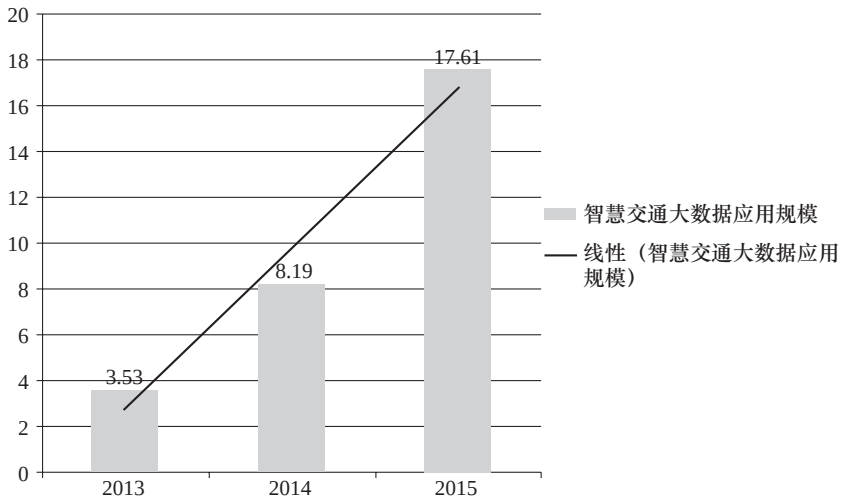


图 2-16　2013 ~ 2015 年中国智慧交通大数据应用规模预测

2.5.2　轨道交通大数据技术创新

我国拥有全球最大的人口基数，幅员辽阔，需求极其复杂，只有我国才会产生世界上最大量的客运线路和最多数量的客户结合在一起独特的需求。目前，传统铁路通过新线路的建设来提高铁路运力和实现安全保障，这样会造成资源的闲置和浪费，与我国轨道交通的长期发展目标不符。因此，以技术为创新动力，提升对既有资源的利用率，提高铁路运输能力、效率、安全以及顾客体验，发展数字铁路将成为未来发展的主流。

2014 年，工业和信息化部软件与集成电路促进中心授权北京泰乐德信息技术公司为国家软件与集成电路公共服务平台——数字铁路与城市轨道交通一体化创新服务平台的承建单位，联合轨道交通业内外相关厂商、用户、高校及研究机构，合力建设我国轨道交通领域基于大数据技术的一体化创新服务平台。该平台是国家产业公共服务体系的重要组成部分，也是数字铁路建设的重要推动力量。因此，建设该平台对我国数字铁路的创新和发展具有重要的社会意义和经济意义。

1. 数字铁路与城市轨道交通一体化创新服务平台的定位

数字铁路与城市轨道交通一体化创新服务平台与轨道交通的运输管理系统是相辅相成的互补关系，与轨道交通既有的监测系统、业务系统是互补互充的对接关系，是定位于轨道交通既有设备和系统的上层顺延平台和增值应用平台（见图 2-17）。

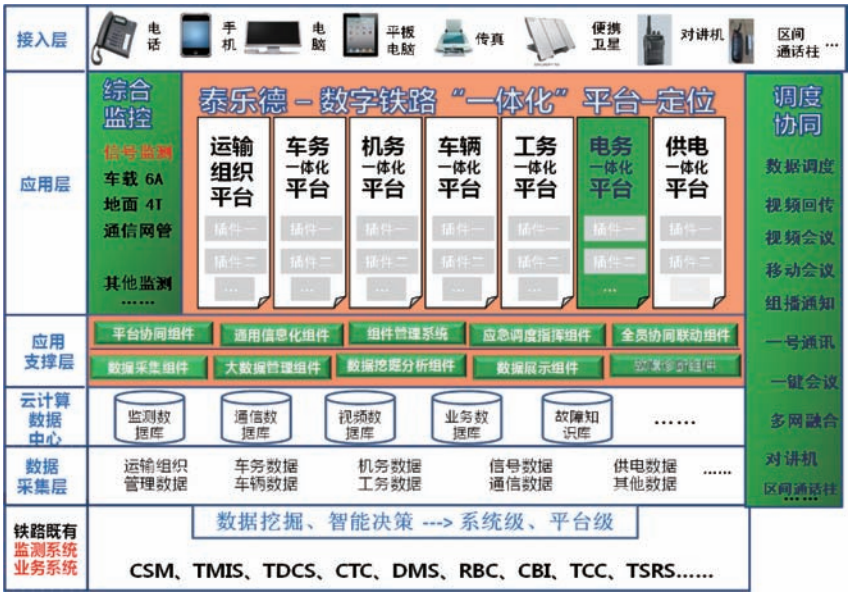


图 2-17　一体化运维平台的定位

2. 数字铁路与城市轨道交通一体化创新服务平台的组成

基于通用组件与定制插件技术路线的数字铁路与城市轨道交通一体化创新服务平台的组成见图 2-18，它可以统筹解决轨道交通运维需求的实际问题：

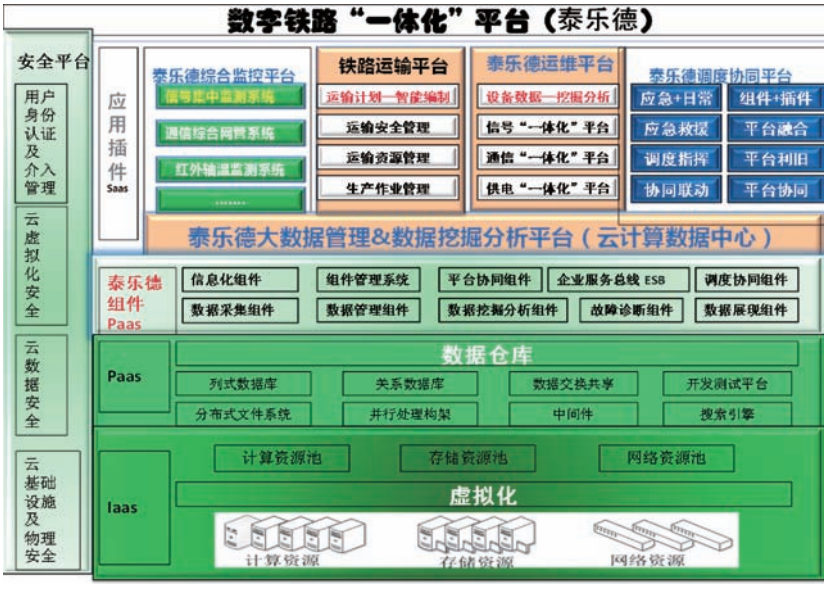


图 2-18　一体化运维平台的组成

综合监控。从源头抓隐患，运维关口前移——告警。对综合监控平台将既有的分散孤立的各监测子系统实现集中管理、信息同步、各取所需和数据共享等，支撑“状态修”运维。

数据挖掘。从源头抓隐患，运维关口前移——预警。大数据管理与挖掘分析平台针对既有设备、系统的各种数据资源，实现按需采集和按需展示，对海量数据进行大数据管理和挖掘分析。利用北京泰乐德信息技术有限公司科技创新的符合轨道交通信号监测数据特点的数据挖掘组件，如决策树、时间序列、SVM、贝叶斯、粗糙集、关联规则、神经网络等对海量监测数据进行挖掘分析，并与设备厂商、现场用户合力开展故障诊断等联合科研，自动生成维护建议，支撑“状态修”运维，从而在积累的过程中逐步减少长期困扰轨道交通行业设备运维的“过剩修”和“失修”难题，逐步推进“计划修”向“状态修”转变。

状态修。从过程抓隐患，运维链条延伸——状态修。综合监控平台采集的各种既有监测系统的告警和大数据管理与挖掘分析平台产生的预警，都将按照预先设定的条件和规范，自动推送给相关的运维人员和管理人员，启动闭环运维，并留存可追溯的依据。

计划修。从过程抓隐患，运维链条加固——计划修。全面实现轨道交通领域当前的计划修运维模式的信息化、闭环化、规范化。

综合督办。从宏观和微观两个方面抓隐患，全面实现轨道交通运维管理和运维作业的信息化、闭环化、规范化。

作业防护。综合利用物联网技术、移动互联网技术、既有监测系统的数据，保障作业人员的人身安全，减少路内伤亡事故，实现设备的全生命周期管理，提升设备管理能力。

网络办公。结合轨道交通运维和管理工作的特点，实现各部门各类日常工作的信息化。

调度协同。综合利用移动互联网、物联网等技术，结合轨道交通运维的移动作业、移动沟通、运维现场与中心需要远程互动等需求特点，量身定制三大通信子系统：移动指挥调度系统、移动视频会议系统、移动运维协同系统；量身定制三大信息化系统：会议组织管理系统、应急预案系统、应急资源系统，这些系统共同组成调度协同平台。

其中，移动指挥调度系统将铁路既有的应急通信系统的通信功能全面实现移动化；移动视频会议系统与铁路既有视频会议系统互补，实现移动的音频和视频会议；移动运维协同系统利用移动智能终端实现运维现场情况的实时转播和汇报、现场专题报告的实时制作与同步分享等。

3. 大数据技术助力实现轨道交通运维“智能化”

大数据技术能够将隐藏于海量数据中的信息和知识挖掘出来，为社会经济活动提供依据，从而提高各个领域的运行效率，大大提高整合社会经济的集约化程度。基于大数据的状态监测、数据挖掘和故障诊断是国家软件与集成电路公共服务平台——数字铁路与城市轨道交通一体化创新服务平台的关键技术（见图 2-19 和图 2-20）。泰乐德公司在国内和国外数据挖掘与故障诊断的科研成果的基础上，总结了公司为其他行业研发数据挖掘故障诊断系统的实践经验，将其与轨道交通运维需求相结合，创新研发了轨道交通信号等数据分析处理软件系统，并对数据中心架构设计、数据仿真、数据采集、数据展示、数据管理以及适合轨道交通数据特点的数据进行挖掘分析与故障诊断，在关联规则、决策树、贝叶斯、神经网络相关技术领域都取得了阶段性的成果，并将该成果运用到实践中，希望未来在数据源积累的过程中，不断创新和发展轨道交通技术。

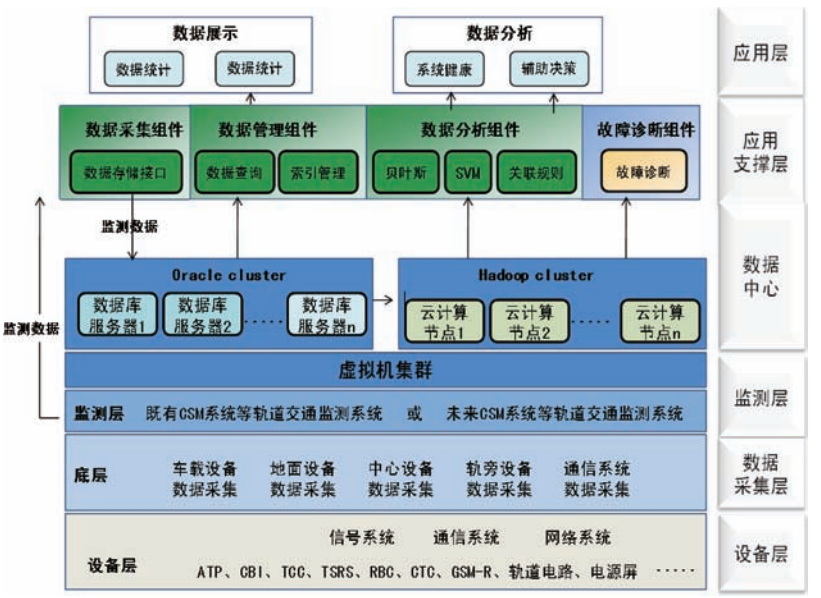


图 2-19　泰乐德 – 轨道交通监测 – 大数据管理与分析平台架构图

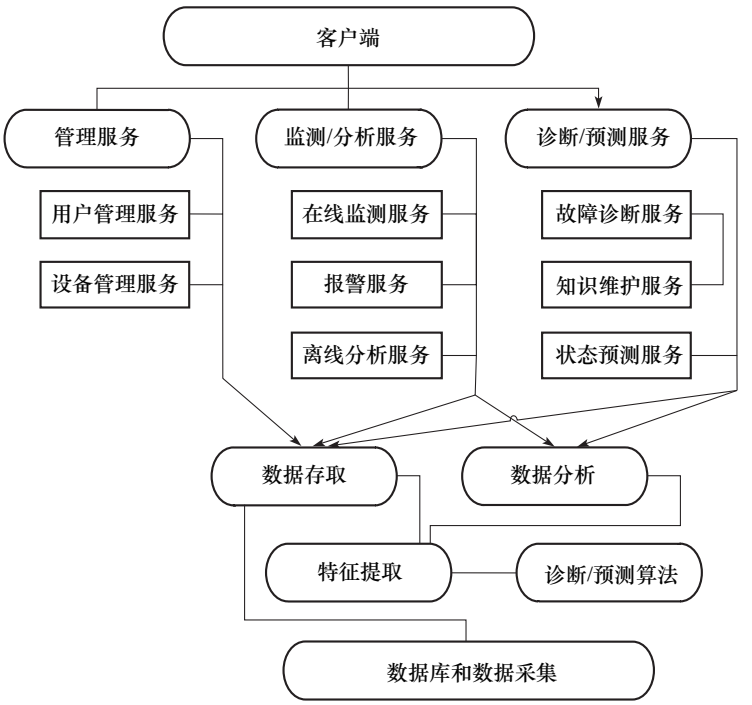


图 2-20　状态监测与数据挖掘故障诊断业务流程图

2.5.3　轨道交通大数据应用

大准铁路是国家“八五”计划重点建设项目“准格尔项目一期工程”三大主体工程之一，大准铁路东起山西省大同市，西至内蒙古鄂尔多斯市准格尔旗薛家湾，正线全长 264 公里，途径两省六旗县（市），是已形成的“西煤东运”大通道——大秦线的向西延伸，属一级单线电气化铁路。随着点岱沟 – 九苏木双线自动闭塞开通及巴准 – 大准 – 准池 – 朔黄的贯通，大准线运量将达到 2 亿吨，货流密度大，对通信信号设备的运维管理工作提出了更高的要求。按照大准铁路近期 2 亿吨 / 年运量计算，故障每延时 1 小时，少运输货物 2 万吨，收入将会损失 90 万元。据统计，2013 年大准铁路信号故障延时 863 分钟，2014 年（截至 9 月份）故障延时为 470 分钟，其中 70% 为道岔和轨道电路故障，因此，大准铁路希望借助泰乐德的数字铁路与轨道交通一体化创新服务平台，利用大数据技术降低故障延时。

信号微机监测系统是保障大准铁路运维的有力手段，但是在实际运用过程中，监测系统误报警多、智能分析薄弱以及监测信息受众范围的局限性导致了信号微机监测系统利用率不高。此外，铁路通信网络系统是铁路运输和运维的支撑系统，与铁路各部门、

各系统都是牵一发而动全身的重要关系，由于信号设备是列车运行的控制系统之一，直接影响行车，因此解决铁路通信系统与信号系统之间的跨系统问题是大准铁路运维现场的主要需求。从避免中断行车的角度讲，当前通信信号设备运维模式已不适应新的运输形势的要求，大准铁路急需综合化、智能化、一体化的铁路电务运维平台，有针对性地开展维修和维护工作。

为解决大准铁路信号段、通信段面临的实际问题，泰乐德公司技术人员深入大准铁路现场进行实地调研，总结信号系统的主要问题如下：

道岔。包括电动转辙机定（反）位表示电路故障、电动转辙机启动电路故障和道岔机械故障，典型故障如滑床板未清理干净导致的道岔动作电流曲线值升高、不平滑，以及尖轨夹入异物或道岔缺口未调好导致的道岔动作电流值升高、道岔不锁闭。

信号机。信号机点灯电路故障，典型故障如信号机灯丝电流超高、发光二极管故障。

轨道电路。包括轨道电路故障和轨道电路钢轨绝缘破损，典型故障如引接线跳线、箱合接续线松动导致的轨道电压下降或波动。

自动 / 半自动闭塞电路。包括自动 / 半自动外线断、自动 / 半自动外线混线和自动 / 半自动外线反接，典型故障如接车站收不到闭塞外线电压或发车站收不到闭塞回执信号。

技术人员利用轨道交通监测大数据管理与分析平台（泰乐德轨道交通一体化运维平台的组成部分之一）的数据采集系统，采集到了大准铁路信号微机监测系统的监测数据和既有的通信数据，并借助于分布式文件系统对数据进行存储和管理。在数据存储的基础上，借助于泰乐德“一体化”平台中的数据挖掘分析系统，实现数据预处理、数据预警判断、故障判断和辅助决策建议等，在数据逐步积累的基础上，逐步解决大准铁路的实际问题。

泰乐德轨道交通监测大数据管理与分析平台综合运用大数据、云计算、铁路设备、运输等多种技术，充分利用铁路既有信息系统的各种数据，解决轨道交通监测数据存储问题，为铁路的运输组织和铁路的设备运维提供客观的、科学的信息支持。提高铁路设备的运维能力，实现保障安全的数字铁路建设目标；提高铁路运输的组织能力，实现提高运力的数字铁路建设目标；伴随运输组织能力和设备运维能力的同步提高，逐步实现挖潜增效的数字铁路的建设目标。

对轨道交通领域而言，大数据是一种全新的技术，我国在铁路信号监测数据挖掘分析与故障诊断、TMIS 车流推算等方面虽然已经初具成效，但是轨道交通领域产生的绝大多数数据都还没有被充分利用，包括客流信息、物流信息、旅服信息等，这一领域具

有巨大的科研价值和市场价值，还需要团结各方力量，共同推动我国轨道交通大数据技术的开发和应用。

2.5.4　交通管理大数据应用

随着城市的迅速发展，交通拥堵、交通污染日益严重，交通事故频繁发生，这些都是各大城市亟待解决的问题，目前智能交通成为改善城市交通的关键所在。为此，及时、准确获取交通数据并构建交通数据处理模型是建设智能交通的前提，而这一难题可以通过大数据技术得以解决。

对于车辆的稽查与管理是交通管理部门的基础业务，也是对机动车所有者与驾驶员生命与财产安全的保障，因此交通管理部门对于基础业务的及时性与准确性一直有较高的需求，而在实际工作中，受限于原有关系型数据查询技术，进行实时的车辆信息比对与查询会严重消耗系统资源，所需时间也较长。而且近年来，我国投入使用了大量高清卡口设备，存储的车辆拍照信息也呈现爆发式增长，大数据技术高数据吞吐量和高度容错性的先天优势可以使查询达到秒级的速度，有效提升了交管部门的业务水平。

以丹东某交警项目为例，该项目一期建设的卡口电警设备每天产生 300 万条过车记录和过车图片，目前整个系统过车图片数量超过 2 亿条，随着时间的累计还会更多。2013 年 12 月，基于 Hadoop 大数据平台的 UniHadoop 系统在丹东试用，满足了 30 亿条过车记录的检索和数据挖掘业务。

在前端，1050 台图像采集设备中，有 700 多台内置算法的高清一体化智能卡口和电子警察抓拍单元，部署于市内、高速、收费站、国省干道，以光纤连接到中心设备，其中 1.5PB 容量的 IPSAN 存储可以将照片保存 6 个月，视频保存 15 ~ 30 天。而中心管理平台统一以地图作为窗口进行相应的功能和业务展示，单级平台容纳百亿级数据量，查询和统计能在 3 秒内返回，提供基于大数据的各项性能展示，以及轨迹碰撞、拥堵分析等智能研判。该项目在应用中虽然数据量成倍增加，但查询和统计时间仍为 3 秒。该项目投入使用一个多月后，通过对上亿条数据的计算与分析，得到的数据价值有：

- 查获假牌、套牌出租车 7 台，报废车 2 台。
- 查获使用伪造、变造号牌车辆 31 台，故意遮挡号牌车辆 18 台。
- 提供线索破获各类案件 16 起，其中杀人案 1 起、抢劫案 1 起、盗窃案 5 起、治安案件 4 起、交通事故逃逸案件 4 起，布控成功并查获刑事案件涉案车辆及人员 1 起。
- 为群众找回失物 14 起，挽回直接经济损失 10 余万元，间接经济损失千万元。

- 在一起恶性重大事故逃逸案中，犯罪嫌疑人换了 7、8 辆车，最后抛车销毁，但因系统准确地从第一辆车就进行锁定和布控，因此仅 5 个小时就抓获嫌疑人。

面对海量的交通信息，交通大数据的开发应用需求日益凸显，交通大数据时代的来临是智能交通发展的必然趋势，这将为众多企业提供更多的发展机遇和空间。当然在这个进程中我们也将面临诸多挑战：比如交通数据资源分割和信息碎片化；交通信息模式复杂，数据种类繁多；缺乏统一标准和市场化推进机制；基于大数据的交通信息服务产业链、价值链尚未真正形成等。这些问题都有待我们继续探索和解决。

2.5.5 交通运输大数据应用

总部位于美国亚特兰大的全球最大的包裹快递公司 UPS（United Parcel Service）以其高效而闻名世界，5 个工作日在全球的送件量就能达到 15.8 亿件。之所以能达到如此高效的节奏，UPS 拼的不是货车速度，而是靠一个简单的规定：货车不能向左转。为此，UPS 的司机宁愿绕个圈也不往左转。

因为执行尽量避免左转的政策，2010 年，UPS 货车在行驶路程减少 2.04 亿公里的前提下多送出了 350 000 件包裹。这是因为全球大部分国家的路面是实行左驾右行政策的，当货车与相反方向的车辆交叉行驶时，会导致货车在左转道上长时间等待，不但增加油耗，而且发生事故的比例也会上升。于是，UPS 的工程师给不同区域的货车司机绘制了“连续右转环形行驶”的送货路线图，帮助司机选择最适合的道路，长期固定执行这个路线图后，效益也慢慢呈现出来了。

不过，在具体行驶中，UPS 送货车也会偶尔左转，比如在车流量低的居民区，以保证司机的行车便捷迅速。一个简单拒绝左转弯的规定，让 UPS 这样的快递公司实现了高效出货率，不仅带来了可观而且流转更快的现金流，同时因为对环境做出了贡献，企业形象也得到了加分。

事实上，这些都要归功于 UPS 研发的道路优化与导航集成系统（On-Road Integrated Optimization and Navigation，简称 Orion），Orion 算法诞生于 21 世纪初，并于 2009 年开始试运行。该系统的代码长达 1 000 页，可以分析每种实时路线的 20 万种可能性，并在大约 3 秒内找出最佳路线。到 2013 年底，Orion 已经在大约 1 万条线路上得到使用，这让公司节省了 150 万加仑燃料，少排放了 1.4 万立方公吨的二氧化碳。

UPS 基于城市车流数据分析而提出右转的线路图，其实也是一份针对公共交通环境提出的优化解决方案，既可以让每一辆车自觉扮演维持马路秩序的角色，又可以在不增

加成本反而提高效率的前提下，一定程度上帮助解决汽车尾气排放问题。

大数据时代的海量数据对交通运输业会产生较大影响，无论是托运商、零售商、供应商、运营商还是社交网络、个性化网站、移动设备等，行业内或将出现结构变化。大数据技术不仅仅能告诉我们发生了什么，还能通过大数据分析技术预测未来将要发生什么，进而采取各种措施提高效率，获得经济和社会效益。

2.5.6　交通大数据发展趋势

智慧交通是将先进的信息技术、数据通信传输技术、电子传感技术、控制技术及计算机技术等有效地集成于整个地面交通管理系统，从而建立的一种在大范围内、全方位发挥作用的，实时、准确、高效的综合交通运输管理系统。

全面提升交通大数据技术，以应对快速增长的交通数据。据统计，当前交通运输行业每年产生的数据量在 PB 级别，存储量预计可达到数十 PB，以北京市交通运行监测调度中心（TOCC）为例，目前 TOCC 已整合接入行业内外 27 个应用系统，共包括 6 000 多项静动态数据、6 万多路视频，其静动态数据存储达到 20TB，每天数据增量达 30GB 左右。面对增长迅速的海量数据，在云计算、大数据等技术的支撑保障下，未来的交通管理系统对数据存储能力、大数据计算能力、对海量的图像和视频数据的分析能力提出了更高的要求。

提升交通服务“软实力”。服务是交通运输的本质属性，随着我国交通基础设施的不断完善，提升交通服务“软实力”变得更加迫切。我国传统交通信息孤立、系统分散的特点已经成为制约我国交通发展的瓶颈。因此，在大数据分析技术的支持下，丰富交通信息服务内容，深入挖掘交通数据的价值，提高区域协同服务能力，以信息化、智能化为牵引，推动现代信息技术与交通运输管理和服务的全面融合，针对行业需求特点，推动个性化、综合化、智能化、一体化的交通创新服务是交通发展的未来方向。

充分利用物联网，提升数据综合利用率。充分利用物联网，将交通路况检测、GPS、交通监控视频等零散信息进行有效的联系、汇聚和发掘，提高数据的综合利用效率，使其能够真正支撑交通系统的运营管理，提高交通运行效率和安全水平，为公众提供更加便捷、高效的出行环境。

开放与安全兼顾。推进智能交通系统的数据标准化建设，整合数据资源，建立交通信息数据库，成立交通信息数据共享平台，为跨部门、跨区域的交通数据开放奠定基础。同时，还要加强交通数据安全防范措施，提升数据监管和保护能力，维护数据的安全使用。

2.6 健康医疗大数据

2.6.1 健康医疗大数据应用现状

随着计算机网络和信息技术的发展以及现代医学技术的不断进步，与医学和健康相关的数据正在急速增长。如何高效收集、处理、存储、交换和挖掘海量的健康大数据，从而为医护人员的及时和正确诊断、个人健康的监测护理与诊疗建议、医疗相关机构的管理与决策提供大数据分析和系统支持，已经成为跨医学和计算机科学领域的一个重要的研究和产业发展方向。国家相关部门在健康医疗大数据方面部署了一系列相关项目和课题，包括科技部 863 计划和国家科技支撑计划部署的一系列相关项目、中国科学院重点部署项目医学影像信息大数据相关研究以及地方资助项目（如山东省资助的医疗大数据管理及分析应用系统项目）等，旨在基于大数据技术推动医疗健康和生物等相关产业的发展。

健康医疗行业涉及从医疗卫生机构、医疗器械企业、医疗管理部门到具体每个个人共四种不同类型机构实体和人群。健康医疗信息化和服务水平影响着每个人生活质量。健康医疗行业信息化的主要应用现状如下。

首先，随着信息化技术的飞速发展，医疗行业的信息化步伐不断加快。国际上已开始利用大数据挖掘与分析来减少医疗浪费和改善医疗效果。国内医疗行业经过多年的建设和发展，目前医院已经普遍建成了以医院信息管理系统（HIS）、电子病历（EMR）、实验室信息管理系统（LIS）、医学影像系统（PACS）以及放射信息管理系统（RIS）为主要应用的综合性信息系统，极大地提升了各级医院的医疗服务水平。在医疗数据规模和种类急剧增长的情况下，传统的医院内各系统间的数据交换、存储、处理和服务模式，都要适应新形势下的医疗健康大数据服务需求面临巨大挑战。具体包括：如何高效且经济地存储、处理上述种类多（非结构化、半结构化和结构化）、规模大（从 TB 级到 PB 级甚至更多）的医疗健康大数据，为医院日常业务提供支撑；如何更好地实现医疗健康大数据的共享和交换，提高医疗健康数据资源的使用率，方便患者就医，降低个人就医的时间和经济成本等。

其次，大众个人健康服务还处在探索阶段，迫切需要在健康服务方面进行新的尝试和探索，提升医疗健康服务的水平。目前，虽然各类保健器械和医疗服务开始进入社区和家庭，人们的健康消费模式从以往单一的基本医疗消费逐步向医疗、保健和提高身体素质等多种形式并存的医疗健康消费模式转变，但是每个人能享受的医疗卫生资源十分

稀少，有限的医疗资源无法满足不断增长的医疗需求，迫切需要通过信息新技术来改善健康服务水平，为用户提供高效、低成本的医疗健康服务。当前，国际上出现了以云计算为核心思想的健康医疗护理解决方案，该方案包括医疗物联网、医疗门户网站、智能手机终端及健康监测软件、数据管理、IP 及无线通信模块、终端设备的嵌入式软件以及短程无线网关的硬件产品等一系列软件和硬件，基于云计算中心模式共同提供移动医疗健康监测与诊疗服务。

再次，医疗管理部门及医疗相关企业和机构的管理与科学决策，期待更多健康医疗大数据挖掘与分析技术和应用的支持，从而得以充分利用大数据分析的优势来指导相关政策的制定、流行性疫病防控，甚至是医疗服务相关企业的产品规划与决策等。

在健康医疗领域，大数据挑战是指在电子健康和医疗相关数据集十分巨大和复杂的情况下，以致很难用传统的软件、工具和方法来管理和处理这些数据。具体来说，健康医疗领域的大数据包括：临床诊断数据，临床决策支持相关数据（如医生撰写的病历、处方、医学图像，以及其他和实验、药品、保险有关的数据），病人电子健康记录数据，各种医学传感器产生的数据（如移动传感器监测到的心电图数据），社会媒体产生的健康医疗相关数据（如 Twitter、微博、博客、网页上的健康和医疗相关数据）等。健康医疗大数据分析和利用具有重要价值。例如，通过发现数据之间的关联，掌握数据模式和变化趋势，可以改善医疗效果，挽救生命并降低医疗成本。又如，可以基于大数据进行疾病的早期预测、管理人口健康以及检测卫生保健欺诈，根据历史数据，我们可以预测哪些病人需要选择外科手术，哪些病人做手术无效，预测病人的治疗并发症风险等。

2.6.2　国外健康医疗大数据分析的应用

在医疗健康服务过程中产生了大量的数据，如病人保健、治疗记录等数据。尽管大量的医疗健康数据以硬拷贝的形式存储，但当前发展趋势是将这些硬拷贝数据进行数字化。这些数字化的医学大数据具有十分广泛的应用，例如临床决策支持、疾病监测、人口健康管理。仅以美国健康系统为例，2011 年美国健康数据达 150 艾字节。大数据分析每年可以使美国节约大约 3 千亿美元的保健费用，还可以减少浪费，改善医疗效率。

美国医疗联盟网络由超过 2700 个成员医院和健康系统组成，包括 40 万名各类医生。该网络收集和存储大量临床、财务、病人和供应链数据，这些大数据为大约 330 个医院提供医疗决策支持服务，改进医疗过程，挽救了大约 29 000 个生命，减少医疗开支

达 70 亿美元。例如，哥伦比亚大学医学中心利用大数据分析脑损伤病人的生理数据流的相关性，从而为专业人员提供重要且及时的信息以治疗并发症。密执安大学健康系统将大数据分析应用于规范化输血管理。应用表明，大数据分析减少了 31% 的病人输血，并且每月减少医疗成本达 20 万美元。

北约总医院是一家有 450 张床位的加拿大多伦多社会教学医院，利用实时大数据分析方法提供临床、行政、财务服务，从而帮助病人更有效地康复。

意大利博洛尼亚 Rizzoli 整形外科研究所通过大数据分析家庭内成员之间的临床变化，及时发现每个人呈现出的不同程度的症状。这种分析可以减少住院病人达 30%，并减少医学图形检查达 60%。未来它们希望结合大数据和基因分析，开发更有效的治疗方法。

加拿大儿童医院利用大数据分析婴儿容易出现的致命性院内感染从而改善医疗效果。通过收集临床检测装置产生的生命体征数据，医护人员通过大数据分析方法，可以提前 24 小时识别潜在的感染迹象。

2.6.3 大数据技术提升传统医疗信息系统效率

医院日常运作每年都会产生海量的医学和健康相关数据，这些数据难以用常规方法进行分析处理。针对这些健康医疗大数据的分析处理，对政府卫生政策制定、公共卫生机构疾病流行规律掌握和有效防治、医院运营决策、临床医生科研等具有非常重要的意义。但是，目前医学数据的搜集、分析和利用仍然不够方便快捷，对数据的处理仍然需要相当的人工干预或操作。通过充分发挥大数据技术的作用将极大推动医学信息化的发展。下面从肿瘤登记方法的演变，通过医院采用大数据技术自动从医院电子数据库整理和统计分析肿瘤资料，来介绍大数据技术如何提升传统的医疗信息系统效率。

1. 肿瘤登记面临的挑战和方法演变

做好肿瘤登记能起到有效防治恶性肿瘤的重要作用。传统的登记方法面临许多问题，资料的收集与整理不仅非常辛苦、繁琐、费时，且不断累积的资料难以用人工甚至是半人工方法进行整理，资料编码复杂且困难，难以确保准确性，登记人员疲于资料的收集、核对、编码和上报，更是无暇充分利用登记的资料。

随着信息技术和医疗信息化发展，网络、电子病历、HIS 和 PACS 等系统得到普遍应用，传统的手工登记方法逐渐被现代电脑系统登记方法取代，网络直报代替了人工卡片上报，大大减轻了肿瘤登记人员工作量。但目前的登记方法也存在明显不足，网

络直报没有减轻临床医生的负担，花费的时间仍较多。如何减轻临床和登记人员的压力与负担，实现肿瘤自动登记需求，需要实现对非结构化的医学大数据的收集和处理的自动化。

2. 肿瘤登记软件的出现和应用

国内外都非常重视肿瘤登记软件的研发。国内河南省肿瘤防办最早开展此方面研究，而全国肿瘤登记中心 2002 年汉化了国际癌症研究所的 CANREG4 登记软件，并在全国推广使用。自此全国多地陆续开展了肿瘤登记技术的研发与应用。同一时期，国外也有许多医疗机构开展了肿瘤登记技术的研究实现与统计分析。国内的肿瘤登记技术研究主要集中在资料收集后的处理，如录入、整理、储存、编码和统计分析，缺乏自动处理功能。国外相关技术虽然能达到一定的自动化效果，但由于语言和环境不同，难以在国内直接应用。

3. 大数据技术实现肿瘤非结构化数据登记过程自动化

某家三甲医院约有平均 3000 条 / 天 PACS 诊断记录，一个月约 10 万条诊断记录数据，需要人工干预判断，该医院存有十年的病人数据。全国有 1431 家三甲医院，其余的三乙到二丙医院共有 9798 家。如此大量的数据若采取人工干预将需要耗费大量的人力，工作人员需要在 PACS 诊断中人工筛选和整理数据，80% 的工作都消耗在了这些数据的收集工作上。

肿瘤自动登记技术可自动进行数据的发现、提取、整理、编码和统计分析。首先借助自然语言处理技术对肿瘤非结构化数据进行预处理，即从医院电子数据库中自动发现、提取、整理和编码所需肿瘤资料，其次建立整理后的数据仓库，再次根据需要进行数据自动分析，最后自动展示分析结果。基于大数据处理技术架构搭建起可方便扩展的肿瘤数据的自动处理和分析应用，极大地提升肿瘤登记的自动化水平。

通过与手工查找方法的对比分析，肿瘤自动登记技术完成的工作结果显示其准确性相当高，对比 3 个月数据测试结果的灵敏度为 98.91%，特异性为 99.22%。肿瘤自动登记技术应用后，以往 15 天的工作现在只需要 15 分钟，极大地提高了工作效率及质量。

肿瘤自动登记技术具备以下优点：

- 不影响原有信息系统的运作，适用于不同信息系统的数据获取。
- 运行速度快，数十秒即可完成查找过程。

- 准确性高，灵敏度和特异度均达 95% 以上。
- 极大地减少了统计和分析工作量。

4. 基于大数据技术的肿瘤自动登记理念在医学领域的推广应用

肿瘤登记自动化对肿瘤登记无疑是一场革命，将为肿瘤防治带来巨大影响。肿瘤监测通过大数据的采集、分析处理和存储实现了肿瘤的自动化和统一处理。类似地，其他疾病（如传染性疾病、慢性非传染性疾病）的监测或者其他医学过程的监测，以及相关医学数据的提取和分析，也非常有可能通过大数据技术来实现某种程度的自动处理。

2.6.4 大数据在区域化医疗卫生管理分析中的应用

经过二十多年的卫生信息化建设，目前卫生信息化正朝着面向人群健康的区域卫生信息化建设阶段发展，对信息化服务模式、海量大数据处理、节能环保等提出了更高的要求。医疗卫生主管部门通过数据分析平台更全面地掌握区域内健康、诊疗、疾病及经济信息，为区域医疗规划和管理科学决策提供数据支撑。随着医疗数据的不断积累，如何通过大数据分析提升区域医疗机构的运营管理能力，从而提升区域医疗机构整体服务能力和服务质量成为关键问题。

当前医疗卫生信息化还处于各自开发和各自建设阶段，而医疗数据纷繁复杂，包括业务流程、财务、药品、设备、耗材、医疗质量、临床等各种数据，以上数据分布于各个医疗厂商系统内部，存在数据异构及信息孤岛情况。因此，区域卫生机构内各系统在实现互联互通、数据共享和应用方面面临着各种挑战：

- 如何获取各医院数据
- 如何整合不同厂家的数据
- 如何确定有价值的数据
- 如何通过数据发现问题
- 整合分析数据后如何与医院共享

依托大数据分析技术从纷繁的临床数据、财务数据和运营数据构成的数据海洋中提炼高价值数据，帮助区域卫生机构深入开展医疗资源、就诊动态、疾病预警、患者分析和用药分析等工作。

区域化医疗健康数据分析平台（见图 2-21）的总体目标是将不同医院的不同系统数据通过关联和整合建立一个统一的医疗健康大数据分析平台，服务于区域化医疗、健康相关的管理、分析和决策支持功能，提升区域化医疗健康整体水平。

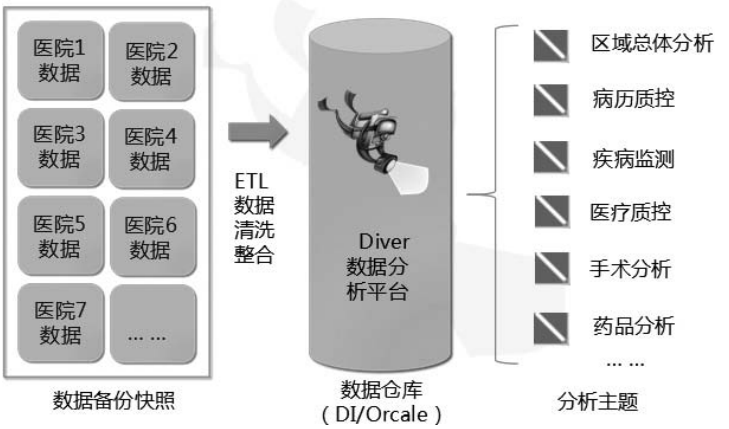


图 2-21　区域化医疗健康数据分析平台示意图

大数据技术为区域卫生机构提供决策支持。在区域医疗卫生数据分析平台上存储着区域内全体居民的健康信息、电子病历、治疗方案、医务人员工作量等信息，区域卫生机构主管部门对这些信息的统计、分析和挖掘对区域内公共卫生政策的制订有着积极的意义。例如通过大数据分析人员流动数据来监测传染区人员的主要流向并提出预警提示，主要流向的当地疾控中心可及时准备相应的人员和药品，为当地人接种疫苗，这样就避免了盲目的全面布局情况和资源浪费。

区域医疗卫生机构可通过大数据统计、挖掘和分析实现各类医疗、健康相关的一系列分析功能，如图 2-22 ~ 图 2-24 所示。

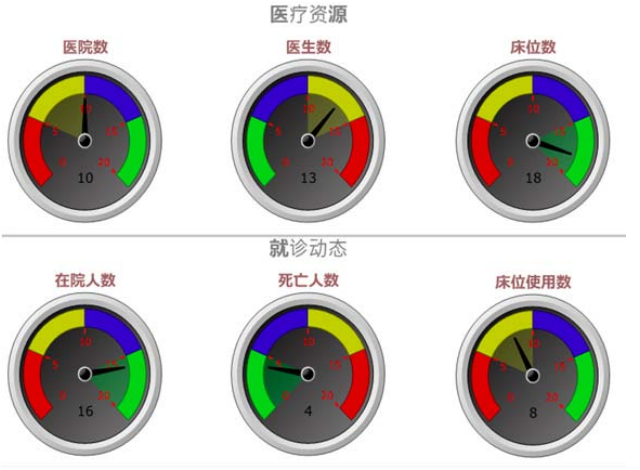


图 2-22　医疗卫生总体情况分析：医疗资源、就诊动态、疾病预警

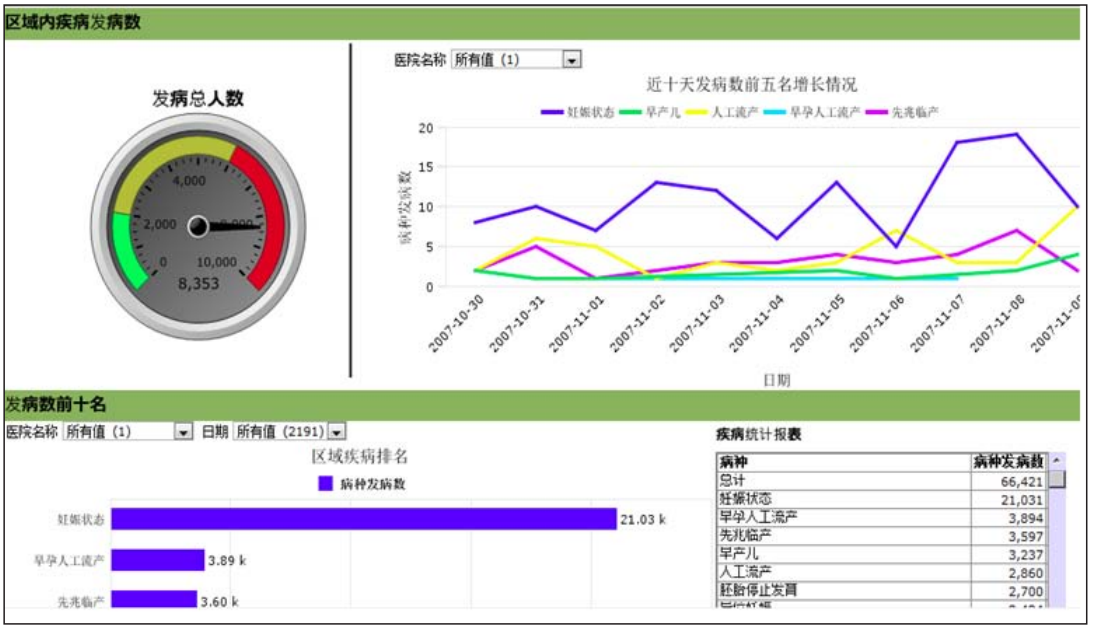


图 2-23　疾病监测：传染病（甲类、乙类、丙类传染病）和慢性病

图 2-24　医疗质控：医院工作效率监控、医疗质量监控和医院经营状况监控

2.6.5　基于互联网大数据的疾病指数预测应用

近年来，我国突发公共卫生事件频发，从 2003 年的“非典”、2004 年的 H5N1 禽流感、2009 年的 H1N1 禽流感再到近年的 H7N9 禽流感，每一次疫情爆发均对人民的生命财产及心理造成了重大的损害。随着信息技术的发展，信息传播的方式多样化，信息传播的速度和范围有了质的飞跃，突发公共卫生事件发生后，对群众心理和社会的影响越来越大，对其危害的评估与控制也越来越难。2009 年以来，随着微博等社交新媒体在中国的兴起与发展，互联网逐渐成为信息传播的主体，个人之间的信息传播更加便捷，信息传播的特征也与以往有所不同。因此，基于互联网大数据来进行相关疾病流行的预测变得可能。

国际上，谷歌 2009 年对甲型 H1N1 流感爆发的预测比美国疾病控制与预防中心（CDC）提早 1 ~ 2 周，当时震惊了整个医学界和 IT 领域的科学家，研究报告发表在《Nature》上。在国内，百度公司 2014 年 7 月上线了“百度疾病预测”[㊀]，借助最新大数据技术为用户呈现身边的疾病信息，不仅可以了解当前的流行病态势，还可以看到未来 7 天的变化趋势，提前做好预防措施。在构建流感预测模型的过程中，中国疾病预防控制中心的流感监测结果起到了参照作用。国内 TRS 公司也开展了基于微博的疾病预测工作。

下面以根据微博数据对流感的分析预测为例，介绍基于互联网大数据的疾病指数预测的应用。

流感的传染性强、发病率高，容易引起暴发流行或大流行，且处理不当易造成较严重后果，一直以来都是公共卫生管理的重点。以流感为例，以新浪微博中全部微博达人和身份认证用户所发原创微博为样本，选取与流感相关的感冒、发烧、发高烧、咳嗽、鼻涕、流感、输液、吊瓶、鼻塞、流涕作为检索关键词，监测 2011 年 01 月 01 日 ~ 2012 年 10 月 22 日之间的所有数据，展开多维度分析。本次监测数据总量为：2011 年，615 449 910 条微博，含上述关键词的 2 858 212 条微博；2012 年，500 343 713 条微博，含上述关键词的 2 430 082 条微博。分析中所用的结果值均做过归一化处理，即结果记录数 / 当天数据总量。

监测结果的多维度分析如下。

1. 微博流感时间分布

按照每日提取的微博流感相关用户数，得出时间趋势图如图 2-25 所示。

图 2-25 中，纵轴为每日微博中提到流感相关关键词的微博人数，横轴为日期（图中横、纵轴说明下同），跨度为 2011 年 01 月 ~ 2012 年 10 月。可以看出，9 月中旬开始到年底为秋季流感多发季节，3、4 月为春季流感多发季节。

㊀ http://trends.baidu.com/disease/#。

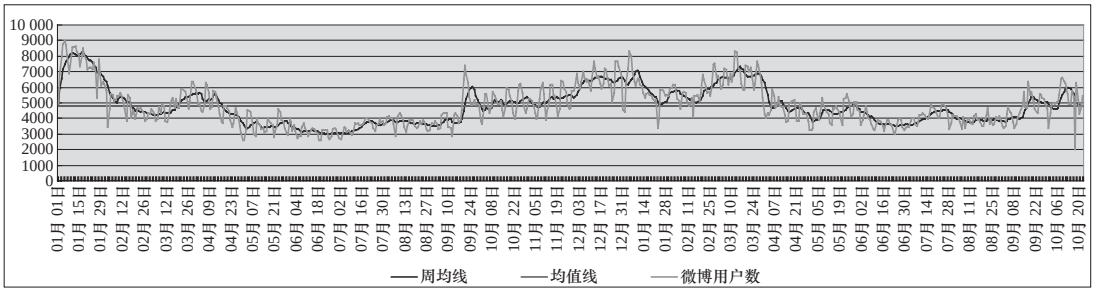


图 2-25　微博流感时间趋势图

把 2011 年与 2012 年的曲线叠加后，得出时间趋势图如图 2-26 所示。

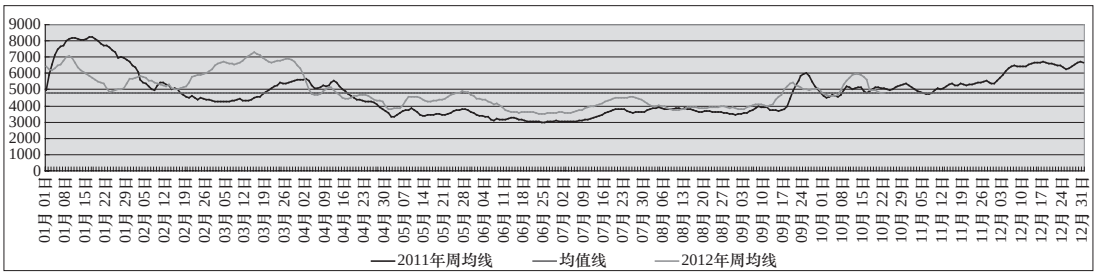


图 2-26　微博流感整年时间趋势图

从图 2-26 中可以看出：

- 2012 年 01 月份流感强度没有 2011 年同期强，但 02 月下旬和 03 月较 2011 年强，但峰值有所回落。
- 2012 年秋季流感爆发点比 2011 年提前 1 周，2011 年是 09 月 20 日，2012 年是 09 月 14 日。
- 2012 年秋季流感趋势平稳，没有大规模爆发迹象，今后将维持高位运行，爆发峰值逐步推高。

2. 微博流感地域分布

根据博主的地域属性可以做地域分析，以北京为例，如图 2-27 所示，北京发生流感的高峰期在冬季，以每年的一月中下旬为甚。北京 2012 年 01 月流感强度明显低于 2011 年同期，2012 年 03 月则高于 2011 年同期，2012 年秋季流感较 2011 年提前。

类似地，提取出全国其余省市的数据，可以画出微博流感的全国地域分布图。以冬季（2012.02.01）和夏季（2012.06.01）为例做对比，见图 2-28。全国大多省市在冬季都为流感高发季节，相对来说北方省市较南方更易发生流感；而每年的夏季则是流感少发季节，相对来说，南方省市较北方更易发生流感。

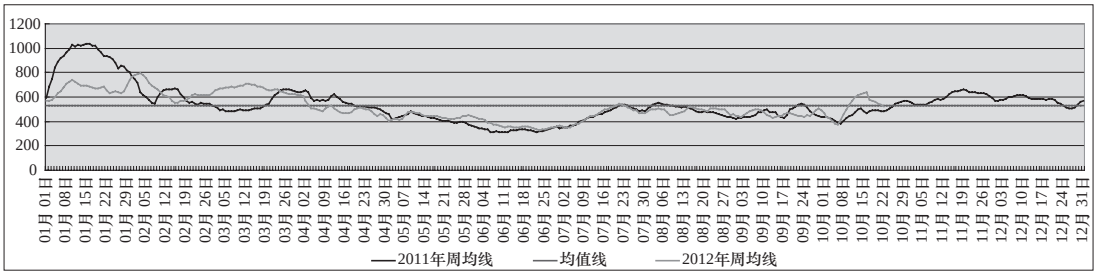
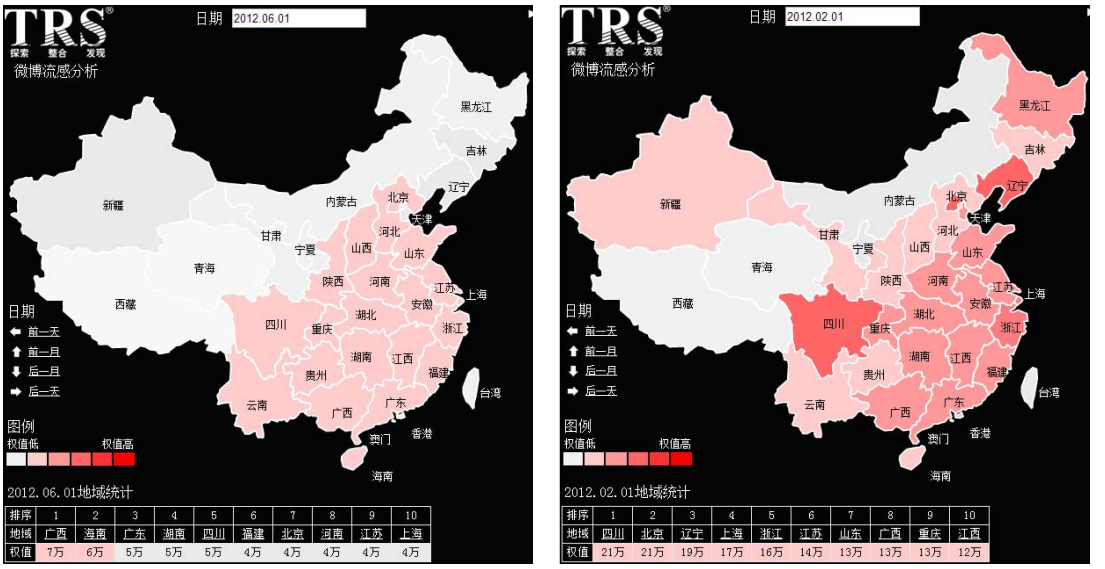


图 2-27　北京微博流感博主整年时间趋势图



冬季（2012.02.01）　　夏季（2012.06.01）

图 2-28　微博流感的全国地域冬季和夏季分布图

基于微博数据的流感分析可以较真实地反映当前社会的流感人数、地域分布和性别分布，因此可以在一定程度上预警流感的爆发，当检测到某一区域流感人数连续明显有别于其他区域和该区域历史数据时，可做出流感即将爆发预警。由此可见，对微博进行实时动态监测，及时发现网络上与公共卫生相关的突发性事件和敏感舆情，第一时间进行危机预警，并对舆情信息进行持续跟踪，实现对公共卫生领域的有效掌控，将推动公共卫生事业的良性和快速发展。

2.6.6　健康医疗大数据发展趋势

随着信息技术的快速发展，医学与健康相关大数据的高效存储和处理成为可能。如

何基于大数据技术来实现健康医疗相关大数据的高效收集、处理、存储、共享和挖掘分析，从而为医护人员的及时和正确诊断、个人健康的监测护理与诊疗建议、医疗相关机构的管理与决策提供大数据分析和系统辅助支持，成为跨医学和计算机科学领域的一个重要的研究和产业发展方向。

如前所述，健康医疗行业涉及从医疗卫生机构、医疗器械企业、医疗管理部门到具体每个个人共四种不同类型机构实体和人群。相应地，对医疗健康大数据发展有以下趋势预测。

医疗卫生机构将继续依托大数据架构和技术，提升现有的医疗信息化水平，实现异构医疗健康大数据的高效处理和服务，医疗健康数据具有种类多（非结构化、半结构化和结构化）和规模大（从 TB 级到 PB 级甚至更多）的特点，大数据技术可以为医院日常业务提供更好的支撑，提高医疗信息化的性价比和可靠性。同时，为应对医改的不断深入，医院和社会对医院综合信息资源的处理与应用需求不断提高，在医疗数据规模和种类急剧增长的情况下，传统数据模式无法适应新形势下的服务需求。在医疗健康大数据挖掘和分析的基础上，更多大数据技术进一步为疾病的诊断和治疗等过程提供辅助支持，更好地服务于诊疗过程。

随着社会经济的发展，高质量的个人医疗健康服务需求迫切。同时，由于医疗健康资源有限，特别是高质量的医疗健康资源不足，因此利用大数据技术更好地实现个人健康服务、个性化诊疗辅助服务具有很好的前景。基于云计算技术的健康医疗护理服务将为个人提供移动医疗健康监测与诊疗推荐服务。

医疗管理机构，包括区域性医疗卫生管理机构乃至全国卫生管理机构，都将更好地利用大数据技术辅助实现管理决策、资源调度、流行性疫病防控与应急处置的合理性和科学性。

由于医疗健康大数据和个人密切相关，其安全和隐私保护问题在管理制度和技术保障两方面都尤为重要。基于云计算平台的开发和应用使得大量用户相关数据被存储在云端进行处理，并基于此提供查询和搜索服务。因此，除了要解决医疗大数据处理和查询搜索服务时的可靠性、可扩展性和效率外，下一步必须考虑大数据存储、处理、搜索和使用的安全和隐私问题。

2.7 政府大数据

2.7.1 政府大数据应用现状

政府数据是一笔巨大的财富，也是开启智慧政府大门的钥匙。智慧政府离不开政府开放数据，而政府开放数据形成的生态圈，则将有力地推动智慧政府又好又快地发展。

政府数据总量及种类庞大繁多，与民众生活密切相关。虽然政府拥有这些高价值数据，但是其数据资产的利用和运营仍处于较为原始的阶段。

2009 年 1 月，美国奥巴马政府签署了《透明和开放政府》备忘录，提出要制定开放政府计划，建立前所未有的开放政府。2009 年 5 月，美国开放数据门户（Data.gov) 网站上线，正式拉开了全球政府开放数据的大幕。

自 2009 年美国开放数据门户（Data.gov) 开通上线以来，许多国家纷纷跟进，制定开放数据战略或政策，加快了数据开放的步伐，英国、法国、德国、日本、新西兰、澳大利亚、加拿大等国家先后出台了相关举措。2013 年 6 月，八国峰会（G8 Summit）期间签署了《八国集团开放数据宪章》(G8 Open Data Charter，简称《G8 开放数据宪章》)，标志着开放政府数据已经成为全球共识。

Open Knowledge Foundation 发布的 2014 年全球开放数据指数显示，英国、美国分居全球政府开放数据排名的前两位，在开放数据的深度和广度方面遥遥领先全球（见图 2-29）。2014 年，中国在这份榜单中仅仅排名 63 位（见图 2-30）；而 2013 年，中国在这份榜单中的排名是 34 位。

图 2-29 （2014）全球开放数据指数[㊀]

㊀ 数据来源：Open Knowledge Foundation 2014/11/27。

图 2-30 （2014）全球开放数据指数中国排名㊀

政府开放数据的滞后，已经对中国大数据产业的发展以及智慧政府的构建造成了一定的影响。虽然存在这些制约条件，但是国内智慧政府的建设并未止步，而大数据在其中的应用更是越来越深入。

2.7.2 政府大数据入口整合

1. 建设背景

信息发布、电子政务服务、公众互动是政府门户网站的三大功能。经过多年的建设，厦门市政府信息化程度获得了较快发展，但还存在各部门惠民服务相对独立的现象，客观上形成孤岛，百姓需多次注册，无法便捷享受信息惠民服务。此外，公共服务均等化、均质化也还需进一步深化。

以“建设惠民服务门户，实现一站式服务，推动惠民公共服务进家庭，促进公共服务均等化、均质化”为目标，从百姓最为关注的医疗、教育、社保、交通、政务等方面入手，厦门市把一站式公共服务门户的建设快速提上了日程。

㊀ 数据来源：Open Knowledge Foundation 2014/11/27。

2. 公共服务门户简介

“i 厦门”是一个一站式公共服务平台（见图 2-31），它充分运用物联网、移动互联网、云计算、大数据、导航、电子商务等信息技术手段，更好地分析、融合政府及社会各类惠民服务资源与应用，从而面向本市市民、企业、社团、政府四类用户对惠民服务各种需求做出智能感知与及时响应，为全社会持续提供一体化、个性化、便捷化、安全的在线惠民服务。

图 2-31　“i 厦门”一站式公共服务平台

3. 公共服务门户特点

围绕“全面感知、深度融合、服务创新、协同运作”的建设理念，拓尔思整合了来自市公安局、图书馆、气象局、卫生局、社保管理中心、计生委等部门的 150 余项服务，同时将 20 多项市民服务搬上微信，涵盖电子政务、市民生活、健康、教育、文化、交通、社保等领域，极大地方便了市民的生活。“i 厦门”公共服务门户的主要特点有：

实名账户一证通行。市民可以用有效身份证件号码在平台或线下注册个人实名账户，实名账户登录后，个人办事、证照、账单、健康档案等服务一页搞定，最终实现个人实名账户在全市所有政府应用中的通行。

网上办事智能增效。市民个人主页基于个人实名账户，实现事前清楚、事中少走、事后评议，办事全过程服务均为个人量身定制。市民在线办理业务时，系统智能感知用户身份，无需用户重复填写个人信息，可以方便地跟踪办事进度和办事结果，如图 2-32

所示的“一孩生育”一站式公共服务。

个人服务按需定制。市民通过个人主页可以查阅来自政府部门和公用事业单位的个人数据，包括个人应用、我的账单、我的办事、我的证照、我的收藏等，全面汇聚为“我的资料库”。同时，市民个人主页还从一个人的全生命周期、家庭、社区三个维度，主动推送待办事项与公共服务。

多屏融合服务随行。针对各种终端的屏幕特点以及各类人群使用不同终端的场景，平台可以支持网页、手机、PAD、智能电视等多个渠道访问。

图 2-32　一站式公共服务：一孩生育

4. 结语

“i 厦门”一站式公共服务门户改变了以往政府门户网站建设中“重建设轻整合”、“重管理轻服务”的弊病，从入口整合入手，进一步探索数据整合、服务整合，这种探索和实践对于大数据时代政府门户网站的建设具有积极的现实意义。

2.7.3　政府大数据惠民服务

1. 背景简介

迪拜是阿联酋人口最多的城市，也是中东最富裕、最国际化的都市。迪拜实行自由和稳定的经济政策，在各国之间以及国际工商界赢得了良好的声誉，这就鼓励本国资本和外国资本投资于商业、工业和服务业等各个经济领域。2013 年，迪拜游客人数首次超过 1100 万人次，较上年同期增长 10.6%。

为了 2020 年的世博会，迪拜酋长亲自启动“智能迪拜”战略计划，要在三年内将迪拜建成世界上最智能的城市。“智能迪拜”计划包括 100 个举措，拟提供 1000 项智能化政府服务，包括六个方面：智能生活、智能交通、智能社会、智能经济、智能管理与智能环境。通过该计划，迪拜将改变人们的生活方式，使整个城市的服务均可通过智能手机解决。

2. 迪拜智能政府公司启航，“智能迪拜”全面启动

“智能迪拜”将迪拜所有政府部门纳入其中，并且设立了专门的机构——迪拜智能政府公司（Dubai Smart Government’s Corporate），由该公司全权负责“智能迪拜”的投资和建设。迪拜智能政府公司综合性地运用了集中式和分布式的方法，以便更好地组织和利用政府资源，同时开展面向政府内部和面向公众的智慧城市建设。迪拜智能政府公司定位如下：

- 支持服务部门。迪拜智能政府公司是“智能迪拜”的一部分，同时面向各个政府部门提供建设所需要的人力资源、采购、财务、管理、法律等服务。
- 战略制定和管理部门。制定并执行“智能迪拜”计划，了解“智能迪拜”建设的最新进展，修订发展战略。
- 商务开发部门。负责“智能迪拜”的技术、资源引入和商务开发，促进“智能迪拜”的可持续发展。
- 基础设施管理部门。负责提供建设“智能迪拜”所需要的基础设施服务。
- 政府资源规划部门。负责提供 GRP（Government Resources Planning）服务，并进行相应的管理。
- 智能服务提供部门。负责将组成“智能迪拜”的各类服务整合起来，提供给公众。

3.“智能迪拜”建设现状

“智能迪拜”将“可接入性、内容、功能性、可用性、界面、无缝性、可靠性”作为

衡量指标，把公众的满意度、公众的采用率作为参考要素。在这一指导原则下，“智能迪拜”的建设速度十分之快，目前已经推出了 16 项突出服务：

- 单点登录所有服务
- 移动紧急支付
- 网上支付电费
- 网上注册车辆
- 移动终端查询司法进度
- 24 小时获得健康证书
- 网上申请健康卡
- 在线法律咨询
- 电子方式提出意见和建议
- 电子投诉
- 道路电子通行
- 网上注册公司
- 手机追踪车辆罚款
- 短信预定出租车
- 网上规划迪拜行程
- 游客在线投诉

迪拜未来计划尽可能将所有政府公共服务电子化，实现足不出户即可办理所有政府公共服务。

4.“智能迪拜”的启示

相比欧美发达国家通过开放数据和政府服务外包建设智能政府的思路，“智能迪拜”的建设主要采用设立专门机构全权负责的方式，并且适当引入市场化的运营管理机制，并且在短时间内取得了明显的成就。毫无疑问，“智能迪拜”的建设方式对国内也有一定的借鉴和参考意义。

2.7.4 政府大数据社会治理

1. 背景简介

电子商务已经深刻地影响着我们的日常生活，使人们足不出户就能够享受互联网带来的各种便利。但是电子商务在其中一个领域却是例外——医药零售业。与 2013 年中国医药市场 11 463 亿元的总体规模、2571 亿元的医药零售规模相比，医药电商只占据了很小的份额。有数据显示，2010 年医药电商市场规模为 2 亿元，2012 年超过 16 亿元，2013 年为 42.6 亿元[㊀]。

与此同时，通过网络制贩假药的活动也日趋猖獗。以微博、微信平台为核心主导的社会化媒体在网络中的影响力日渐趋强，网络营销也随之突飞猛进地发展，各门户网站、搜索引擎、微博等社会化媒体推广成为各行业商业活动的重要宣传手段，食品、药

㊀ 数据来源：中康资讯，中国医药行业六大终端用药市场分析蓝皮书（2013 ~ 2014），2014。

品、化妆品、保健品、医疗器械等营销信息以文字链、展示广告、软文、搜索引擎推广链接等形式高频出现在各种宣传渠道中，其中虚假信息和不法广告占据了很大比例。在 2012 年 6 月 28 日国家工商总局等五部门联合召开的贯彻实施《大众传播媒介广告发布审查规定》电视电话会议上，国家工商总局表示，对主要商业网站的广告监测表明，我国网络广告违法问题十分严重，一些网站发布的医疗、药品、医疗器械、保健食品、化妆品等广告的违法率高达 90% 以上。

2. 打击网络非法制贩假药刻不容缓

根据国家药监局规定，网上售药必须具备《互联网药品交易服务资格证》和《互联网药品信息服务资格证》，但是仍旧存在大量的非法网站违规从事药品交易，非法医药广告横行网络。另一方面，不法分子利用隐蔽手段，通过正规医药电商购买感冒药制售冰毒的案例在国内更是层出不穷。总体来看，利用网络制贩假药、非法买卖药品具有以下几个特点：

宣传欺骗性强。虚假广告信息充斥网络，有的冒用科研机构、医疗单位甚至是政府部门名义，有的编造进口药、特效药的信息和神奇疗效，有的专门设计了针对老年病、慢性病的骗术。

网上渠道多样、灵活易变，产品追责难。网上药品非法销售把汇款、快递、第三方物流等作为主要售假方式，规避日常检查、产品抽验等监管措施，公众在受骗后难以投诉、难以索赔。

隐蔽性强，案件查处难。不法分子通过注册虚假网站、提供虚假身份、宣传虚拟产品等进行伪装，有的不法分子服务器设在境外，有的组织跨区域作案并和监管部门打游击战，致使网上售假问题屡打不绝、屡禁不止。

3. 利用大数据技术打击网络非法制贩假药

针对网络非法制贩假药欺骗性强、渠道多样、隐蔽性强的特点，拓尔思配合监管部门，利用互联网舆情大数据监测技术，从多个维度强化对网络制贩假药的监测。

针对全国三品一械 240 余万家生产经营企业的监管，互联网舆情大数据监测系统支持 18 万种现有药品、40 000 种化妆品、42 000 种医疗器械产品及新产品的实时认证和备案，实时掌握三品一械的全国生产经营状态，建立三品一械所有流通路径的可追踪、可回溯、可分析、可监管的统一业务平台。

对于互联网舆情大数据监测的信源，其主要来源于分为以下四类：

广告宣传。造假者通过招聘网站、网页推广、论坛软文、博客等发布的宣传内容。

群众反映。受害者在线上投诉网站的投诉内容以及食药总局网站接收的投诉内容等。

举报信访。受害者在网络论坛、贴吧、问答社区、个人博客等举报、信访内容。

媒体曝光。官方及各类商业媒体的曝光、官方微博的曝光等。

互联网舆情大数据监测平台对搜集的信源数据进行统一的存储、清洗、排重、标引，形成有效的、实时更新的数据集，建立并完善虚假网络信息分类体系与识别词典，通过危害指数等量化评估模型，以便在海量的非法信息中识别出影响力大、危害度高、老百姓痛恨、管理部门头疼的重点打击对象。

4. 结语

互联网舆情大数据监测平台厘清了食药领域三品一械生产销售机构、国家监管机构、广告发布平台的关系，切断生产销售机构与广告平台的直接交易，构建生产销售机构提出广告申请、国家监管机构审核通过、广告发布平台发布广告的良好生态，通过数据中心的 API 接口服务，实现对三品一械网络广告交易的监管，在打击违法产品、违规网站，切断其利益链条方面取得了较好的效果。

2.7.5 政府大数据宏观经济管理

1. 工商大数据简介

工商部门在履行市场主体准入与监管职责过程中，采集和掌握了包括企业、个体工商户等各类市场主体从出生到消亡整个生命过程中的主客体及行为信息（见图 2-33）。对应于自然人户籍登记，工商部门的市场准入工作就是确立“经济户籍”（曾称为“经济户口”）的过程，市场主体登记管理的信息经过汇总（通过互联网和现代化的电子信息技术）形成了“国家经济户籍库”。

随着企业信用分类监管系统的建立和完善，总局、省级局、地市局、县级局、工商所五级联网已经贯通，工商总局和省级局两级数据中心已基本建立，全国工商行政管理部门数据信息共享和数据交换基本能实时完成。目前，工商部门掌握的“国家经济户籍库”有一千多万企业法人和数量庞大的非法人机构（截至 2014 年一季度末，共有 1560.63 万户企业，4510.46 万户个体工商户），从主体总量保守估算，占国家法人库的 90% 以上。

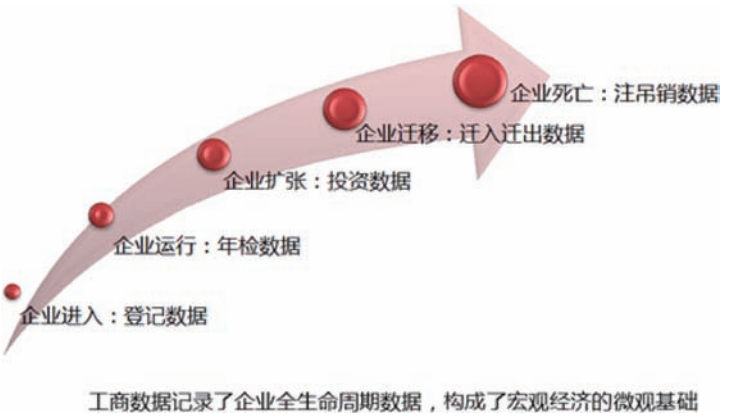


图 2-33　企业全生命周期数据情况

与其他部门相比，工商数据具备以下几个特点：

全面性。工商部门掌握了企业、个体工商户、农专等全部市场主体的信息，是全量数据而非抽样数据，是区域经济的“户籍库”。

全周期性。工商部门掌握了企业从出生、成长、监管乃至死亡的全生命周期的数据，是连续性的时间数据而非截面数据，能够更好地反映企业乃至区域经济的成长变化规律。

细粒度。工商数据是细化到每个市场主体的数据，而非汇总后的统计数据，能够进行细致深入的分析。

即时性。工商数据变动基本上能够在数据库中实时反映出来。

对经济运行具有先行性。工商数据中市场主体的发展变动能够在很大程度上反映资本的流动情况。资本作为经济活动的源泉，经常在经济活动中扮演着先行者的角色，资本流动体现了对未来实体经济走势的判断。

因此，立足工商数据资源，借助大数据分析技术，通过相关研究探讨市场主体发展变动与经济发展的关系，从工商角度反映区域经济发展的特点，评估各种经济政策的影响，对相关部门准确判断经济形势、有针对性地制定政策、合理调配资源具有重要意义。

2. 基于工商大数据的企业发展工商指数构建

宏观经济管理中利用先行指标的变化来预测经济走势，对决策的前瞻性和针对性具有重要意义。一个好的先行指标应当能够及时提前反映经济走势的变化，且具备数据真

实可靠、获取经济快捷、编制科学合理等特性。

工商企业登记数据作为市场投资行为最直接的反映，能够用来构建经济预测的先行指标。国家工商行政管理总局联合龙信数据组织“企业发展与宏观经济发展关系研究”课题组，依托工商企业全量登记数据，借助大数据分析技术，构建了企业发展工商指数，该指数与 GDP、公共财政收入等经济指标具有显著的正相关性，并表现出稳定的先行关系，能够提前 1 ~ 2 个季度预判宏观经济走势。基于企业发展指数开展的宏观经济及区域财政收入预测研究，是我国政府领域大数据挖掘的首个成功案例，多次在国务院经济工作会议中得到应用，对宏观经济决策的前瞻性和针对性具有重要意义。

如图 2-34 所示，2008 年 6 月全国财政收入增幅下滑，经济危机影响凸显，而企业发展工商指数在 7 个月前的 2007 年 11 月就开始下滑，提前反映了经济危机的影响。2009 年 1 月企业发展工商指数出现反弹，比 2009 年 7 月财政收入增速出现反弹提前了 6 个月。对于 2012 年的经济下行引发财政收入增幅的放缓，企业发展工商指数也提前 2 ~ 5 个月有所体现。研究还发现，企业发展工商指数与 GDP 增长率也有类似的先行关系。这表明企业发展工商指数能够先行感应经济形势变化，是宏观经济的晴雨表。

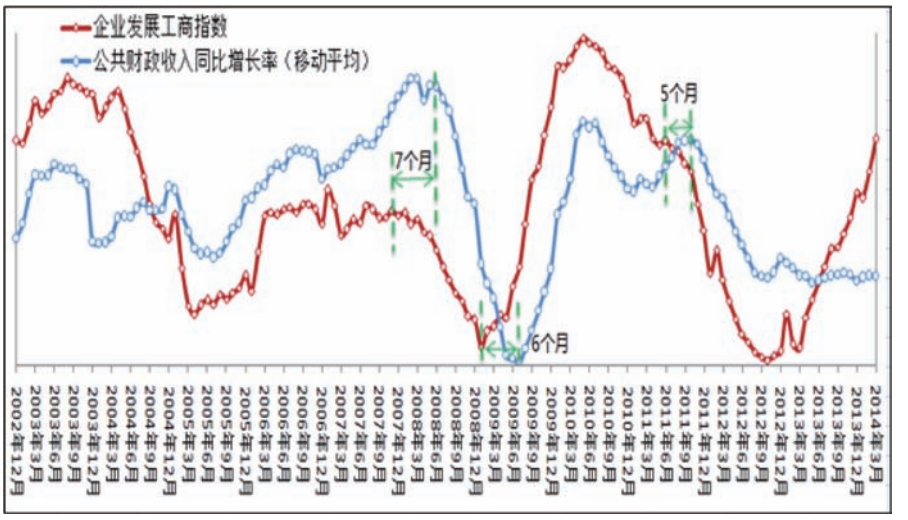


图 2-34　全国企业发展工商指数和财政收入（移动平均）调整后同比增长趋势图

企业发展工商指数的编制过程包括指标初选、数据预处理、指标分类和指数合成四个阶段。初始指标的选择尽可能涵盖企业的成立、成长、发展、消亡的全过程，并同时

考虑不同类型企业的发展变化特点。初始指标选择了不同维度（规模、行业、主体类型）的企业新设、注销、投资额和注册资本等 33 个指标。

为了能够准确发现不同指标与宏观经济的相关关系，对原始数据进行了预处理，主要包括对数据的异常值处理、季节因子调整、数据标准化以及量纲转化等，以此减少孤立事件、季节波动以及随机波动对分析结果的影响。通过对处理后的数据与 GDP、公共财政收入等基准指标进行时差相关分析计算，根据其与宏观经济的相关变化关系，将初始指标分成了先行指标、同步指标、滞后指标和弱相关指标四类，并选择对宏观指标具有先行关系的 10 个指标作为企业发展工商指数的基础指标，并对基础指标值进行主成分分析，将主成分的系数标准化后作为指标合成的权重，最终合成了企业发展工商指数（企业发展工商指数的取值范围在 0 ~ 100 之间，其环比变化能够较好地反映企业发展趋势），这种方法能够最大程度地反映指标的变动信息。

2.7.6　政府大数据发展趋势

智能政府不仅仅是电子政务的智能化，还包括智能服务和开放数据，三者缺一不可。开放数据能够让政府以较小的投入延展便民服务的深度和广度，从而撬动庞大的大数据产业链条，带动整个产业的兴旺。因此，智慧政府的建设不仅需要各级政府开放思想、开放数据，还需要引入设计合理的市场机制，才能极大地加快智慧政府的建设。我们大胆预测智慧政府的发展趋势如下。

1. 用 API 整合智能政府大数据

政府数据资源的整合是智慧政府建设的前提和基础，这一点毋庸置疑。但是不可忽视的是，当越来越多的政府部门意识到“数据是一种财富”的时候，反而阻碍了内部数据资源的整合。那么如何跨越政府数据资源整合的障碍？引入市场机制不失为一种解决方案。

政府数据资源可以用 API 接入的方式实现形式上的数据资源整合，而数据集的管理和维护仍旧由现有部门实施。因为政府数据的多样性和复杂性，实现政府数据的统一管理和维护并不一定经济可行。另外，根据 API 的调用实时计费，从而促进政府部门之间的数据交易。同时各个政府部门的 API 接口还可以有选择地面向公众和企业开放。人们可以像使用水、电、天然气一样，按照自身对数据的需求和使用量支付费用，也能够使政府更多、更好、更快地开放数据。

2. 政府开放数据第三方门户网站将有力助推智能政府建设

事实上，国内政府的数据开放并不像排名表现得那样糟糕。政府正在不断地开放数据，但是这些数据处于零散的碎片化状态，其可用性并不令人满意。政府开放数据第三方门户的兴起，将有力地解决这一问题。政府开放数据第三方门户将分散的政府开放数据加以搜集、精炼并归类，形成可用的数据集，供公众和企业使用。

第三方开放数据门户的意义在于：一方面，这些开放的、精炼的、可用的数据集可以引导公众、企业积极地利用政府开放数据参与智能政府建设，从而让智能政府的建设进度得以大大地加快；另一方面，第三方开放数据门户能够以更加优异的服务带动整个大数据产业链的发展和成熟。

2.8 农业大数据

2.8.1 农业大数据应用现状

农业大数据作为大数据的重要分支，是大数据理论、技术、方法在农业领域中的专业化实践和应用。

农业涉及农资、育种、耕地、播种、灌溉、施肥、防治病虫害、收获、仓储、农产品加工、农产品物流、销售、畜牧业生产管理等内容，贯穿整个农业生产管理、消费过程中的各个环节都会产生大量的数据。农业大数据不仅涉及生产全过程中各个环节产生的数据，还涉及跨行业、跨专业、跨领域的数据。

农业大数据应用依托部署在农业生产现场的各种传感节点（环境温湿度、土壤水分、二氧化碳、图像等）和无线通信网络，完成农业大数据采集、传输、存储、处理等环节的数据管理，结合大数据分析挖掘技术，最终实现农业生产环境的智能感知、智能预警、智能决策、智能分析、专家在线指导，为农业生产提供精准化种植、可视化管理、智能化决策。如今，国内外农业大数据快速发展，其典型应用体现在以下方面。

培育良种。对人类营养状况数据、生物群体的基因组等数据进行分析，通过对农作物的基因组进行测序，培育一些营养价值较高的作物品种，有助于提升人们的健康水平。

精准种植。精准种植是基于 3S 技术（遥感技术、地理信息系统、全球定位系统）实施一整套现代化农事操作技术与管理的系统，主要用于土壤肥力的精准化监测、农田边界图智能管理、病虫害精准定位和防治、精准施肥和灌溉等。

农业生态环境监测。生态环境监测主要是对农作物生长相关的土壤、水质、气候、气象和灾害等情况进行全面监测，并对它们之间的复杂关系进行分析，可以判断不同生态环境对农作物生长的影响。

天气预测。通过分析历史天气变化规律，建立天气识别模型，结合当前气候特征和近期天气情况，对某地未来一定时期的天气状况进行预测分析，对农业生产和日常生活具有重要指导作用。

农产品与食品安全监测。通过对农产品与食品的产地环境、产业链管理、产前产中产后、仓储加工、物流等数据进行监测，并通过对影响农产品与食品安全的关键性指标设置警告、分析数据、发布预警、寻找警源、消除警情等一系列操作，实现对农产品与食品的安全监控。

农产品物流。农产品物流涉及农产品的收购、储存、加工、包装、运输、卸载、搬运、配送等环节，通过整合、分析各个环节的数据，不仅可以连接农业主体和消费需求主体，还能实现农产品保值增值，甚至可以为整个物流管理提供有力的决策支持，如物流中心选址、最优化配送路线、合理管理库存等。

农产品市场追踪。通过对农产品销售价格、销售量、销售需求、消费者购买行为数据进行分析，可以判断农产品的供需、价格变动以及消费者的购买习惯等。

我国的农业大数据应用虽已取得阶段性成果，但在整个农业产业链推广过程中仍存在许多问题。而农业大数据作为农业信息化的发展趋势，是新一代信息技术的集中反映，是一个具有无穷潜力的新兴科技产业方向。

2.8.2　农业监控预警

近年来，规模种植为提高人们的生活水平带来了极大的便利，得到了迅速的推广和应用。种植环境中的温度、湿度、光照度、CO_2 浓度等环境因子对作物的生产有很大的影响。然而，在大面积种植中，定期检查、灌溉、排水、施肥等工作没有一套指导标准，更多的是靠人为判断，判断的差异性严重影响种植产量和质量。针对上述问题，龙信思源（北京）科技有限公司提供的智能农业监控系统融入国际领先的“物联网 – 移动互联网 – 云计算”技术，借助个人电脑、智能手机等终端设备，实现对农业生产现场气象、土壤、水源环境的实时监测，并对大棚、温室、灌溉等农业设施实现远程自动化控制（见图 2-35）。同时，结合视频直播、智能预警等强大功能，帮助广大农业工作者及时掌握农作物生长状况及环境变化趋势，为用户提供高效便捷、功能强大的农业监控解决方案。

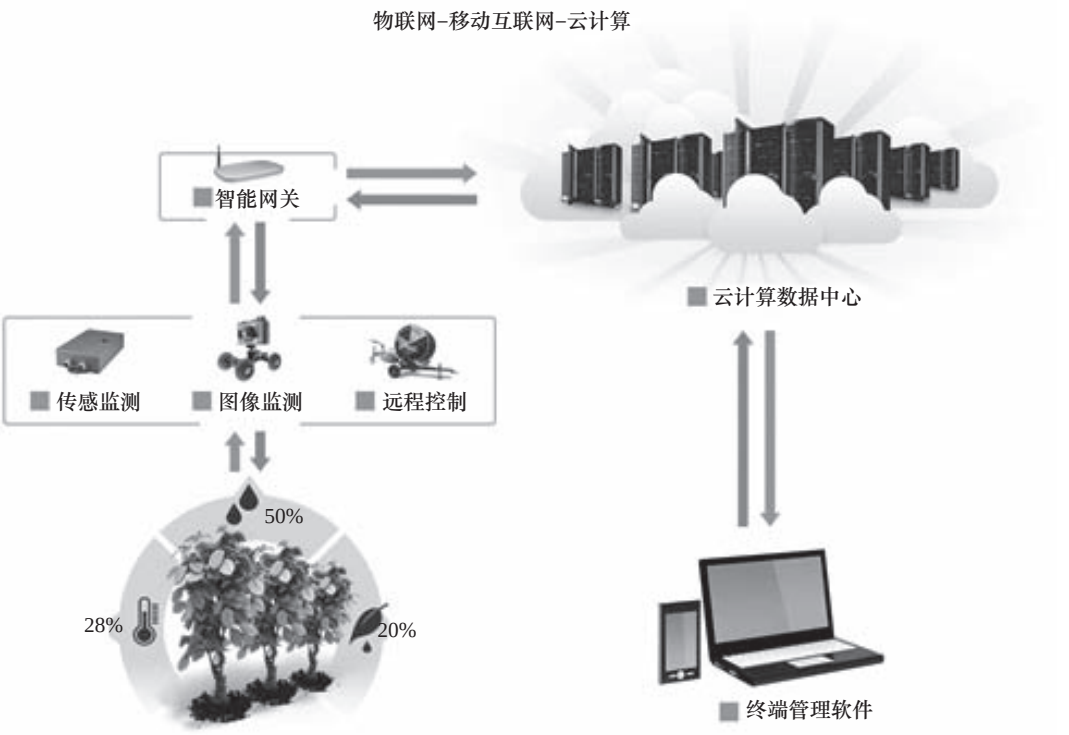


图 2-35　智能农业监控技术背景

该系统通过定制开发的智能终端设备监控农业生产过程中的各类指标（包括气象环境、土壤情况、设备状态等），通过高清摄像机或者照相机远程监控生产园区中的一系列智能终端设备（降温、加湿、抽风、施肥等），将数据汇聚到云计算数据中心，实现农业信息检测和标准化生产监控，帮助用户精确了解农作物生长、病虫害、土地灌溉及土壤空气变更情况等，并结合农产品的生产流程与标准指标设置预警反馈，最终实现全程监控和预警机制（见图 2-36）。

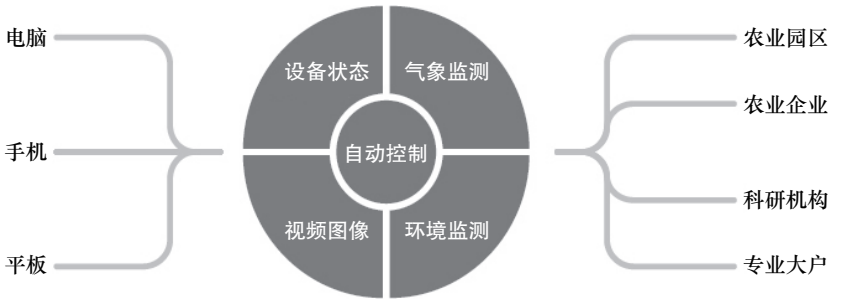


图 2-36　智能农业监控系统

气象监测。田间的“一站式”气象站采用太阳能供电，集成了多种传感器，实时监测各种气象信息（风向、风速、光照、温度、降雨量等），并通过智能网关直接将数据信息传回云数据中心。全程采用全智能化设计，一旦设定监控条件，可完全自动化运行，无需人工干预，最大程度避免人工操作的随意性，同时明显降低现场劳动力，帮助用户实现对农业设施的精准控制。气象监测示例如图 2-37 所示。

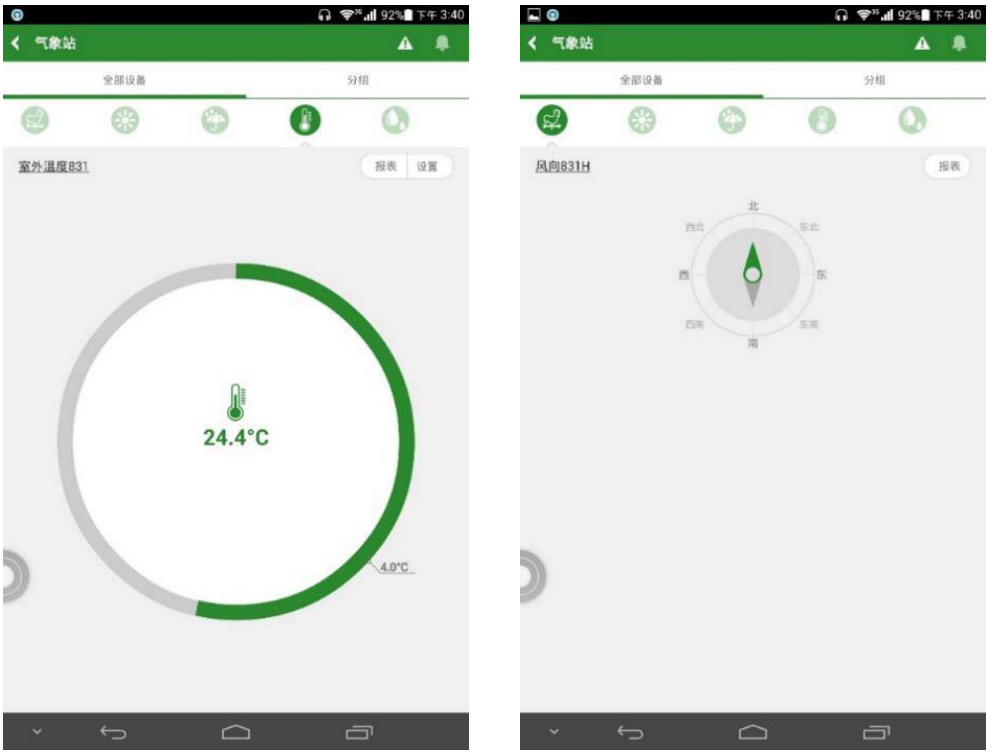


图 2-37　气象监测示例

环境监测。采用“一站式”监测站，实时监测各种环境信息（空气湿度、土壤湿度、CO_2 含量、土壤 pH 值等），并通过智能网关直接将数据信息传回云数据中心，提高监控效率，帮助用户实现生产流程的标准化管理。环境监测示例如图 2-38 所示。

视频图像。综合运用传感器、控制器、智能相机、智能摄像头、RFID 等高端物联网设备，实现对农业生产活动中从物到人的 360° 全面监控，监控范围包括：现场视频、高清图片、环境质量、设备状态、人员定位等，并根据设定条件对各种异常情况进行自动预警、任务跟踪与远程控制。视频图像示例如图 2-39 所示。

图 2-38　环境监测示例

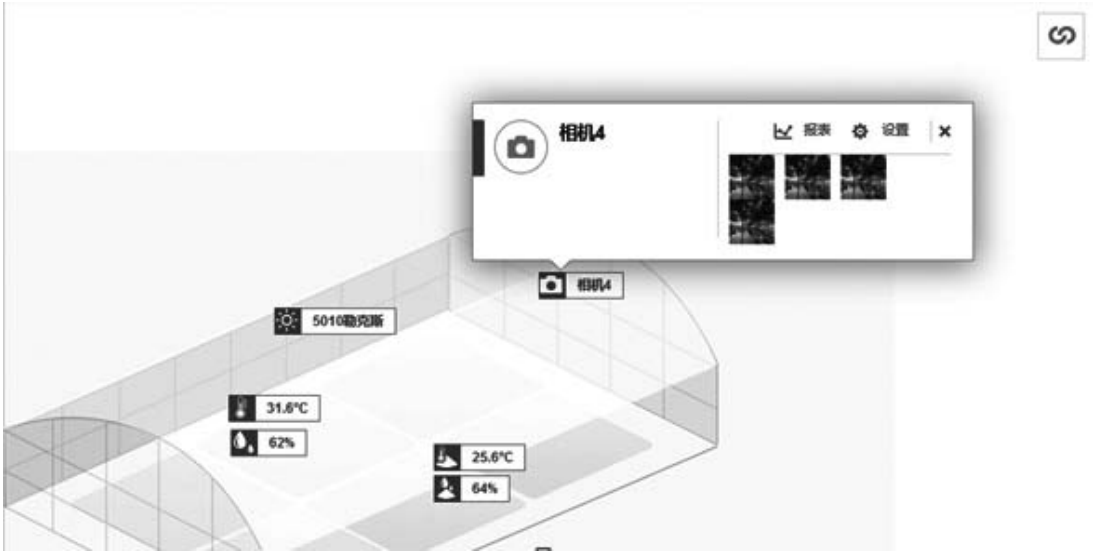


图 2-39　视频图像示例

设备状态。实时监测生产现场各种设备运行状态（灌溉记录、排风记录、流量、水压等），并通过智能网关直接将数据信息传回云数据中心，及时为用户反馈设备运行情况，优化农作物生长环境。设备状态监测示例如图 2-40 所示。

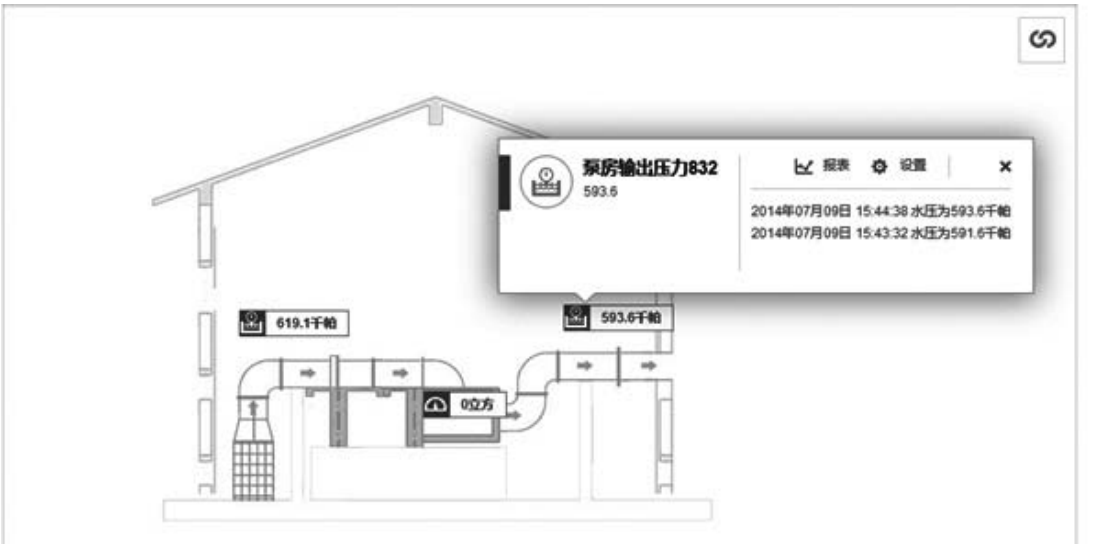


图 2-40　设备状态监测示例

自动控制。融入国际领先的“物联网 – 移动互联网 – 云计算”技术，借助个人电脑、智能手机等终端设备，通过对农业生产现场气象、土壤、水源环境的实时监测，实现对大棚、温室、灌溉等农业设施实现远程自动化控制。自动控制示例如图 2-41 所示。

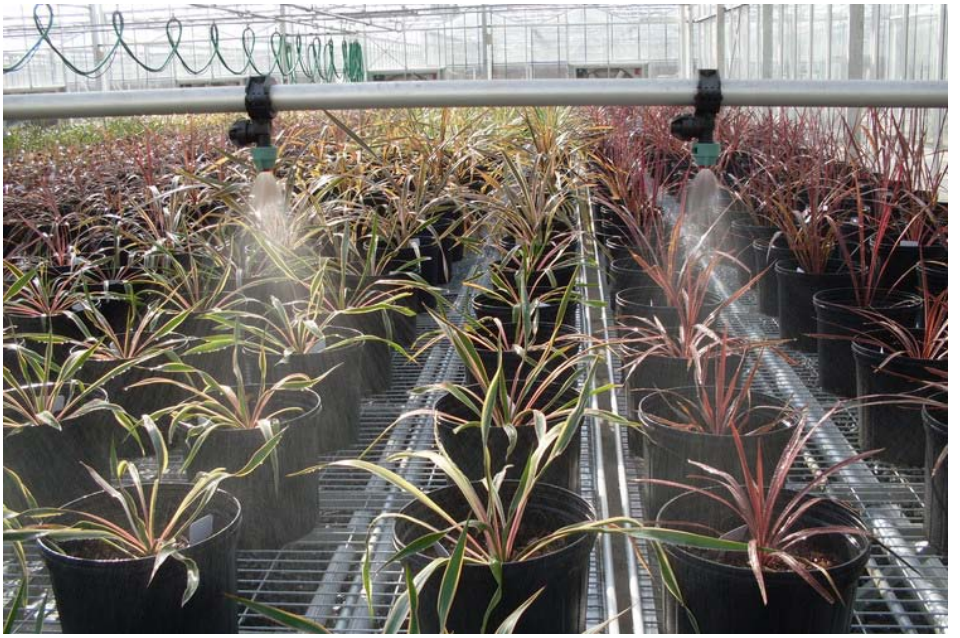


图 2-41　自动控制示例

总之，信息作为现代农业的基本要素，将很快向农业生产管理深层渗透，信息服务技术向着网络化、集成化、智能化、个性化、低成本化方向发展，海量信息低成本获取、传输、大数据建模技术的瓶颈突破以及智能农业监控系统的应用，将使智慧农业逐步走进农家，为中国实现农业现代化助力的同时，也为广大农产品消费者带来更高的效率和更多的便利。

2.8.3 农业精准种植

随着科技的进步和生产力的提高，我国农业有了较大发展，粮食产量也有了大幅提高，但人们在农业生产管理过程中仍存在诸多的问题：不合理使用化肥、大量喷洒农药、大水漫灌等，不仅对农作物生长不利，而且还会造成土壤板结、盐碱化程度加重等后果，进而影响后续的农业生产和收成，陷入恶性循环。目前，大数据处理技术以及农业生产过程中产生、积累的海量农业数据，为农业的发展带来了新的机遇。龙信思源（北京）科技有限公司作为协作单位参与了科技部“渤海粮仓增产增效”项目并承担了其中部分大数据支撑研究，下面以此研究的相关成果为内容进行阐述。

1. 数据现状与应用需求分析

目前，团队采集的农业数据主要包括土壤数据、作物（小麦）生长数据、气候数据、种植与生产数据、病虫害防治数据、气象数据等，数据的采集主要采用两种方式：人工采集和自动采集。人工采集主要是当地的科技人员按指定的采样标准并通过专门的数据采集网站对这些数据进行记录，如图 2-42、图 2-43 所示；自动采集主要是利用农田信息采集物联网设备对数据进行实时采集，如图 2-44 所示。

图 2-42 数据（人工）采集系统主菜单

基于对农田作物长势及生长环境数据的采集和积累，依托于大数据分析和挖掘技术，分析采集点的土壤养分含量、播种品种、播种日期、上茬农作物的种类及产量情况对小麦生长的影响，通过构建大田作物（小麦）生长因素分析模型，找出影响大田作物（小麦）生长的关键性指标，为改良作物生长提供决策依据及改良方向，进而为农作物的增产增效服务。

图 2-43　数据的人工采集

图 2-44　农田信息物联网自动采集系统

2. 分析方法与过程

通过对数据和业务问题的理解，由于作物（小麦）品种多样、长势不一，如果不对麦苗的长势情况进行分类划分，会使得数据分析或挖掘的难度很大、效果较差，因此在对这些数据进行分析或挖掘前，需要对数据进行预处理，根据麦田划分标准并结合麦苗长势数据，将麦田划分为三个等级：一类麦田、二类麦田、三类麦田。

从宏观和微观层面对影响麦苗生长的指标进行统计分析，并采用有效的可视化方式

对统计分析结果进行展现，其关键在于从单指标角度分析每个指标与麦苗长势之间的关系，如图 2-45 ~ 图 2-48 所示。

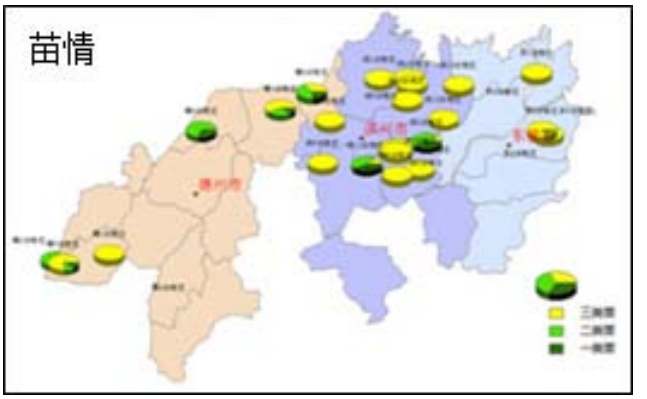


图 2-45　样本点苗情布局图

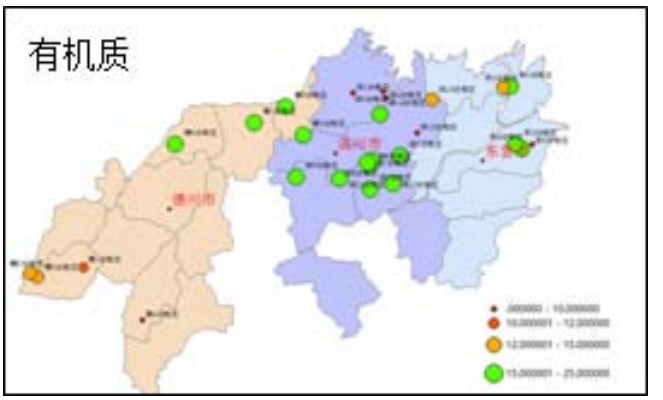


图 2-46　样本点有机质（宏观）布局图

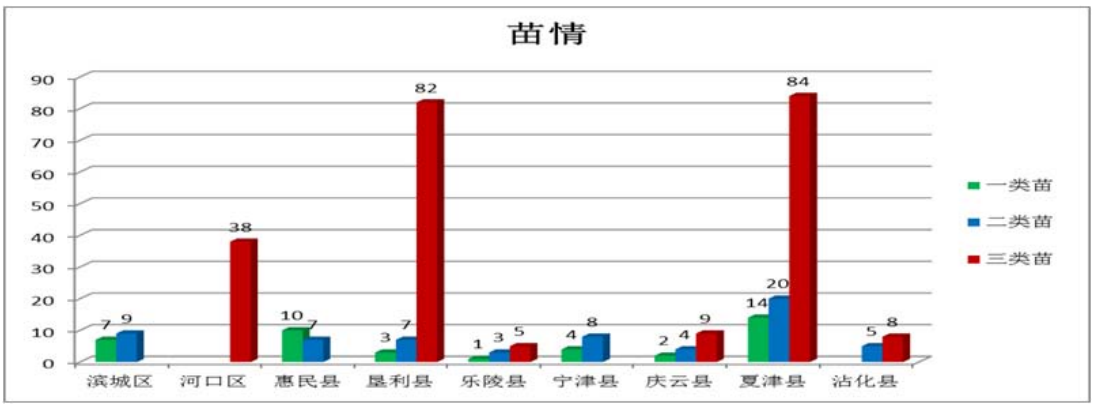


图 2-47　样本点苗情统计分析图

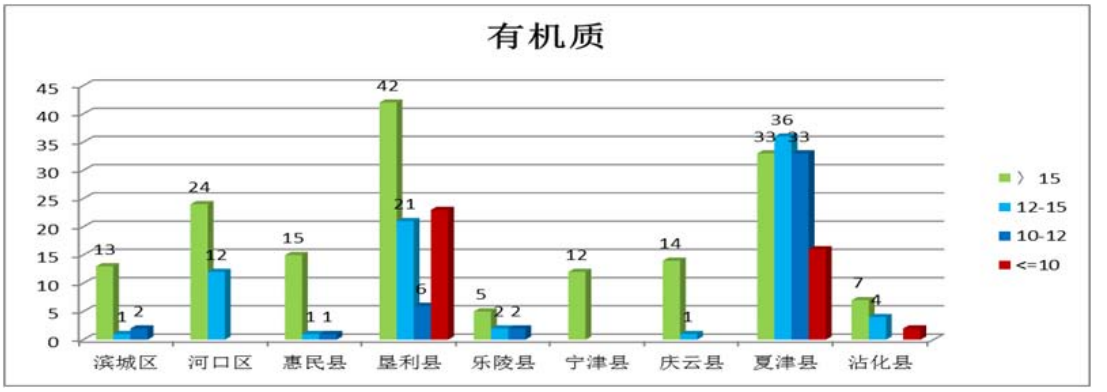


图 2-48　样本点有机质（微观）统计分析图

在对影响麦苗生产的各个指标进行统计分析的基础上，结合数据预处理、数据挖掘等技术，采用有效的挖掘算法——决策树 C5.0 和 Logistic 回归分析（目标变量即麦田等级为离散的分类变量，对麦苗影响因素的分析问题转化为对麦田等级划分的分类问题），对影响麦苗生长的各个指标进行动态的、关联性的深层次分析和挖掘，通过构建大田作物（小麦）生长因素分析模型，分析出各种组合模式（作物品种、土壤养分含量、种植模式、施肥量、灌溉量）下对麦田等级划分的有效性，从而找出影响麦苗生长的关键性指标。模型构建流程如图 2-49 所示（模型构建使用 IBM SPSS Modeler 数据挖掘软件）。

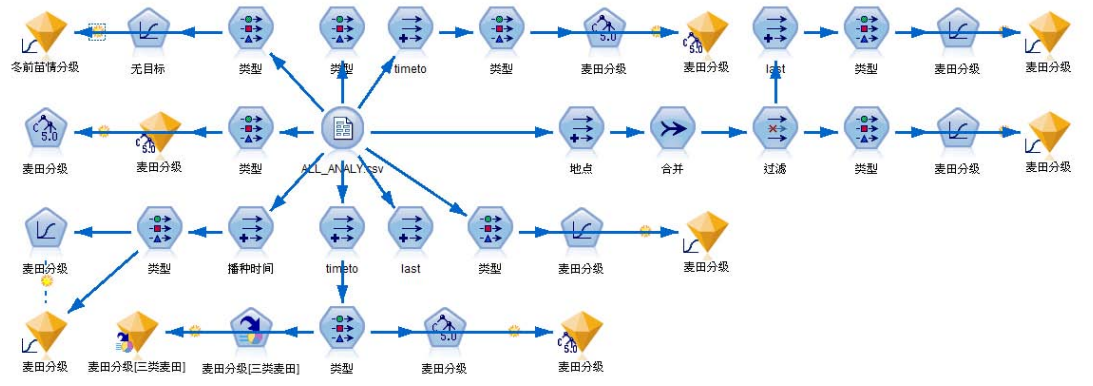


图 2-49　大田作物（小麦）生长因素分析建模流

模型构建完成后，需要对建模结果进行分析和评估，确定一系列影响作物生长的分析指标体系（组合模式），对训练结果和验证结果进行比较，最终找出影响作物生长的关

键性指标。

3. 应用成效

通过构建大数据采集系统，对农作物苗情、土壤养分、含水量等数据进行收集，可有效解决田间数据的实时采集问题。通过宏观和微观分析，可以对影响作物生长的单指标（土壤养分、含水量、盐碱程度等）进行全面透析，有针对性地对土样代表的地块进行施肥、灌溉、控盐渍化等综合指导，同时可以根据苗情布局情况，对不同等级的麦田进行分类指导，如因地因苗制宜、科学运筹肥水、抗旱保苗、促弱转壮、控盐保墒、构建合理群体、加快转化升级，最终奠定丰收基础。

通过构建大田作物（小麦）生长因素分析模型，初步发现对试验对象麦田等级有重要影响的指标，在农作物的生产管理过程中优先关注并改良这些指标，并在示范工程中取得良好的成效，为农作物的增产增效提供了有力的数据支撑。

2.8.4 农业大数据发展趋势

大数据时代，农业大数据不仅充满了未知的挑战，也让我们充满了更多期待和憧憬。在新一轮农业现代化建设中，要将农业大数据纳入国家农业信息化发展战略，夯实智慧农业的基石，让大数据创造出真正的智慧，支撑农业大数据的稳健发展。要密切跟踪国际大数据前沿技术，积极抓住发展契机，基于政府强有力的推动和引导，做好顶层设计，实现有序发展。围绕国家农业特点和重大需求，梳理农业大数据重点发展领域，创新农业大数据关键技术，重点培养和支持一批农业大数据的应用与示范项目。尤其是要通过大数据与农业的融合，不断加强基于农业物联网成果的示范应用，推进现代农业跨越式发展。

2.9 地理信息大数据

2.9.1 地理信息产业大数据应用现状

地理信息产业是 21 世纪的朝阳产业，属于高新技术产业，是以现代测绘技术、信息技术为基础，结合计算机技术、通信技术和网络技术形成的综合性产业。具体包括空间地理科学技术、信息科学技术、航天航空科学技术，是遥感技术、地理信息系统、卫星定位系统与测量技术应用的产业化结果，是国民经济信息的重要组成部分。2014 年 1 月 22 日，国务院办公厅以国办发（2014）2 号印发了《关于促进地理信息产业发

展的意见》，这一意见的提出，预示着地理信息产业在新一轮的行业整合方面将面临重大机遇。同年 7 月，国家正式印发《国家地理信息产业发展规划（2014 ~ 2020）》，对于推进我国地理信息产业的蓬勃发展具有重要指导意义。该规划强调到 2020 年，中国地理信息有关政策法规体系基本建立，结构优化、布局合理、特色鲜明、竞争有序的产业发展格局初步形成。科技创新能力显著增强，涌现出具有核心竞争力的龙头企业和较好成长性的创新型中小企业，并拥有一批具有国际影响力的自主知名品牌。产业保持年均 20% 以上的增长速度，2020 年总产值超过 8000 亿元，成为国民经济发展新的增长点。

我国地理信息产业发展经历了一个比较曲折的发展过程。在 20 世纪 90 年代之前，我国测绘和地理信息的发展注重于为政府管理和国家安全提供服务，市场在测绘地理信息资源配置中没能发挥基础性作用。因此，地理信息产业发展受到限制。上世纪末，我国社会主义市场经济体制不断完善，卫星遥感、卫星定位和地理信息系统等技术应用不断业务化，在国民经济和社会发展的各领域中不断深入，并加速向人民群众的日常生活渗透。地理信息产业快速发展的条件基本具备。在这一背景下，国家测绘局开展了第一次测绘发展战略研究，正式提出发展地理信息产业。此后相继开展了促进地理信息产业发展的政策研究，形成了一系列促进我国地理信息产业发展的具体思路和措施。在国家测绘局和有关部门的大力支持下，地理信息产业获得快速发展。据行业协会不完全统计，截至 2013 年底，企业达 2 万多家，从业人员超过 40 万人，年产值近 2600 亿元。2006 年以来年均增长速度超过 25%，形成了一批具有一定市场竞争能力的地理信息硬件、软件和数据产品。地理信息在国土资源、环保、电力、公安、交通、水利、铁道、民政、农业、林业、公共应急等领域得到广泛应用，并和汽车、手机及其他多种导航终端结合，将应用和服务延伸到了社会大众衣食住行的各个方面。地理信息技术不仅已成功应用于水利环保、能源矿产、气象环保、国土房产等行业中，而且成为国家数字城市与智慧城市建设的核心平台。目前，我国各行业对地理信息的需求旺盛，地理信息企业不断壮大，地理信息产业园区开始出现，已有八家企业在国内外上市，一些企业开始并购重组迅速发展，并“走出去”参与国际市场竞争。“十二五”期间，我国工业化、信息化、城镇化、市场化、国际化深入发展，推进经济发展方式转变，实现经济社会科学发展成为我国“十二五”的主要任务，对地理信息服务提出了新的更高的要求。云计算、物联网等高新技术应用不断深化，其与现代地

理信息技术的融合日趋紧密，推动不断产生新的地理信息服务。地理信息产业发展也进入了全面构建数字中国的关键期。

地理科学面对的是一个复杂的系统，是天然的大数据。1986 年钱学森院士在现代人类知识体系中将地理科学归结为自然科学与社会科学之间的桥梁科学，研究整个地球表面同人类息息相关的大气对流层、岩石圈上部、水圈、生物圈和人类圈环境。所以上至卫星遥感数据，下至地震传感数据以及我们常见的统计、环境、水利、资源、土地等领域的数据都属于地理数据，所以地理信息技术需要处理的范围广、数据源多、数据类型多样，其数据量巨大是不言而喻的。

地球表面的信息量巨大，感知手段多样。目前的遥感数据处理能力已经不可想象，还要考虑多波段、多时相、多产品、历史数据、中间数据、重叠区、雷达、点云数据等问题。遥感数据之外，北斗定位系统的建立、移动互联网和物联网的快速发展也会导致包括来自车辆、风力、雨量、温度、湿度等各种传感器以及个人网络活动的高频空间关联信息的数据洪流涌入，并且要求快速处理响应。

地理信息数据涉及领域广泛，多渠道的数据积累形成海量数据。数据总量较大甚至巨大，通常以 GB、TB 或 PB 为基础单位。数据类型繁多，包含结构化、半结构化甚至非结构化数据，且非结构化数据所占份额越来越大。数据产生速度飞快，主要基于手持移动终端、互联网、物联网、车联网等平台产生。数据是潜在待挖掘的资源，基于“价值密度与数据总量呈反比”规律，大数据的价值密度较低。

由此可见，不论是大数据还是地理信息行业，无论是技术研究还是应用，都紧紧围绕“地理信息数据”为中心，确立其核心地位。

1. 地理信息资源为大数据提供重要依据

地理信息资源涉及各行各业，有多种类型的空间数据。主要包含地图（矢量和栅格）数据、影像数据、派生数据、公众数据、开源数据、标签数据、元数据、分类和未分类数据、视频和传感器数据等。地理信息数据的获取途径主要包含以下几种：

人工获取资源。外业采集人员通过野外测量、以手工绘图的形式记录信息。

机器获取资源。通过卫星、航空大飞机、无人机对地观测获取低、中、高精度影像资源信息。可以通过激光雷达扫描、光谱成像、视频摄像、传感器感知等手段获取丰富的数据源。

人机交互合作获取。通过网络手段上传的文字、图片、视频等数据，通过地理名词

语义搜索与提取、个人终端定位数据收集、个人导航数据和车联网数据累计，以及各种方式引起的数据库更新。

其他方式获取的数据。在其他领域、利用其他的方式获取的数据信息。

2. 空间数据挖掘是大数据应用的组成部分

数据挖掘通过分析从大量数据中寻找规律，这种规律将形成新的知识，发挥其应有的价值。传统的空间数据分析方法主要通过基于空间数据的分析运算以及基于空间数据和非空间数据的联合运算等。后来逐渐发展，包括对地理空间本身特性的深层次分析、对空间决策过程的模拟、对复杂多因子空间系统时空演化的模拟和预测等。例如在城市规划方面，通过对地市基本地物、气候环境等自然数据和经济、人口等社会人文信息的联合挖掘，可以对城市规划提供强有力的支撑，为城市的智能管理与科学发展给出合理关联意义，预测数据相关对象的发展趋势，这与大数据的应用方向不谋而合。空间数据挖掘是凸显大数据价值、盘活大数据资产和有效利用大数据的重要思想和技术。从数据中提取信息，从信息中挖掘知识，在知识中萃取数据智能，提高自学习、自反馈和自适应的能力，实现人机智慧。这也正是大数据应用的根本内生动力，即“数据即资源”[㊀]。

大数据下的地理信息已经得到了广泛应用，服务于我们的生活、工作，并带来便利。电子地图、卫星导航、遥感影像，这些地理信息产业链上的新生事物正在创造奇迹，效益已经显现。

2.9.2　大数据在智慧环保中的应用

自从 2008 年 11 月 IBM 提出“智慧地球”的概念后，国内外智慧城市的建设风起云涌，与此同时智慧环保、智慧交通、智慧医疗、智慧物流、智慧交通等各细分行业也开始了相应研究，将“智慧”技术应用于各个领域。

智慧环保是结合物联网技术、超级计算机和云计算技术，对大气、水体水源、噪声、放射源、废弃物等进行感知、处置和管理，建设成为一个集智能感知能力、智能处理能力和综合管理能力于一体的新一代网络化智能环保体系，达到“测得准、传得快、算得清、管得好”的总体目标，旨在推进污染减排，加强环境保护，实现环境与人、经济乃至整个社会和谐发展。

㊀ 引自：周顺平，徐枫，大数据环境下地理信息产业发展的几点思考，P208, 1672-1586 (2014) 01-0045-06, 2013. 8。

智慧环保的总体架构包括：感知层、传输层、智慧层和服务层（见图 2-50）。实现智慧环保需要以下几项关键支持技术：

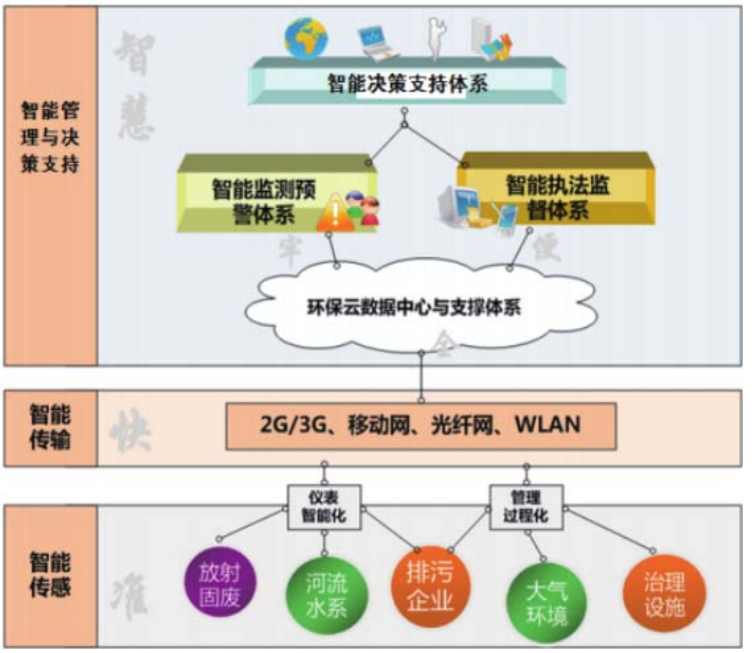


图 2-50　智慧环保体系架构

物联网技术。环保物联网技术是指通过各种传感设备（传感器、射频设备技术、全球定位系统、红外感应器、激光扫描等）采集声、光、热、电、力、化学、生物、位置等各种信息并与互联网、无线专网进行交互传输信息的一个巨大网络。

云计算技术。云计算技术是网络计算、分布式计算、并行计算、效用计算、网络存储、虚拟化、负载均衡等传统计算机技术和网络技术发展融合的产物。

智能 GIS 技术。采用了多维 GIS 融合技术，将“时间维、空间维和仿真技术”相结合的三维 GIS 平台，真正实现“物联网前端感知、应用时态分析、管理虚拟仿真、多维 GIS 空间分析”一体化的 GIS 可视化应用创新模式，将三维 GIS 的发展带入了多维 GIS 时代。

“天地空”一体化遥感监测技术。地面遥感车、气球、飞艇、火箭、人造卫星、航天飞机和太空观测站等多个观测地球平台相互配合使用，搭载各种用途的传感器，实现对全球陆地、大气、海洋等进行立体、实时观测和动态监测，是未来获取地球表面和深部时空信息的重要手段，也是智慧环保获取基础数据的重要来源。天空地一体化遥感监测技术是智慧环保实现的重要基础支撑技术之一。

海量数据挖掘技术。海量数据的搜集、强大的多处理器计算机、数据挖掘算法作为支持数据挖掘的基础技术已逐渐发展成熟。目前常用的数据挖掘方法包括神经网络法、遗传算法、决策树方法、粗集方法、覆盖正例排斥反例方法和模糊集方法等。

环境模型模拟技术。地理信息系统与环境模型进行集成应用为环境决策提供技术支持已经成为环境保护决策的重要发展趋势。环境模型模拟技术的最终目的是要还原一个实际系统的行为特征，模拟其物理原型的数学模型。

大数据和环保结合解决数据存储、管理、加工处理、共享、分析过程中存在的一系列问题，有效提高环境信息化水平，通过大数据分析，人们可预测和洞察未来会发生的事情：

帮环保部门更好地预测未来走向。大数据通过挖掘分析，科学地建立评估和预测预报模型，让环保管理者可预测未来发展趋势，降低环境管理风险，能够在管理调整尚未展开之前给出相关的确定性答案，让管理措施做到有的放矢。

帮环保部门更好地管理污染源企业。通过大数据技术，可以实现污染源企业的精准锁定，在污染源的生命周期过程中，对于每个节点所需要的每一类数据的挖掘方式、具体情况等，都可以进行搜集分析，形成基于污染源管理的数据资源分布可视图（见图 2-51）。

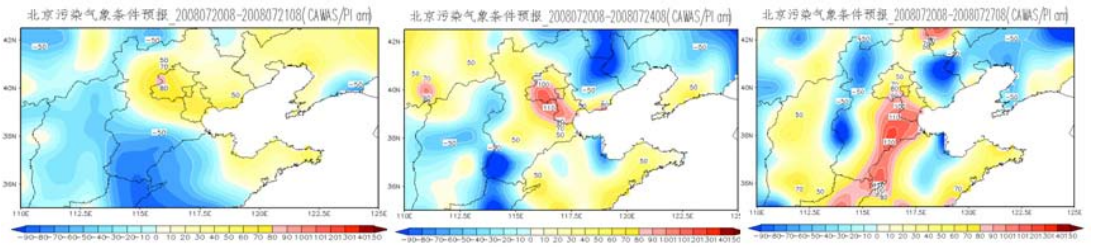


图 2-51　污染气象条件指数预报

帮环保部门更好地做好公众服务。社交信息数据、公众互动数据等帮助环保部门进行公众服务的水平化设计和碎片化扩散。经济学家 Richard H.Thaler 曾经提出一种观点：“个人观点的微小变化都可以演变为所有人的群体行为模式的重大变革。”环保部门可以借助社交媒体中公开的海量数据，通过大数据信息交叉验证技术分析数据内容之间的关联度等，进而面向社会化用户开展精细化服务，提供更多便利、产生更大价值。

2013 年以来，中国雾霾天数创 52 年之最，PM2.5 成为最热门的环境名词。大气污染、固体垃圾排放和水污染已经让我们的天空不再蓝、空气不再清新、河流不再清澈，过去唯“经济发展”的管理习惯与环境保护意识的淡薄已经让现在的我们尝到了恶果，过度的污染和环境破坏已成为影响居民健康安全的重要因素，所以推动环境保护相关法

规政策出台，加强对于资源节约、废物减排和治理污染技术的研发和执行已经迫在眉睫。借助大数据采集技术，将收集到大量关于各项环境质量指标的信息，将其传输到中心数据库进行数据分析，直接指导下一步环境治理方案的制定，并实时监测环境治理效果，动态更新治理方案。通过数据开放，将实用的环境治理数据和案例以极富创意的方式传播给公众，通过一种鼓励社会参与的模式提升环境保护的效果与效率。

在这场与雾霾的战争中，利用遥感的大尺度连续监测 / 导航的精准定位及地理信息的分析溯源和展示，可以追溯雾霾的发生轨迹、确定雾霾的重要污染源分布并预测雾霾的演变趋势，实现雾霾的监控与预报，从而最大程度地改善雾霾状况，缩短打赢治霾战争的周期及降低反复发生的几率。

公众还可以通过移动端上传周围环境数据，配合 GPS、无人机、卫星、雷达、激光扫描仪、气象监测站与环境监测站生成大量的专业数据，整合这些碎片化的信息，基于数据分析将产生更大价值的数据服务，反过来能为用户提供更好的服务。

1. 环境大数据的布局

布局环境大数据要整合数据获取、数据分析及数据应用所关联的一系列环节。针对环境大数据布局提出的实施方案如图 2-52 所示。

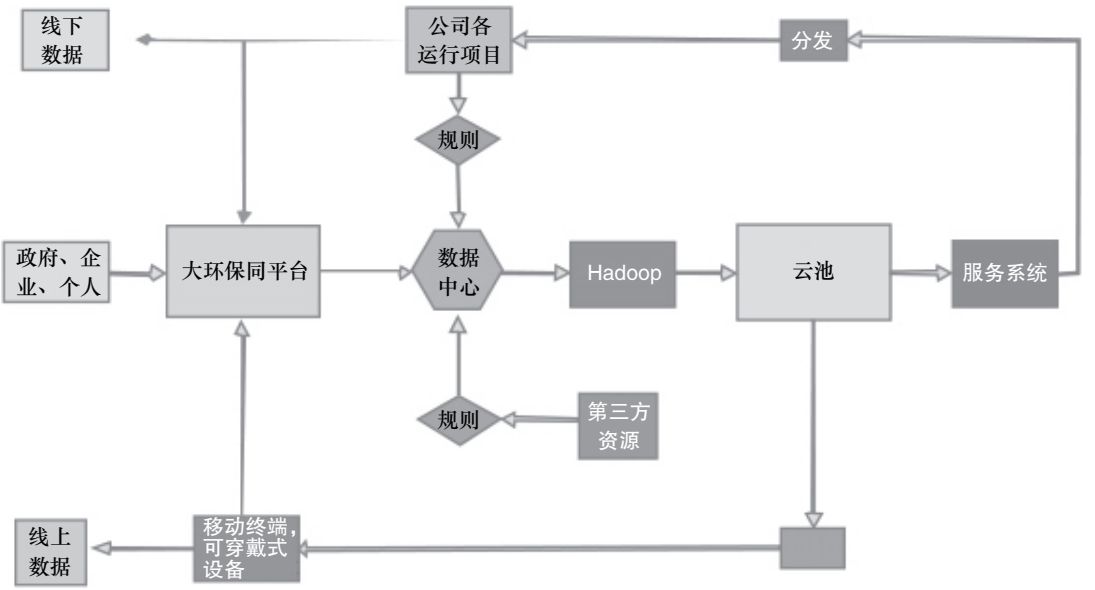


图 2-52 环境大数据布局

大环保网作为环境数据聚集平台，结合可穿戴设备、移动端入口、政企项目以及广

泛合作的第三方内容提供商，共同构建环境大数据生态系统。大数据为平台服务提供信息支持，而服务的落地也有利于有效信息的不断采集，形成数据循环。

线上数据。手机内置传感器、可穿戴设备的发展使得环境检测可以不完全依靠专业而笨重的仪器。例如，手机麦克风测噪声，智能口罩内置离子感应器测 PM2.5。通过 UGC 或众包模式，利用手机及可穿戴设备采集环境数据。

线下数据。项目运行产生的数据，即环保部门、企业项目在日常运营过程中生成的大量环境监测数据。整合这些项目的监测数据获取渠道，不但能获得大量有价值数据，而且经过对聚集的数据进行分析产生有价值的服务数据，也能促进项目的开发建设。

第三方资源包括内容服务商提供的天气数据，以及研究机构、科研院所提出的新的理论技术方法。

2. 环境大数据的获取及应用场景

目前在空气质量预报中，污染源是最基础和最重要的信息。但是调查污染源是十分耗时耗力的工作，环境大数据的构建能较好地解决这一问题。

- 政府用户在大环保网上公布的重点监控企业排放信息。
- 移动端用户上传的小规模污染源。
- 遥感应用中心提供的在秸秆焚烧、沙尘季节的卫星遥感监测数据（见图 2-53 和图 2-54）。

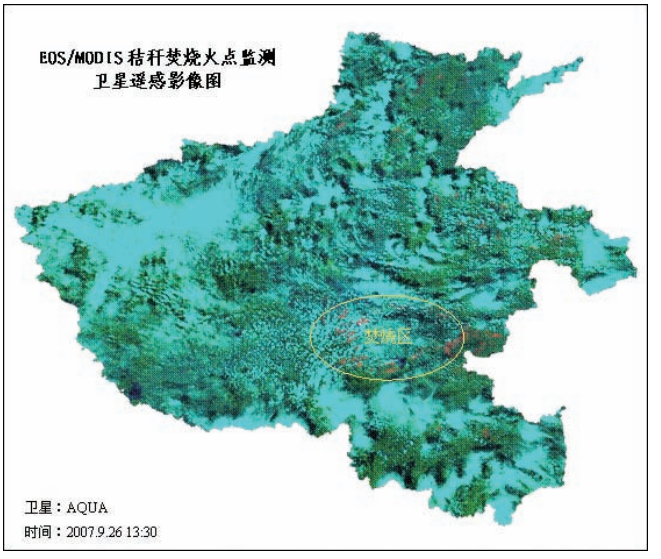


图 2-53　秸秆焚烧火点监测

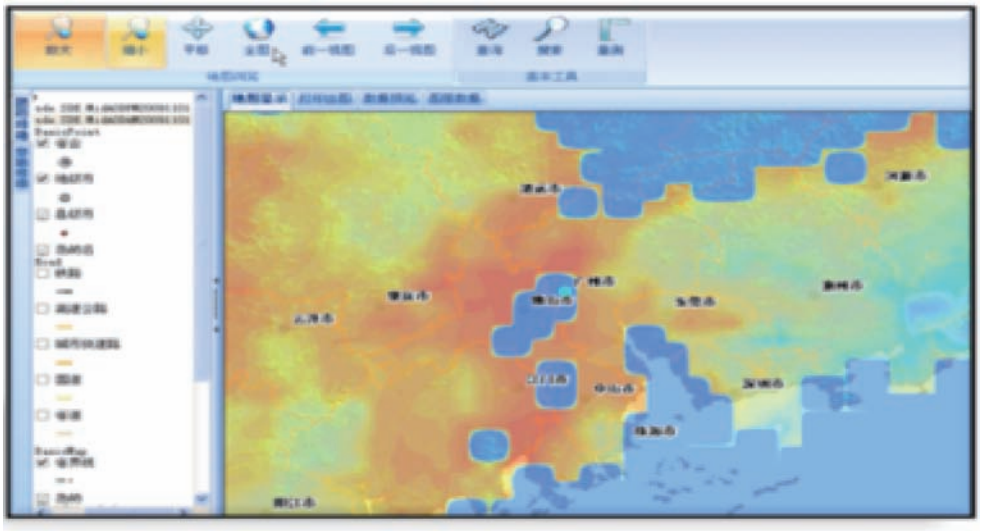


图 2-54　空气质量卫星遥感监测系统

数据中心获取这些数据后，经过模型计算，为用户发布更准确的空气质量预报。

3. 环境大数据建设

依据各类数据标准与规范，通过数据交换、整合、导入、录入等手段，全方位收集环境监测、监督、管理中使用的各种基础类、背景类、业务类的数据资料，并对这些资料进行规范化、标准化处理加工，形成体系完整、时间跨度长、专业覆盖全面、科学系统的环境信息资源体系，为环境监管、治理、规划、决策等提供最强大的数据服务和信息共享支撑。建设流程大体分五个步骤：一是环境资源信息标准规范建设，包括环境监测信息资源共建共享管理办法、环境监测信息资源共建共享技术规范、污染源统一编码规则、环境质量基础信息编码规则与编码方法、元数据库数据字典、环境数据中心基础数据库数据字典、环境数据转换与清洗规则以及其他相关的管理办法和技术规范等。二是环境资源目录建设，清点梳理数据业务过程中涉及的环境信息，进行科学编码和分类分级，划分资源责任单位，建立资源与业务的有机关联。三是环境基础数据平台建设，设计创建环境基础数据平台，收集整理环境元数据、标准基础数据并入库。四是交换共享平台建设，建设环境信息传输、交换、处理、共享和监管的统一平台。五是环境资源信息应用建设，基于整合处理后的环境数据进行分析和展示，为环境管理提供科学的辅助决策。环保大数据建设体系如图 2-55 所示。

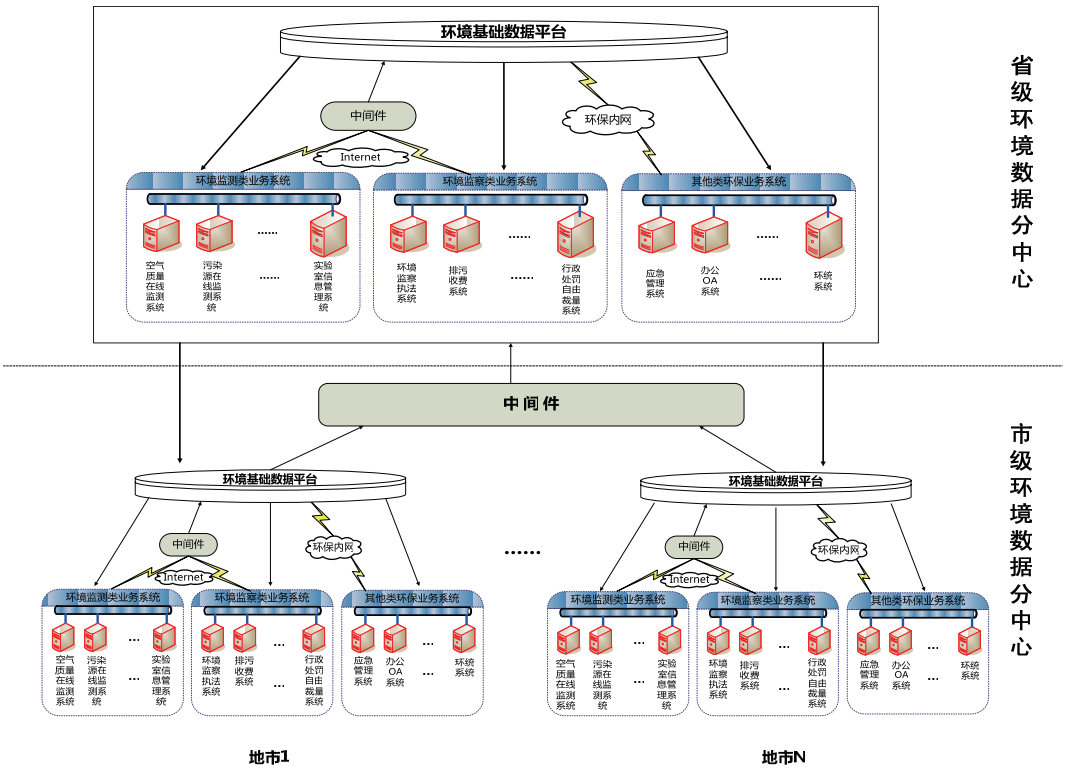


图 2-55　环保大数据建设体系（省级）

伴随着大数据时代的到来，人们的需求逐渐从数据存储、数据处理向数据应用和数据运维服务过渡，同时传统环保行业的数据处理模式已然不适应新一代数据中心的发展需要。而随着大数据技术的不断发展，一种全新的环保行业数据解决方案成为可能。大数据技术逐渐趋向成熟，一旦完成数据的整合和监管，大数据爆发的时代即将到来。我们现在要做的就是选好方向，为迎接大数据的到来提前做好准备。

2.9.3　大数据在互联网地图中的应用

1. 地图与大数据的结合发展

现在人们的各种行为，85% 都与空间地理位置有关，地图与人们的生活息息相关。电子地图种类繁多，如互联网地图、导航电子地图、街景地图、地形图、栅格地形图、遥感影像图、高程模型图及各种专题图等。

现代地图已经不仅是一种简单的信息传递，还是信息的增值，增值的数据应用于各

行各业，服务于公众。地图大数据可以从现代地图中收集与整理信息，结合数据分析与挖掘技术，建立数据预测模型，建立基于 LBS（位置服务）的系统平台，应用于不同行业，如移动、电信、联通、电力、环保、公安、城管、石油石化、银行、林业、水利、旅游、交通烟草物流等行业。

2. 互联网地图大数据助力传统行为方式

（1）求婚方式改变

2014 年 11 月 25 日，腾讯发布的一条新闻，日本东京的高桥靖花费半年时间，主要采用徒步，偶尔使用自行车、汽车、轮渡等出行方式游遍日本，用 GPS 记录下自己的全程足迹，完成了一副“ MARRY ME”的 GPS 图像和一幅一箭穿心的图注（见图 2-56），并把足迹记录上传到 Google 地图上，连点成线，用这幅 GPS 画作向女友求婚，成功抱得美人归，同时还获得了吉尼斯世界纪录最大 GPS 画作的“奖励”。

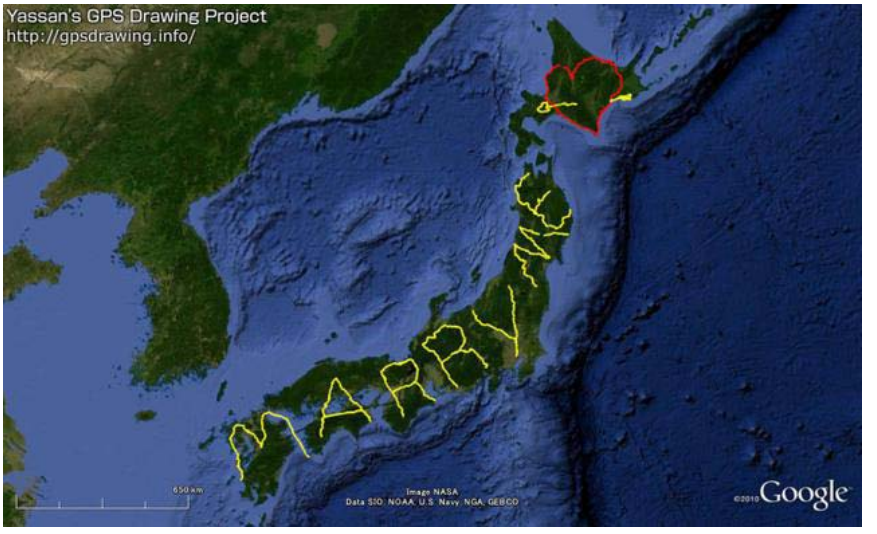


图 2-56 “MARRY ME”的 GPS 地图

这种地图规划路线，戴上 GPS 开始行走，把足迹记录上传到 Google 地图上，采用空间数据定位与集成管理，并通过 GIS、GPS、RS、物联网技术，将各类数据叠加到地图与地理空间位置上进行信息集成，最后连点成线显示出一幅图像，这就是地图大数据对传统行为方式的改变。

（2）出行方式改变

打车难一直是困扰广大北京市民出行的一大难题，一方面，乘客抱怨打车难，另一

方面，司机烦恼收入少。

“滴滴打车”和“快的打车”改变了传统打车方式，在移动互联网时代引领用户的现代化出行方式（见图 2-57）。2013 年，“滴滴打车”和“快的打车”与入口级应用运营商高德地图、百度地图达成合作，开启了与地图类应用合作联运的新模式，此类软件对乘客与出租车的位置定位都是基于电子地图，进行功能策划和平台应用开发，从而改变传统打车市场格局，颠覆路边拦车概念，利用移动互联网，将线上与线下相融合，从打车初始阶段到下车时线上支付车费，画出一个乘客与司机紧密相连的 O2O 完美闭环。

图 2-57　地图在打车软件中的应用

同时，人们在使用打车软件时，理论上会形成庞大的数据量，这些数据可以把每个人的地理坐标、工作地点、家住在哪儿、希望去哪儿都做出可靠的统计。因为，当人们在使用打车软件时，真实的地理信息都毫无保留的透露给打车软件。通过这些数据，人们的消费习惯、消费半径、消费能力将会清晰地展示在商家面前。如何获取数据、如何分析运用数据、如何在庞大的数据中挖出真金白银，将是未来商家研究的重点课题。如此看来，现在的出租车抢的是用户，而未来抢的是大数据。

探讨这些新地图应用，都是通过空间位置信息，结合人们的行为建立功能平台，发布和共享地图信息，改变人们的思维方式和行为方式。在互联网平台下发布与共享这些海量大数据，要求地图信息数据全、精、准、新。展望互联网电子地图发展趋势，未来的电子地图会结合公众行为大数据，利用数据分析与挖掘技术、云计算、物联网，让更多信息显示到电子地图上，从而有助于改变和影响政府决策和方针，改变和影响企业的

经营决策和经营方向，改变和影响人们的行为习惯等方方面面。

2.9.4 地理信息产业大数据发展趋势

大数据时代下，地理信息产业挑战与机遇并存。不仅需要地理信息系统数据采集、筛选、存储、分析与显示技术的升级，而且要彻底扭转传统地理信息系统重视数据管理与显示、轻视数据分析的状况。大数据的意义只有将地理信息产业的核心关注转移到用户价值上之后才能得到体现，否则大数据不如小数据，因为实现同样的价值前提下前者成本只会更高。

1. 地理信息发展职能的扩展

我国的地理信息产业发展经历了曲折的过程。前期，测绘和地理信息的发展注重政府管理和为国家安全提供服务，地理信息只是符号和数字的代名词，市场在资源配置、地理行业的需求方面受到限制。现阶段，地理信息的使用者不仅局限于静态的、定期更新的信息，而更加倚重于公众参与更新的实时或准实时信息。大数据中包含空间位置信息的数据量激增给地理信息的发展扩大了“交际圈”。

人工智能技术的发展使得具有更接近人的思维和表达方式的语义信息也能在计算机系统中得到表达，这使得和地理范畴有关的语义信息在互联网上呈现爆炸性增长，这些信息使得地理信息厂商在位置服务领域找到新的突破点。此外，物联网技术下的传感器数据、视频监控流媒体数据等，这些海量资源将有力拓宽地理信息产业发展的平台。

传统的地理信息技术需依据大数据的特征进行技术升级和改造，包括海量数据的预处理、高效分析与评价等，例如需要建立数据分析过程中的人机互动机制，提升分析结果的可视化水平等。如果能在上述应用领域取得进展，那么对于一般海量事务性数据的数据挖掘，将产生可以借鉴的技术模型，从而在大数据环境下引领整个信息产业技术的进步。

综合上述事实和预期，地理信息行业要结合实际需要，顺应时代潮流趋势，在新资源领域主动承担责任，拓展行业职能，提升行业影响力。

2. 地理信息产业技术的升级

地理信息产业的技术始终走在 IT 技术的前端，从最初的计算机数字化制图到多源异构型空间数据库的出现，再到 WebGIS、云 GIS 环境下地理信息服务的提出与实现，地理信息产业不断吸收数据的精华以为我所用。

大数据和地理信息产业技术的升级有着特殊的“互生互促”发展过程，即地理信息

产业底层硬件的进步带来了大数据，而又由于大数据的出现，促使空间数据库和 GIS 软件等计算层技术的革新与优化，与此同时，行业外的数据分析经验也引进到地理信息产业，使基于数据挖掘技术的服务有了更多的价值增长点。

大数据环境下新技术的出现使国内的技术发展出现更多的追求目标，我们可以借助大数据这一战略资源，通过市场换技术的方式引进国外先进的软硬件技术，通过技术的吸收达到行业技术的跨越式发展。

3. 地理信息服务模式的变革

原有行业应用集成服务和商业模式都发生变革。随着互联网技术的提高，BTOC 商务模式在消费领域成为主要的发展模式，以信息为载体，在网络端将企业与客户直接串联，达到服务的精确化和高速化，这也是地理信息服务商发展的必由之路。大数据的服务对象不再局限于政府部门或者对地理信息有特殊需求的企业和单位，而是面向所有对位置信息有需求甚至仅仅是有兴趣的个人，这使得服务端的需求呈现“大客户化”，数据庞大、类型众多的服务需求将严重挑战传统服务模式的承受度，使得服务商在服务资源的可伸缩性、服务效率的平衡性、服务类别的兼顾性上都需要进行改革和创新。

大数据应用深度挖掘的规则和趋势，将地理信息被动式服务逐渐升级为提供用户所需的产品信息和“数据挖掘后信息”，后者将是一项全新的服务单元，配合地理信息产业职能的扩展，提供吸引决策需要的辅助性数据和信息，将自身也纳入到决策方之中，成为客户的新型伙伴[㊀]。

2.10　新媒体大数据

2.10.1　新媒体大数据应用现状

全球数字化和信息化的发展进程越来越快，信息社会开始进入了加速的阶段，计算机技术和互联网技术越来越广泛而深入地影响到社会的各个层面。在信息化发展的进程中，“大数据”成为一个显学，大数据与新媒体之间的关系是媒体业界和学界都非常关注的课题。由于新媒体本身衍化和转型都非常快，这个课题带来了更多复杂的结构和分析，我们试图在模糊中把握规律，在探索中进行思考。

1. 传统媒体与小数据

在新媒体发轫之前，电视、报纸、广播和杂志是传统主流媒体，对于社会发展做出

㊀ 引自：周顺平，徐枫，大数据环境下地理信息产业发展的几点思考，P208, 1672-1586 (2014) 01-0045-06, 2013. 8。

了重要的历史贡献，都拥有自己的辉煌历程。传统媒体和数据之间是怎样的关系呢？

我们先分析收视率的例子。电视媒体的发展过程中，需要解决一个行业问题，比如《爸爸去哪儿》、《快乐大本营》这样的节目价值如何？到底有多少用户看这些节目？电视广告的价格如何来确定？这些问题催生了电视行业的典型数据问题——收视率问题。能不能用大数据的理念来解决收视率问题呢？理论上没有问题，我们只要收集齐每个观众收看不同节目的数据，就能够得出电视节目的市场效果和价值。在目前的收视率市场，收视率调查公司虽然开展了基于大数据的收视调查，但由于在全面获取大数据、运营商提供的数据公正性认证等方面的现实问题难以解决，传统抽样调查仍是目前主流的实践应用。传统的收视率调查方法获得的是小数据。第一，其操作的理论基础遵循统计学原理，通过对抽选样本的调查来推及总体的收视情况。如在北京，选取 500 户样本家庭（超过 1500 个 样本个人）来代表北京的观众情况。这就是基于抽样调查的小数据的理论基础。第二，采集方式上，目前国内及全球主流的方式为基于个人测量仪的硬件自动化采集，即在样户家中的电视机上安装个人测量仪，该设备会自动回传样户家庭中每个成员的精确到秒的收看行为数据。除此之外，在一些欠发达地区，由于市场需求等因素的影响，仍保留了人工填写“日记卡”的传统方式。第三，对采集回来的数据进行数据处理、加权等，并与节目数据、样户背景信息整合，最终形成针对不同地区、观众、频道、节目等反映电视收视市场的收视率。

随着电视传播技术的数字化发展，收视率调查技术也在不断向数字化转型，目前，个人测量仪及声音匹配技术可同时适用于模拟与数字信号，是当前多元电视传播环境中适用性较强的调查技术。只是我们现在再来看十几年前的技术创新，觉得是有一些地方需要调整了。

同样，报纸媒体的数据应用来自于订阅情况，还有一个非常重要的沟通手段，叫作“读者来信”。杂志媒体的情况和报纸媒体类似。广播媒体应用了收听率，情况和电视媒体类似，这些就构成了传统媒体与数据的关系。

2. 新媒体产生了大数据

随着计算机技术和互联网技术的发展，新媒体快速发展起来，类型多种多样，让人眼花缭乱。第一阶段，计算机技术和固网技术构建了传统的互联网新媒体；第二阶段，移动智能终端和无线网络技术又搭建了移动互联网新媒体，这两个互联网结构催生了现代意义的大数据。

新媒体技术体系为现代意义的大数据带来了技术保证。第一是计算能力。云计算和

个人计算机体系为我们提供了越来越强大的计算能力，新媒体正在推动将上千万台服务器集中在一起，为社会提供像水和空气一样廉价的计算能力，有了计算能力，新媒体才能够具有数据处理的基础。第二是传输能力，新媒体需要各种数据，这些数据的采集和分享需要非常优化的网络技术和网络条件，将数据传送回数据平台，再将数据平台的数据分享给各种终端。第三是技术能力。新媒体将各种数据采集回来，需要按照一定的业务逻辑处理，建立各种数据库，设计各种业务流程。数据挖掘、数据仓库、机器学习等技术是我们专门开发数据的技术体系。

新媒体传播体系为现代意义的大数据提供了传播平台。美国麻省理工学院的媒体实验室——《连线》杂志对于新媒体给出的定义是新媒体就是“所有人向所有人的传播”。这种传播方式形成了大数据的各种平台，也带来了大数据的各种结构。例如微信平台，就是基于移动客户端的即时通信社区平台，意图建立“连接”，体现了所有人向所有人的传播。所有人向所有人的传播带来了“大数据”，一方面是人人都是传播者，每个传播者都传播了大量信息，亿万人传播的信息都在处理、存储和发布，形成了海量数据的来源，大数据内容和应用信息源自于此。另一方面，人人都是受众，每个受众都在使用新媒体，在新媒体上留下了各种使用痕迹，受众使用行为构成了另外的大数据，英文“track”非常好地概括了新媒体用户行为。

3. 大数据推动新媒体发展

“大数据”这个概念来自于信息化和数字化发展潮流之中，是新媒体产生的数据的概括，“大数据应用”也是新媒体经营中关注度很高的一个问题。在阿里巴巴，它的“大数据业务”是一个独立的业务板块，是未来的战略发展方向。大数据应用对于新媒体业务会形成什么样的影响呢？可以说潜力巨大、影响深远。目前，大数据在以下方面对新媒体产生了深刻影响：

（1）大数据影响新媒体内容

在新媒体的业务经营中，判断内容产品价值和进行内容运营决策都要依托大数据。以视频网站为例，视频网站每年都会购买一批电视剧，有古装的、爱情的等不同类型，什么时间段推出哪个电视剧是靠什么决策的呢？除了电视收视率外，就是后台每天电视剧的访问量和访问趋势，这些决定了网站留给这个电视剧的位置和资源。进一步地，根据用户的访问量和观看时长，又进一步决定了网站每年上亿的购剧预算到底花在哪家公司和哪部作品上。大数据决定了哪些内容会有投入、哪些内容更有前途、哪些内容会分配更多资源，而后就进一步影响了围绕新媒体的内容创

作和生产体系。

（2）大数据影响新媒体广告

目前，广告经营是新媒体核心经营模式之一。新媒体广告一样需要数据来证明其具有什么样的商业价值和什么样的价格。举个例子，实效广告越来越成为新媒体广告的一种业态，它就是基于大数据的。企业投放广告时一直爱问一个问题，就是“我投放的广告有什么效果?”在大数据的支持下，新媒体广告已经能够回答这个问题。第一层面是曝光率，有多少人看到了这个广告；第二层面是关注度，有多少人点击广告并且感兴趣；第三层面是行动率，有多少用户打了电话或者是领取了优惠券；第四层面，特别是在电商平台上，直接能够看到有多少用户通过看这个广告产生了购买，形成了多大的销售额。这种大数据支持在广告主和媒体之间形成了良好的互动关系，企业能够得到更好的销售效果，媒体也更清楚如何服务广告客户。

（3）大数据推动新媒体服务用户

新媒体从诞生开始就继承了美国商业理念中“用户至上”的理念，全心全意服务客户，极度重视用户体验，这就是新媒体公司的理念。在各个互联网公司战略级别的会议上，如何吸引用户、如何提高用户黏度、如何提高用户体验仍然是核心议题。为了提升用户体验，大数据是非常重要的技术手段。“个性化定制”就是新媒体服务用户的一个利器，这个利器依托于大数据。新媒体公司通过长期用户行为数据的积累，开始逐步了解用户的信息行为，根据用户的信息行为，新媒体公司不断尝试推送新的内容和应用，这就是大数据非常典型性的应用。新媒体公司利用大数据不断尝试“预测”用户行为，甚至用户还不知道自己要干什么的时候，新媒体公司就知道了，能够非常好地帮助用户准备相关信息。

在全媒体技术的强力渗透下，大数据已经成为移动互联网时代不可或缺的重要内容，并对传统电视产业逐渐形成不同程度的解构。面对如此境况，电视媒体应积极调整媒体战略，建构电视领域的大数据资源库，利用大数据指导电视节目策划，加强传统媒体与新媒体之间深层次的合作互动，通过跨界生产寻求电视创新的新突破，实现电视产业的数据化转型。

2.10.2 基于大数据的收视率测量

1. 传统收视率测量方法

美国是世界上最早开展收视率调查的国家。1947 年，Hooper 公司首次在纽约采用

同步电话法进行电视收视率调查。进入 20 世纪 50 年代，AC 尼尔森公司将原本用于收听率调查的受众测量仪移植到电视机上，开始收视率调查，但测量仪法并不能提供人口统计方面的数据。随后，一家美国私人视听率调查研究机构使用日记法解决了这一问题。从 20 世纪 60 年代起，AC 尼尔森公司推出即时存储测量仪，为美国全国电视网提供收视率调查服务。中国目前最主要的收视率调查公司是央视 – 索福瑞媒介研究有限公司（CSM），采用的收视率的统计方法包括日记卡法和目前国际上最先进的个人测量仪法。日记卡法是指通过由样本户中所有 4 岁及以上家庭成员填写日记卡来收集收视信息的方法，日记卡上记录每个家庭成员每天收看电视的情况（包括收看的频道和时间段）。个人测量仪法则通过安装在电视机中的个人测量仪自动采集用户收看电视时的各种行为数据，根据电视传输信号的不同，个人测量仪会采集到不同类型的数据并直接回传到调查机构的服务器，针对采集到的数据，再分别采用 DFM、Si-code 或声音匹配（Audio Matching）等数据处理技术，最终与样户家庭中的个人背景信息、节目信息等进行匹配，从而综合地准确反映用户的收视行为。

然而，无论是日记卡固定样组测量法或者是个人测量仪调查法，都是建立在抽样调查的基础上。尽管收视率调查样本户是科学抽样产生，但抽样调查方法在实际操作中存在一些固有缺陷，如市场对调查精度的需求及其可接受的成本支出决定了样本量的有限规模、需要样本户高度配合、在法律不健全的市场环境中存在样本污染风险等。随着数字双向有线电视的推广普及，频道数量激增，业界需要更为精确和安全的收视率测量方法。

2. 大数据基础上的大样本收视测量法

大数据的出现为丰富传统的收视率调查方法提供了技术基础与想象力，但从目前的研究进展及实际应用来看，数字化进程及机顶盒数据的自身缺陷使之难以取代传统的抽样调查方法。首先，双向互动的数字电视普及率有限且发展极不平衡，特别是在非中心城市及广大农村地区，这使基于机顶盒大数据的收视调查难以客观、全面地反映电视观众的收视情况。在实际应用中，虽然大数据能提高局部精度，但它只能反映数字互动家庭户的收看行为，忽视机顶盒数据的这种局限性，必然将导致对整体收视分析的结构性偏差。其次，机顶盒大数据在收视率调查的应用中仍有许多需要解决的技术性问题，采集的数据仍有缺陷，如是否能准确测得家庭成员、准确开关机时间及收看时间等，这些都是关系到收视率指标统计的最基本数据要素。第三，收视率反映的是个人的电视收视行为，而基于机顶盒采集的大数据及其产生的收看数据反映的是“家庭户”的行为，远非真正意义上的个人收视率。同时，基于大数据的收视数据缺少

观众背景信息，如年龄、性别、收入等，而这些对电视价值评估至关重要。上述情况是目前国际收视率调查业面临的共同难题。面对大数据可能带来的无限商机，问题与难度并不妨碍国际收视率调查机构积极开展相关探索。全球行业性的共识是，机顶盒大数据的缺陷使之无法抛离抽样调查，更大的意义与价值在于二者的整合与融合，因此全球收视率调研机构都在积极寻求二者的有机结合，如以抽样调查数据反映整体与结构，以机顶盒大数据提升局部精度等，以期以创新的思路为传统收视率调查带来新的补充与提升。

依托大数据的收视测量在抽样调查方法的基础上丰富了收视率调查方法的多样性，通过高清互动机顶盒对观众开关机顶盒、转换频道、使用增值业务等操作行为进行精确到秒的准确记录，不但最大限度地保证了数据采集和传输的安全性，而且为实现大样本测量提供了基础支持。

2014 年 11 月 15 日，北京歌华有线电视网络股份有限公司（简称“歌华有线”）宣布建成大样本收视数据实时采集分析系统，可实时采集、分析、展示超过 400 万高清交互用户的收视行为数据。目前，北京约 500 万电视用户中，超过 400 万户已经使用了高清交互数字电视机顶盒，占全国高清交互数字电视用户总数的约 1/5。歌华有线大样本收视数据中心提供的数据显示，随着高清交互节目内容的不断丰富，北京地区有线电视用户每日每户平均收视时长几年来持续增长，从 2012 年的 192 分钟已经上升到 2014 年的 206 分钟，2014 年 11 月 12 日的平均收视时长为 221 分钟。数据还显示，北京地区高清交互用户近两年平均每日开机率稳定，保持在 60% 以上，2014 年 11 月 1 日 ~ 11 月 12 日，北京地区高清交互电视用户日均开机率为 65.11%。歌华有线大样本收视数据研究中心依托海量高清交互数据生产的北京地区收视率数据产品，为政府部门提供专业化的舆情宣传效果评价、舆情导向、舆情影响力度、舆情动态等各个维度数据的分析报告，通过提供节目全时段收视数据，促进建立科学合理的节目评价体系，为电视台节目布局、节目设置、节目调整、节目策划编排等提供参考，帮助电视台和节目制作商提高节目品质，强化节目创新，提高核心竞争力；依托大数据采集，分析电视用户使用偏好，为用户提供个性化的智能电视收视服务，打造电视用户收视指南，使用户及时了解并观看喜爱的节目；通过广告刊播情况统计分析，为广告主提供电视媒体广告曝光量、点击量、转化率等相关数据信息，以便广告主更有效地调整广告投放策略；此外，基于专业的数据挖掘技术使高清交互平台广告投放更为精准，避免广告重复投放，为增值业务运营等提供有力的数据支撑。歌华有线大样本收视数据研究中心 2015

年将逐步实现宽带用户数据、互联网电视数据、手机电视数据的采集，可提供更多维度的收视数据指标[㊀]。

3. 社交网络收视率

尽管基于电视机顶盒回路数据的大样本收视测量方法能够有效提高收视率测量的精度，但它仅能统计出通过传统电视机播放的节目的收视行为。随着智能手机、社交网络、视频网站等新媒体在全球范围内的快速普及，逐渐打破了电视等传统媒体主宰信息传播的局面。社交网络对电视观众的影响日益受到业界重视。电视节目在社交网络上的关注度与传统的收视率同等重要。尼尔森曾联合麦肯锡咨询公司进行调研，发现在一个新的电视节目初次开播的前几个星期内，年龄在 18 ~ 34 岁的最活跃的社交媒体用户群中，社交媒体上的评论每增加 9%，收视率就会相应地提高 1%。因此，社交网络的关注度成为衡量电视节目影响力的有效补充标准。

2013 年 10 月尼尔森公司第一次上线推特收视指数（Twitter TV Rating），通过衡量社交网络的活跃度来评估电视节目的传播效果。尼尔森公司从四个角度来衡量推特收视指数：与电视节目相关的 Tweets 总数，原创发表电视节目的作者数量，和电视节目相关的 Tweets 的曝光数量，看过与电视节目相关的 Tweets 的数量。

2014 年 7 月 2 日，由央视 – 索福瑞媒介研究有限公司（CSM）与新浪微博合力打造的微博电视指数 Beta 版宣告上线，成为国内首个基于社交媒体评估电视节目影响力的大数据分析系统。“微博电视指数”以微博上电视节目的讨论为基础，重点考察口碑影响力和受众覆盖情况，再经过大数据运算和关键词的系统优化，计算出相关电视节目在微博上的阅读量、提及人数和次数。同时，深入数据解读分析将进一步展现微博上讨论该档电视节目的热度和人群特征。作为收视率数据的重要补充，该指数在媒体融合的时代从多维度提供了受众的关注习惯，为电视栏目提供多维度、客观的社交媒体传播影响力评价数据[㊁]。2014 年第三季度，新浪微博日活跃用户增长 30%，达到 7660 万。2013 年夏季的“疯狂综艺季”吸引了 77 档节目参与，在新浪微博上的总讨论量约 5 亿条，其中《中国好声音》的讨论量高居榜首，在收视率上也始终领先。

同时，在网络用户流量日益增长的趋势下，传统电视媒体越来越注重通过微博等社交媒体扩大其节目影响力。截至 2014 年 1 月，来自新浪微博的官方数据显示：微博平台上已有超过 7000 个经过认证的与电视相关的官方微博，其中电视台官微 510 个、电视频道官微 712 个、电视栏目官微 6107 个，电视行业已经自动开启与社交媒体主动合作的模式。

㊀ 歌华有线正式推出“歌华发布”收视数据品牌，http://tv.sohu.com/, 2014.11。

㊁ 微博电视指数：如何让电视节目与社交媒体“联姻”，http://media.people.com.cn, 2014.7。

2.10.3 新媒体视频内容监管

1. 广播电视内容监管

随着广播电视事业的发展，我国先后建成了全国中短波广播监测网、全国卫星电视监测网、海外广播监测网、全国有线电视监测网和全国无线调频广播电视监测网，并对无线、有线、卫星监测网资源进行整合，建立了一个开放的监测信息综合服务平台，存储有大量国内外电视、广播数据。

然而，目前广播电视内容监管仍然采用人工方式完成监听监看、违规情况研判及数据统计汇总等工作，劳动强度大，工作效率有待提高，监管能力有待提升。为了充分发挥监管中心原有资源覆盖全面的优势，在已有的海量音视频数据中快速、高效地提取出可用信息，总结广播电视播出规律，国家广播电影电视总局构建了一个以音视频内容处理与管理为核心的广播电视监管统一自动编目和音视频信息检索系统，提高了相关监管业务的工作效率，更好地完成了国家广播电影电视总局对于广播电视节目内容的监管任务要求。其系统架构图如图 2-58 所示。目前，该系统采用基于 NAS 的并行存储结构，存储空间达到 2PB。存储的录像文件包括：全国上星 207 个卫视频道的所有录像文件，有线数字前端每月向存储系统回传的 92 个频道一周的录像文件，有线模拟前端每月向存储系统回传的 162 个频道一周的录像文件。

针对群众反映集中的卫视频道电视购物短片广告时间长、频次多的问题，2013 年 10 月 29 日，总局发布《关于进一步加强卫视频道播出电视购物短片广告管理工作的通知》。为了确保政令落实到位，总局决定从 2014 年 1 月 1 日起，开展为期 1 个月的治理卫视频道购物短片广告违规问题专项监管月行动。系统基于模式识别、模版匹配等技术，在少量人工辅助的情况下，利用节目间可区分的特征对广播电视流中的节目、广告等进行自动切分、标注，形成可检索的节目单，使海量音视频的重复利用、监管监督成为可能。利用该技术，系统建立了跨频道的广告模版库，面对上千个广播电视频道，精确定位广告的播出频道、播出时长、播出时段、播出次数，并根据广告的重复性在少量人工编辑与审核的情况下，完成广告内容、名称、分类的确定。监管月期间，共完成监看任务 14 批次 231 频道（次），监看节目时长约 4900 小时，发现 122 频道（次）播出违规购物短片广告 541 条（次），录像取证文件 45G。系统筛选出购物短片广告 95371 秒，传统模式的广告监管要对一天 24 小时的节目进行监看，应用本系统，平均每频道只需对约 40 分钟购物短片广告进行研判，大大降低工作量，提高工作效率。

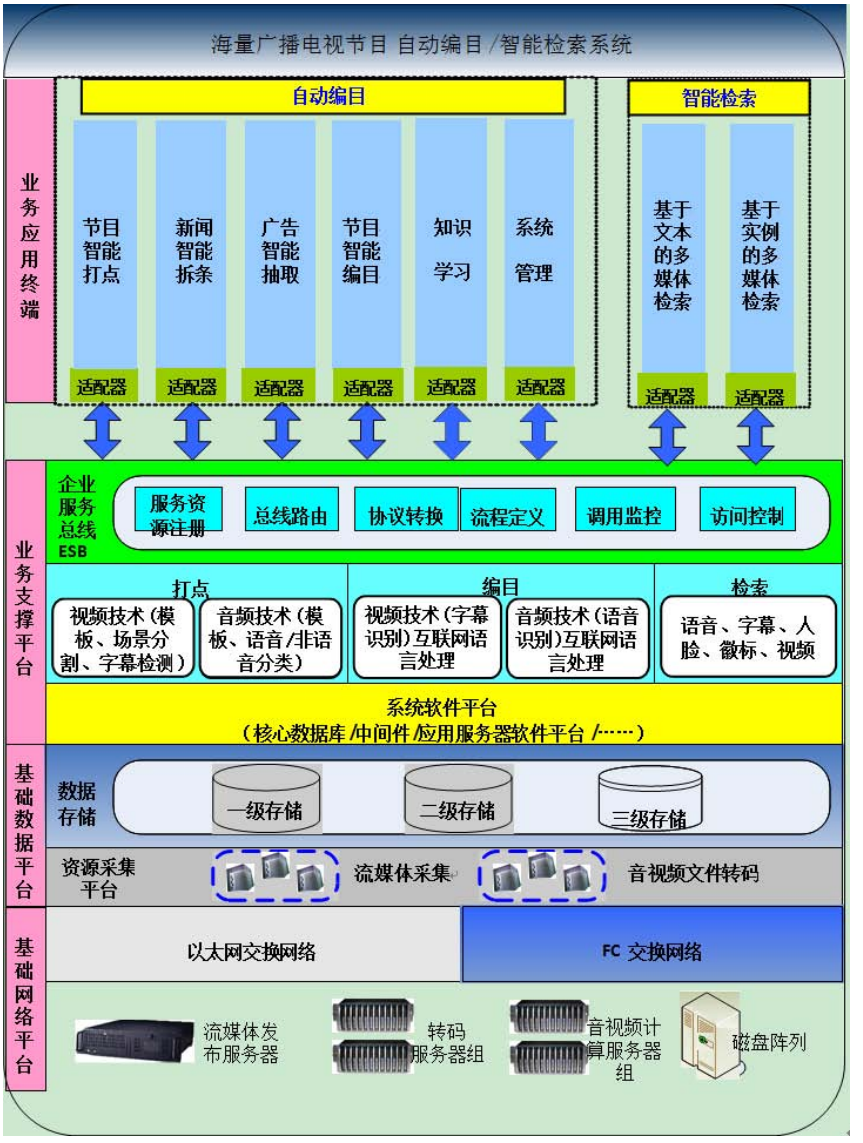


图 2-58　广播电视节目监测自动编目及检索系统逻辑架构图

此外，系统还实现了电视节目信息的编目自动化，将完整的视频流根据不同的广告、节目进行自动切分，包括视频特征提取、镜头检测、关键帧提取、镜头聚类，基于音频特征的切分与分类等。对于已知广告，可利用音视频模版匹配技术，利用已经具有语义标注的模板为重复出现的广告进行分类与内容标注。对于没有模版的新广告，可以利用重复性检测以及广告的视频特征进行筛选。

2. 互联网视频内容监管

我国国内目前每年有上万集电视剧投放，近十万小时电视节目上线和上千部电影进入市场。中国互联网络信息中心（CNNIC）《第 34 次中国互联网络发展状况统计报告》显示，截至 2014 年 6 月，中国网络视频用户规模达 4.39 亿，较去年年底增加 1057 万人，用户增长率为 2.5%。网络视频用户使用率为 69.4%，与去年年底基本持平。据艾瑞咨询公布的数字显示 2014 年 8 月在线视频 PC 端与移动端的有效使用时长分别为 59.3 亿小时和 38.1 亿小时，同比增长分别为 9.8% 和 372.7%。面对庞大的用户群体和无处不在的视频内容，网络视频内容健康监管面临巨大挑战。优酷网、土豆网等网站为用户提供微视频上传、播放和分享服务，其监管更加困难。例如用户将《色戒》、《苹果》等影片被删减的片段通过优酷等视频网站非法传播。此外，一些恶搞视频的传播对青少年的价值观产生消极影响，对社会主流价值体系也是一种巨大消解。

国家广播电影电视总局通过采集分析、模式识别、机器学习等大数据技术，从以下几个方面对网络视频数据进行监测，有效降低了内容监管的人力要求并提高了可靠程度。

（1）从信息获取的角度

互联网监管需要对互联网上的视频网站进行自动发现和跟踪。为了有效发现互联网上的视频网站，可以充分利用互联网上现有的网站数据资源，如网址导航、门户网站，发现网站信息并通过网站之间的链接关系，利用采集器的链接扩展功能进一步发现更多网站，自动发现和跟踪互联网网址，建立超过百万的互联网网站信息数据库。同时基于文本分类、链接分析等技术，分析网站内容的基本特征，对网站的影响力、敏感度、视频节目数量等信息进行评估。通过主动搜索与元搜索相结合，发现互联网上的视频节目，并结合信息抽取技术实现视频节目信息的自动识别、排重和入库。进而基于大数据处理技术，对视频内容进行分类、识别、聚合、统计等，以便于从微观和宏观角度对互联网视频进行有效监管。

（2）从微观的角度

需要对节目内容进行分析和审核。2012 年 6 月 ~ 2014 年 10 月期间，国家广播电影电视总局通过文本分类技术对系统发现的超过 1.3 亿个视频节目进行归类，对其中的社会、时政等类别进行重点监管。对电影、电视剧等视频内容，构建了涵盖国内外 6 万多部影视剧信息的影视资源信息库，并通过影视节目识别技术对互联网影视节目进行识别，发现互联网影视节目相关视频超过 450 万个，相关视频的总点击量超过 150 亿次，其中电影 4.19 万部、电视剧 1.21 万部、动漫 6135 部、电影相关视频 60.81 万个、电视

剧相关视频 123 万个、动漫相关视频 222 万余个。

（3）从宏观的角度

需要通过大数据分析提高对互联网视频舆情的掌握、预测以及快速反应能力。分析每日采集更新的互联网视频内容，结合新闻、微博等数据，利用视频节目数据的本身特征，包括标题、标签、分类、热度、发布时间等信息，以及关键词在不同时间段内的热度散布等特征，基于视频和关键词的双向表示模型，通过文本聚类、数据挖掘等技术对视频、新闻、微博等数据进行聚类分析，发现最新的热门视频事件。对于用户关注的热点事件持续追踪，通过趋势分析图和传播链分析图等技术跟踪热点事件的报道趋势及来龙去脉。对采集信息进行聚合计算，对当天、当周、当月重要的热点新闻、每天热点网络词汇信息进行分析和追踪，及时掌握互联网视频舆情爆发点和事态。图 2-59 是国家广播电影电视总局热点事件视频聚类效果实例图。针对重点关注事件，可以利用微博数据的特点，针对特定人、特定微博、特定事件的深度分析，对可能成为热点的敏感事件进行预测和趋势分析。

图 2-59　热点事件视频聚类效果实例图

2.10.4 大数据指导节目内容生产

传统电视节目内容生产采用确定主题、写稿、编片、审核、播出的单向封闭制作流程，节目内容一旦确定，在节目传播过程中很少发生改变。随着新一代社交媒体开始在世界范围内发展壮大，用户已经习惯在看电视的同时通过社交媒体进行讨论，电视节目给社交媒体带来了海量的话题和信息量。2014 年英国一项研究结果表明 14% 的受调查者通过社交媒体向亲朋好友推荐自己心仪的节目。2014 年马年春节晚会上共有 3447 万名微博用户参与了春晚互动，互动量（原创、转发、评论、点赞）达到 6895 万条，转发量最高的微博是韩国明星李敏镐的春晚拜年微博，转发 43 万次。因此，在媒体融合时代，节目内容的生产流程发生了新的变化。节目制作方可以一边进行节目直播，一边根据观众在社交媒体上的反馈决定节目接下来的走向。以内容生产、调整、播出、反馈融为一体的“制播同步”模式，将成为媒体融合时代电视节目内容生产的常态。

最早将大数据纳入节目制作流程的是英国广播公司（BBC）。BBC 在一些“真人秀”、访谈类的直播节目中，根据节目直播时用户在社交媒体上的节目评论内容统计，决定延长或者缩短相应的主题。印度脱口秀节目《真相访谈》在全球拥有 4 亿来自印度电视网络和 YouTube 的电视观众，以及 12 亿来自其官方网站、Facebook、Twitter 等各种媒体的用户群体，与超过 800 万民众通过 Facebook、网站留言、邮件、短信及电话热线等方式进行互动。每次节目直播时段 Twitter 上平均会产生约 40 000 条来自印度、加拿大等 165 个国家和地区的评论[㊀]。该节目主要关注印度的一些社会问题，大胆曝光印度社会的种种弊病，力图以“残酷的真相”唤醒大众，引发公共讨论，进而推动社会变革。节目编导通过分析网上数以百万计的热点议题及其讨论帖子确定节目主题。在节目播出的进程中，通过形成政府、媒体和公众三方议程的有效互动，促成公共政策的调整和完善。例如，印度国会在探讨儿童性骚扰的一期节目播出后，迅速通过了一项儿童保护法案，主持人阿米尔也获邀到国会听证[㊁]。

美剧《纸牌屋》的热映也很好地诠释了大数据对于新媒体内容制作的影响，并帮助 Netflix 公司扭转了亏损局面，完成了华丽转身。它的成功得益于 Netflix 海量的用户数据积累和分析。Netflix 的订阅用户数据库存有 2700 万美国订阅用户、3300 万全球订阅用户的年龄、性别、居住地、使用服务终端以及用户每天、每周的观看时间等信息。每天用户在 Netflix 上产生 3000 万多个行为、400 万个评分，还会有 300 万次搜索请求。因

㊀ 引自：大数据能为电视媒体带来什么，http://www.qnjz.com/，2014。

㊁ 引自：史安斌，刘滢 . 颠覆与重构：大数据对电视业的影响 . 新闻记者 .2014，(03)。

此，Netflix 根据数据技术推导出《纸牌屋》的关键要素，在该剧的样片出品之前就承诺将至少购买一共两季、26 集的《纸牌屋》，一经播出便迅速走红。

在国内，《爸爸去哪儿》、《中国达人秀》、《快乐大本营》等热度较高的节目也将大数据技术应用于节目策划、播出、营销和后续提升等方面。《爸爸去哪儿》在节目播出阶段对微博、论坛、搜索引擎进行了大量数据分析，了解观众喜欢的父子和桥段，在持续播出中重新分配出镜率并设计更多情节，使节目的热度得到显著提升。《中国好声音》节目播出后，通过数据分析发现互联网用户对某些学员的热议程度很高，因此相应制作了一档姊妹节目《酷我真声音》，通过访谈形式解答观众关心的问题，播出后引起巨大反响。

2.10.5　新媒体大数据发展趋势

自信息传播出现以来，人类的传播内容和方式就受到同时代传播技术的深刻影响。在全媒体时代，科技进步不仅影响传播媒介，更促进着传播模式的改变。新媒体的迅猛发展产生了海量数据，深入挖掘新媒体大数据蕴含的价值，将有力促进传统媒体的行业转型升级。中共中央政治局委员、中宣部部长刘奇葆在出席推动媒体融合发展座谈会中强调，推动媒体融合发展，要紧盯技术前沿，瞄准发展趋势，不断以新技术和新应用引领和推动融合发展；要始终把内容建设摆在十分突出的位置，增强核心竞争力，以内容优势赢得发展优势。因此，传统媒体行业必须要运用新媒体技术和平台，抓住移动化、数字化和网络化的大趋势，对内容、渠道、平台、经营、管理等各个环节进行改造，在原有的品牌优势基础上提升传播优势和公信力，使传统媒体的优质内容能够在新的传播生态环境下获得更好的传播。

然而，新媒体的发展在为人们的生产生活等带来极大便利的同时，也导致了一些问题，如个人隐私泄露、不良信息传播、谣言层出不穷等，未来仍需完善相关法律法规，并采取针对性的措施。

第 3 章　大数据技术进展

结合第 2 章给出的大数据典型应用与案例，不难发现，大数据技术能够将大规模数据中隐藏的信息和知识挖掘出来，为人类社会经济活动提供依据，提高各个领域的运行效率，甚至整个社会经济的集约化程度。目前，大数据领域每年都会涌现出大量新的技术，成为大数据获取、存储、处理分析或者可视化的有效手段；相关的平台和工具也越来越多，给业界提供了更多的选择，未来还会继续出现新的技术和工具。总之，大数据技术卷帙浩繁，难以用一个简单的分类学体系概而括之。3.1 节试图从三个视角（数据的生命周期、技术栈以及近年来通用的技术范例）对大数据技术进行梳理；3.2 ~ 3.7 节结合大数据生命周期以及一般的大数据处理流程，按照自下而上的顺序，介绍各种大数据技术的最新发展和动态。

为方便读者快速了解今年的大数据技术热点，我们根据 Gartner Hyper Cycle 的风格，绘制了一个大数据技术发展趋势图（如图 3-1 所示），图中标记了目前主要的大数据技术和它们在本章中出现的位置，希望能够抛砖引玉，对本章讨论的有关技术进行总结。

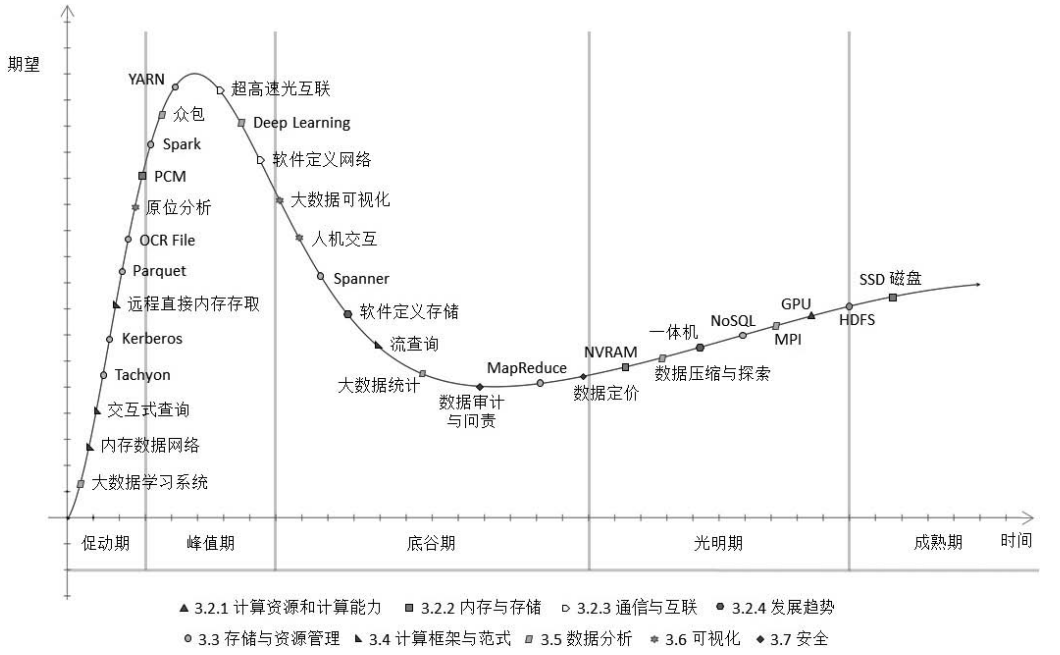


图 3-1　大数据技术发展趋势图

3.1　大数据技术图谱

3.1.1　数据的生命周期

图 3-2 揭示了典型的大数据生命周期：起始于数据源，经过数据获取和治理后进入存储系统，在需要的时候被调出进行计算处理，在分析和可视化后形成数据驱动的决策，或作为数据资源变现。

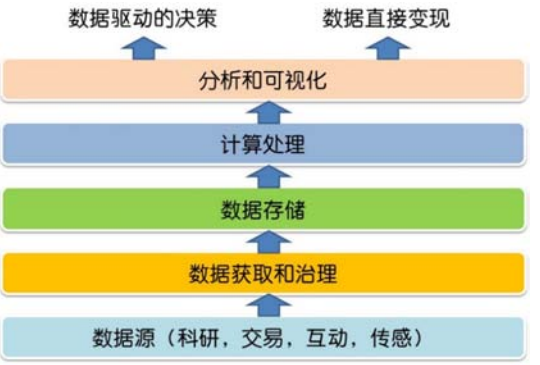


图 3-2　大数据生命周期

传统的集成生命周期管理（Integrated Lifecycle Management，ILM）在大数据时代面临新的挑战。

首先，相比传统的企业数据源（客户、财务和人力资源），新数据源更为多样，包括企业内部（如日志）以及更多外部（社交网络、电商平台和物联网等）数据。而且，数据类型五花八门，生命周期彼此不同，有些数据“有效期”极短，如果不能及时处理，即使存下来也已经失去意义。

在数据获取上，需要形成自觉、全面和客观测量世界的习惯。《Raw data is an oxymoron》（原始数据的说法是一种矛盾修辞法）指出数据不是自然资源，它不是“原始”的，它是由带着文化背景和主观倾向的人产生和解释的，带入了自我选择机制，从前数字时代到数字时代，无一例外。所以，必须尽量地减少主观性。第一，尽量由“机器”来决定采什么、哪里采。拿在程序里加日志为例，可以通过源代码分析工具来自动插入日志的写入点。第二，如果是数据本身带有主观性（如民意调查），那数据采集可能需要设计成多变量（如问很多问题）来约束主观误差。第三，尽量把数据采集和存储工具化，纳入基础框架，而不是来一个业务做一种采集 / 存储方案。

在存储、处理和分析阶段，ILM 需要跟各种新的大数据技术平台整合。有些数据遵从不同的生命周期。比如流数据，它有可能是先处理后存储，甚至无需存储。这些都带来了系统的复杂性。

随着数据量越来越大，一味追求“全量数据”可能会形成投入成本与产出的巨大反差。因此必须要对数据进行价值判断，进行多温度管理，根据数据保留策略或外部数据许可协议，选择归档保留或为无用数据画上句号。

在现实应用中的另一个挑战是，数据的采集、分析与业务脱节，面临组织、文化和技术的多重考验。

3.1.2 技术栈

图 3-3 展示了一个典型的大数据技术栈。底层是基础设施，涵盖计算资源、内存与存储和网络互联，具体表现为计算节点、集群、机柜和数据中心。

在此之上是数据存储和管理，包括文件系统、数据库和类似于 YARN 的资源管理系统。

然后是计算处理层，如 Hadoop MapReduce 和 Spark，以及在此之上的各种不同计算范式，如批处理、流处理和图计算等，包括衍生出编程模型的计算模型，如 BSP、GAS 等。

图 3-3 大数据技术栈

数据分析和可视化基于计算处理层。分析包括简单的查询分析、流分析以及更复杂的分析（如机器学习、图计算等）。查询分析多基于表结构和关系函数，流分析基于数据、事件流以及简单的统计分析，而复杂分析则基于更复杂的数据结构与方法，如图、矩阵、迭代计算和线性代数。

一般意义的可视化是对分析结果的展示。但是通过交互式可视化，还可以探索性地提问，使分析获得新的线索，形成迭代的分析和可视化。基于大规模数据的实时交互可视化分析，以及在这个过程中引入自动化的因素，是目前研究的热点。

有两个领域垂直打通了上述的各层，需要整体、协同地看待。一是编程和管理工具，方向是机器通过学习实现自动最优化、尽量无需编程、无需复杂的配置。另一个领域是数据安全，也是贯穿整个技术栈。

除了这两个领域垂直打通各层，还有一些技术方向是跨了多层的，例如“内存计算”事实上覆盖了整个技术栈：

- 在硬件平台层，主内存计算采用大内存，广义的内存计算还使用了基于闪存或下一代 NVRAM 的介质。不同节点的内存之间要求低延迟、高带宽的数据传输。
- 在数据存储和管理层，内存缓存、内存存储、内存数据库和内存文件系统是主角。
- 在计算处理层，有各种基于内存的计算框架，它们架构在不同的存储层上。比如内存数据网格（in memory data grids）在传统数据库上加上一个基于内存的 MPP

处理层或 MapReduce 层，以 Spark 为代表的内存计算框架则是在 HDFS、Hive Metastore 或 Tachyon 内存文件系统之上实现算子的堆叠。

- 在数据分析和可视化层，有些分析库和工具直接基于上述内存计算层，比如 MLlib 建筑在 Spark 之上，而很多 BI 工具需要仔细设计内存数据结构以实现交互式分析的可能。

对于产业界来说，选择合适的技术栈是最大的挑战。传统企业面临太多的选择，必须根据不同的场景灵活使用。比如对于很多企业来说，最经济的选择是对原有架构进行扩展，如增加队列、采用数据库分片和内存缓存，往往能够满足多数的应用场景。如果不行，转向过渡性的大数据架构也是一种选择，比如采用内存数据库和 Big SQL，毕竟能够发挥原有 DBA 和编程人员的余热。更激进的选择是转向全新的大数据架构，如 Hadoop 和 NoSQL。企业需要适应灰度部署和切换，因为在应用场景中需要传统的 RDBMS 处理 OLTP，也需要数据仓库处理传统商业智能，需要流式的事件处理，也需要批处理式的数据处理（如 ETL）和业务分析。

这需要架构师能够灵活地权衡三个因素：数据量、分析精确性和实时性需求。批处理能够满足前两个因素，流计算更擅长第三个。有很多技术在满足大体量的同时能获得实时或近似实时的响应性（如 Dremel 基于只读数据的即席查询和 BlinkDB 基于采样的查询），但损失了精确性。

另外，较深的技术栈对洞悉系统行为带来很大的挑战。Facebook 发现，一个在上层表现为 IO 读为主的系统，经过技术栈的层层处理（日志、压缩、复制和缓存），到底层时已经表现为以写为主的行为。

3.1.3　通用范例

本章内容基本上是按照技术栈的各个层次组织的，但也有一些范例是通用的、适用于不同层次的。

范例一　把大数据变小（减少空间复杂度）

数据压缩提高内存空间效率，减少网络瓶颈，如 Apache Parquet。

- 列存储使压缩率获得数量级的提升，而它本身也是由大变小的办法，因为查询往往并不需要访问所有列，所以数据加载时无需读入不相干的列，当然，好的 SQL 优化器也对行存储做裁剪。

- 缓存机制也是天然的由大变小。例如分布式缓存根据数据的活性和局部性识别出活跃数据，并将其置于内存中。
- 为了得到计算的实时性，有时要舍弃旧数据、冷数据和静态数据。流式处理便是这样；有时也需要根据数据冷热程度重新分布数据存储，如 Facebook 的 F4。
- 利用大数据的稀疏性，采用稀疏数据结构。
- 最后，大数据的信息密度低、熵高，需要数据清洗、数据治理等方式提升信息密度，便于后续处理。

范例二　把算法复杂性降低（减少时间复杂度）

分布式计算设施对大数据处理的加速是线性的，而很多大数据算法的复杂性是平方甚至是立方的，因此，降低算法复杂性是实现实时洞察的重要手段。

首先是采用更简单的算法。Peter Norvig 早在 2009 年用机器翻译的演进揭示“大的相对低质数据集 + 简单算法”优于“小的优质数据集 + 复杂算法”，推广到其他基于 Web 文本的机器学习，简单的统计算法（如 n-gram 和线性分类器）配上百万以上的特征，要比试图涵盖通用规则的精心设计的语言模型有效。同样的情况发生在基于图像的应用当中，数据的充沛可以弥补简单算法的缺陷，同样，简单算法使得对大规模数据的处理成为可能。算法组合（ensemble）是提升数据分析精度的重要方法。对多种 $O(N^2)$ 的简单算法组合，所得到的整体复杂度还是比 $O(N^3)$ 的单个复杂算法低。

其次，利用采样和近似算法。两者都能在精确度要求不高的情况下有效降低算法复杂度（以及空间需求）：

- 如 BlinkDB 通过离线和在线的两重采样，在 100 个节点的集群上对数十 TB 的数据进行查询，获得小于 2 秒的延迟，但同时误差率达到 2% ~ 10%。Facebook 的 Presto 采用了 BlinkDB 的技术。
- 在近似计算方面，针对 Membership Query、Range Query、Top-k、Frequency Estimation 以及 Cardinality Estimation 等计算，出现了 Bloom filter、HyperLoglog、CountMinSketch 等算法，在误差局限于个位数百分比的前提下，实现时空复杂性几个数量级的降低。Summingbird 子项目 Algebird 实现了大量的近似计算 Monoid。

再如，在很多大数据场景中，数据是很稀疏的，相应地采用稀疏数据算法，可以降低复杂性，把 O(n) 复杂度的算法降低至 O(nnz)。

在更复杂的场景里，数据在高维空间中、但数据点很稀疏，这也催生了一些降低复

杂性的方法。混合建模是其中一种，如把适用于小数据规模的带参数（parametric）模型与大数据规模的无参（non-parametric）模型结合起来，先用后者在不同的低维空间建模，再由前者把这些模型综合起来，这种方法能够有效地解决稀疏性和计算量的问题。还有一种方法是针对高维数据进行降维，然后再对其施以通用算法。

范例三　降低精度和一致性的要求

以 NoSQL 数据库为代表的大数据架构多采用舍一致性来获得响应性和扩展性的设计。CAP（Consistency、Availability 和 Partition-tolerance）理论有十多年的历史，最近几年十分火热，分布式系统既然不能舍 P，就必须在 A（可用性）和 C（一致性）有所取舍。七八十种 NoSQL 数据库，多数采用弱一致性获得高扩展性下的快速响应。典型如 BASE（Basically Available Soft-state services with Eventual-consistency，BASE 英文意思为“碱”），放弃严格一致性，但是获得最终一致性，与 ACID（Atomicity、Consistency、Isolation、Durability，ACID 英文意思为“酸”）针锋相对。在具体实现上，也有不同的选择。BigTable 系支持单行的强一致，对多行退化为最终一致性；Amazon 的 DynamoDB 提供可选择的一致性，根据服务 / 价格水平提供强一致性或最终一致性；还有系统采用分层的一致性，如 Google 的 Megastore，在 Entity Group 内部采取满足强一致性，在 Group 之间则选择弱一致性。

一致性的取舍已经从 NoSQL 数据库延伸到更多的领域，比如如今比较火的分布式机器学习和在线 / 流式机器学习。有些机器学习算法推出了并行实现，如并行 Gibbs Sampling 采用 colored fields 或 thin Junction Trees 进行优化。但更多的机器学习算法具有比较强的串行的计算 / 数据依赖，以至于无法适用于大规模分布式计算环境。Google 的 DisBrief 尝试将细粒度的、频繁的依赖（如单次迭代间的依赖）打破，变成粗粒度的、多个迭代周期发生一次的依赖。它对样本数据切分，对子集独立训练模型副本，定期地将副本更新到全局参数服务器，或从参数服务器获得其他副本的更新。它背后的逻辑是：放松数据同步中的一致性，固然会导致训练过程中丢失数据，从而降低精度；然而，大数据集的学习可以弥补单个数据精度的丢失，而且，放松的一致性允许学习在大规模集群上并行进行。CMU 进一步把这种机制抽象成 SSP（Stale Synchronous Parallel），把参数服务器和参数的有限过期（bounded staleness）模型化。

在线 / 流式机器学习要求在新数据流入的同时可以在线学习、更新模型，同时又能不间断地利用新模型进行分类或预测。这里同样涉及一致性的取舍。例如 Jubatus 的 mix

同步机制在更新全局模型的同时允许训练和分类 / 预测并发进行，虽有数据的丢失，基于上面同样的逻辑，更多的数据可以弥补。

范例四　提高算法对长尾分布和丰富信息维度的容纳性

对于具有丰富信息维度的数据，Norvig 的大数据 + 简单算法理论未必适用。比如，深度模型的突飞猛进一定程度上就说明上述结论也许过于简单化。大数据、尤其是非结构化数据里蕴含的丰富信息维度，需要复杂的、高容量的、具有强大表达力的模型。如，从简单的带参模型到无参模型，从判别（discriminative）模型到生成（generative）模型，从线性模型到非线性模型，从浅层模型到深度模型。

前述的组合算法，或者更一般性的“ Society of Mind”方法，把诸多简单算法组合起来，获得优于单个复杂算法的效果。这在 IBM Watson 主机的 DeepQA 上得到了展现，该系统运用了上百种算法，从不同维度分析数据 / 证据对备选答案的支持程度，通过众愚成智的方式获得更高的准确度。

大数据的一个基本假设是能够倾听每一个个体的声音，也就是说，即使是长尾的低频信号，也是有价值的。然而，现在很多算法，如 pLSA 和 LDA 这样的概率模型，都是基于指数分布假设，因此把长长的尾巴割掉了（把低频信号过滤掉了），因此无法很好地理解小众需求。谷歌基于神经网络的 Rephil 提升了对长尾的容纳性，因而极大提升了 AdSense 的广告相关性。

范例五　人与机器的消长和合作

在 19 世纪末、20 世纪初，computer 别有深意。最著名的是“ Harvard computers”，特指一些善于处理天文数据的女性。20 世纪，computer 开始以机器的形式出现。在前大数据时代，即使机器的计算能力是数据分析的基础，人仍起到至关重要的因素：人提出假设，精挑细选获得数据，建立模型最后检验结果。数据科学家的直觉和灵感是数据分析成功的基础。在大数据时代，数据不再是稀缺资源，数据收集能力进入基础设施，社会个体在无意识中被数据化，机器无需人的介入直接访问散落在各处的数据。这时候人的关键性作用被削弱了。人的假设或许以偏概全，人挑选数据可能带有主观色彩，而机器可以做到不偏不倚不漏。更重要的是，其以机械化的数据挖掘寻找相关性来替代主观假设，让“数据自己找到数据、相关性主动找到你”。Google Correlate 可以发现任意关键词之间的相关性，或关键词与任意输入曲线的相关性，这是传统的人主导的数据分析无法实现的。

在前大数据时代，涉及语义分析的领域人是无法替代的。大数据改变了这一切：人的数字足迹的无意识记录，以及机器学习（尤其是深度学习）晓意能力的增强，逐渐改变机器的劣势。基于大数据的情感分析、价值观分析和个人刻画，在一些经验密集型的领域（如人力资源）正蓬勃展开。

在技术层面，人在机器学习中扮演的角色也在弱化。传统机器学习是特征加模型，模型大同小异，关键在特征。而特征抽取是人工工程，经验丰富的团队可以发掘出更好更多的特征，但也面临边际效益递减的情况，最后就无法提高了。数据样本的增加超出了人工提取有效特征的能力时，机器特征工程的优势就凸显出来了。相比传统浅层学习，深度学习使得机器能够做特征的学习，从海量数据中构造海量的冗余特征，以数十亿计，加上非线性模型，优势一下体现了出来。当然，目前深度学习仍然需要人的经验，亟待理论上获得发展，未来机器的优势更将明显。

可以预见，在数据科学家的巨大缺口下，人与不断发展的机器分析工具将会不断磨合和重新分工。在研究界，MLBase 作出了很有价值的尝试，帮助终端用户选择最佳的特征和模型。大数据的可视化也越来越强调交互式，即可视化不只展示答案、又是激发问题的过程，需要人与机器的协同。华盛顿大学的 VizDeck 把可视化看成是打牌。机器先产生一手牌，每一张牌是一种 / 一部分数据的可视化表示，系统在用户的交互中对“牌”进行排序（像抓牌后按花色排序），进而生成最终的可视化效果。

人与机器的合作还体现在众包，例如 Google 的 reCAPTCHA 和伯克利的 CrowdDB。

在大数据和智慧社会的发展之路上，人与机器的共同进化（coevolution）将是个极其重要的范例。

3.2　大数据基础设施

大数据的特定计算和存储需求对硬件平台创新提出了新的要求。从本世纪初的数据密集型可扩展计算（Data Intensive Scalable Computing）开始，大数据架构与高性能计算、云计算架构交错发展、互相借鉴。总体而言，相比起后两者，数据密集型计算走了一条不同的路。

在处理天文大数据时，Jim Gray 最早尝试使用商品化计算组件（PC 服务器）搭建所谓“裴欧沃夫（Beowulf）集群”。谷歌真正开始大规模部署 PC 服务器组成的水平扩展集群，通过软硬件协同将其变成仓库级计算机（Warehouse-Scale Computer），以应对互联网大数据的指数级增长。这一发展路径强调机柜的高密度，在功耗、散热、空间限制下获

得最大的处理效能。近年来，在商用和研究领域出现了一些新的趋势，下文分别从计算资源和计算能力、内存与存储、通信与互联三个方面阐述，最后阐明整体的发展趋势。

3.2.1 计算资源和计算能力

首先，大核和小核（Brawny vs.Wimpy Cores）的路线之争。小核在能耗上的优势一度有星火燎原之势，不仅是 ARM 服务器横空出世，Intel 也推出了微服务器。但是几年过后，基于小核的架构只是在一些细分领域（如冷存储）得到了有限的发展，而其旗手之一 Calxeda 也黯然关门。谷歌负责基础设施的 Urs Hölzle 指出，大核在多数情况下仍然优于小核。小核架构的拥趸们往往忽略软件开发成本以及集群的整体成本。在性能上，小核架构的响应时间明显更长。总体来说，小核架构只适用于一些特定的应用，只有当其性能接近中等的大核架构时，才有可能在通用负载上形成优势。亚马逊 AWS 的主要架构师 James Hamilton 最近分享了相似的观点㊀，从综合成本模型上看，基于 ARM 的服务器芯片远远落后于 Intel 的主流服务器芯片。

即便如此，小核架构的一些设计也影响了大核架构，比如互联和 SoC。目前，大核也开始走 SoC 的道路。Intel 的至强（Xeon）服务器芯片已经形成了 30 多种半定制化设计，而其 Xeon D 系列将全面转向 SoC 架构。在大数据架构发展历史上，很多技术都是从高性能计算往下走，而 SoC 化应该算是一个异类，它发源于移动和嵌入式架构，并且往上走影响了高端系统的发展。

其次，异构计算以其高计算密度和并行性、相对低功耗已经成为数据密集性计算发展的共识。异构可能在不同层面发生，集成的异构多核、CPU+ 加速器和异构集群是三种不同的实现。三者之中，CPU+ 加速器应用最广。

在加速器当中，又以高端 GPU 和至强融核（Xeon Phi）最为主流，应用到深度学习（如 Caffe、Theano）、关系计算（如 Red Fox）、可视化（如 MapD）等诸多领域。

再次是 FPGA，在包括在线查询分析（如 Netezza）、图计算（如 Convey Hybrid Core）、搜索（微软 Bing）、机器学习等领域得到应用。对于深度学习的某些应用领域（如计算机视觉），美国 NSF 研究㊁认为 FPGA 系统在性能和功耗上优于 GPU。在典型情况下，上述系统存在一些缺憾：加速器卡上内存有限，与 CPU 之间的连接（通常为 PCIe）带宽受限。就编程而言，GPU 和至强融核又优于 FPGA，但 FPGA 具有可重构能力。

㊀ http://www.bloomberg.com/news/2014-11-13/amazon-arm-chipmakers-aren-t-matching-intel-s-innovation.html。

㊁ http://insidehpc.com/2014/08/nsf-study-probe-advantages-f/

第三类加速器——特定领域的加速器如今大行其道，尤以神经处理单元（Neural Processing Unit，NPU）为盛。在客户端有 IBM 的 TrueNorth、高通的 Zeroth 和中科院计算所的“电脑”，前两者以脉冲神经网络为主，强调在线学习能力但识别精度较差，后者属于人工神经网络，依赖于后端的离线学习但识别精度高。在后端，计算所也在尝试使用新的 NPU，“大电脑”集成大容量的 eDRAM 形成以数据为中心的 NPU 架构，加速大规模的机器学习。

集成的异构多核有望得到长足的发展。Intel 自 Sandy Bridge 架构开始，就把 CPU 和 GPU 集成到一个芯片，通过 last level cache 交换数据，这样解决了两者之间数据交换的瓶颈。AMD 的 APU 也采用了 CPU 和 GPU 的异构融合。Intel 从 Haswell 服务器开始支持高速缓存的 QoS 管理，未来能够更加合理地在 CPU 和 GPU 之间划分资源，对两边的计算进行协调。

异构集群早已有之，多指因为计算资源渐进增长而形成的集群内系统的不同构性。这里强调因为特定计算需求而在集群里配置高计算密度的机器，比如亚马逊提供 GPU 实例与其他机器组成集群。但因为 GPU 本身不能自主运行，上述 CPU+ 加速器的问题未能解决。下一代 Intel 至强融核（Knights Landing）将解决这些问题，作为高计算密度的芯片，它既可以作为加速器，同时又能独立启动系统，而无需作为通过 PCIe 依附于 CPU 的从属。

第三，内存与处理器更加靠近。为增加处理器与内存之间的带宽，工业界和学术界在尝试把内存从存储体系上移至处理器。Intel 和 IBM 都开始把 eDRAM 放入处理器芯片，作为最下一级的缓存；把处理器与内存进行三维堆叠的做法已经在学术界研究多年，工业界已经开始试图产品化。

另外，数据的移动在延迟和功耗上极为昂贵，把处理器下移到内存是另一种新的范式切换，“内存中计算”（Processing in Memory）使架构从“以计算为中心”演化到以“数据为中心”，实现高度的数据并行计算，如内存厂商美光的 Automata 处理器。此类架构对软件和工具链要求很高。

3.2.2　内存与存储

内存和存储技术的快速发展大大提高了数据存储的规模和访问的效率。

2008 年，单服务器内存通常只有 8GB，现在主流内存达 96GB 到 192GB，内存容量迅速增加。这不但推动了大数据“内存计算”的普及，而且模糊了内存和磁盘的边界，

越来越多内存被用于 caching，甚至当成 ramdisk 使用。

2008 年，主流配置单磁盘容量只有 1TB，一个机器通常挂 4TB。到 2014 年，磁盘密度大大发展，单服务器挂载磁盘通常达 48TB，2014 年 9 月，西部数据宣布了单盘容量到 10TB，磁盘密度进一步提升。另一方面，磁盘的接口也开始向 IP 化发展，通过 IP 技术，融合对象存储、分布式架构和高密度存储平台，迈向全 IP 数据中心的愿景，典型的产品是希捷 Kinetic 硬盘。

2014 年 9 月 Intel 开发者峰会展示了 2U 服务器可以容纳 1.5TB 内存和 100TB 硬盘，使高密度部署更上一层楼。而同年发布的 Intel E7 v2 系列处理器，能够支持高达 32-socket 的高端服务器，单服务器最高内存容量可达 48TB。这推动了高端大数据一体机的发展。微软在 2014 年 10 月宣布的 Azure G 系列虚拟机能够提供单虚拟机 448GB 内存。

从数据的组织方式的角度来看，在传统的内存模型中，通常直接或间接地通过地址或引用访问数据。这种方式对基于内存计算的大数据处理而言并不适合，因为大数据处理需要的操作级别更高，即通常是对整个数据集合的访问与操作。因此，内存中的数据必须以集合的形式组织，以便于访问与索引。另外，要求支持外部内存池溢出到磁盘，从而增大内存计算的数据规模。

从数据的可靠性与可用性上来讲，内存是易失性的存储介质，在传统的关于可靠性的研究中，通常可以通过多机备份或将数据写入永久存储介质以保证数据的可靠性，但不可避免地会影响系统吞吐率。内存计算中引入的存储级内存，作为一种可能的解决方案，可以提供某种级别的数据可靠性保证，然而依然存在着节点失效等问题。因此，如何结合新的内存特性设计高效的数据可靠性保证机制是一个关键问题。针对内存数据的可靠性与可用性已经有很多研究工作，例如 Spark 和 Tachyon 使用 checkpoint 和 lineage 来支持容错，redis 使用快照存储（RDB）和写操作存储（AOF）来支持容错，RAMCloud 通过多机备份和硬盘备份，SAP HANA 通过 SSD 加速日志处理。

除了内存和磁盘，其他存储也快速发展，SSD 越来越普及，单 SSD 容量已经跃入 4TB，SSD 单位容量的成本只有内存的 1/30（SSD 价格：$0.5/GB，内存价格：$15/GB），延时是磁盘的 1/125（SSD 延迟：80us，磁盘：10ms），费效比非常高。SSD 在大数据的应用场景中越来越多，与内存计算相得益彰，例如 SAP HANA 中 SSD 与磁盘各司其职，Databricks 采用 Spark 在 Daytona Gray 类 Sort Benchmark 上创造了并列第一的成绩[㊀]，则是基于 AWS 的 SSD 存储 I2。

㊀ http://www.csdn.net/article/2014-11-06/2822505

SSD 作为磁盘的一种变体并未跳出存储分层体系的范畴。相比之下，PCIe SSD 和闪存存储（flash storage）选择了更为激进的道路。

PCIe SSD 从特立独行的 Fusion io 到众望所归的 NVMe，以其轻量级栈、低 CPU 开销、直接闪存访问带来的高吞吐量、高 IOPS 在大数据场景中得到广泛应用。

另一方面，闪存存储也得到了长足发展。其中，全闪存阵列代表了一个方向，相比 SSD 对磁盘的简单替换，全闪存阵列是对整个传统、厚重系统软件栈的颠覆，在保证低延迟、高带宽、轻量化的同时，加入自动精简配置、实时压缩和重复数据删除等功能，为上层应用提供语义丰富的 API。除了企业应用场景，闪存存储还在特定应用场景中得到应用，如 Facebook 的 McDipper 作为内存缓存 Memcached 的补充。

在全闪存阵列得到更广泛应用之前，包含闪存和磁盘的混合存储或联合存储则有望大行其道，这对软件的管理能力提出很大的挑战。谷歌的 Janus 智能地把数据在闪存和磁盘之间分配、迁移，闪存只存放 1%的数据，却能服务 28%的读操作。

下一代非易失内存（NVRAM) 也将渐渐走到舞台的中央，其性能接近 DRAM（最短延迟为 DRAM 的 2 ~ 3 倍)，容量大幅增加，数据不易失，可擦写次数更多，字节访问（闪存只能块访问），这些特性将改写整个分布式存储应用的版图。在其作为内存时，我们将不再需要对内存数据做检查点，不需要预写式日志，不需要序列化（写入磁盘持久化）与反序列化（从磁盘载入内存），就能实现更高效的内存数据库。而当其作为存储时，也将对文件系统为主的存储架构产生巨大影响，使其进一步走向对象化。

在追求快的极致的同时，大的极致也带来挑战。谷歌作为世界上最大的磁带机买家之一，利用磁带对 EB 级别的数据进行备份和管理[㊀]，并通过位置隔离、应用层问题隔离、存储问题隔离、存储介质问题隔离等多种混合手段保证数据的可用性。

3.2.3　通信与互联

分布式内存计算首先要减少跨节点内存数据传输，如 Spark 通过 HashPartitioner 实现协同划分、减少跨节点 Shuffle。如果不能避免，那就需要高速的跨节点内存数据传输。目前多数采用软件实现方案，例如 Spark 采用 BitTorrent 的实现来加速数据传输。由于单节点无法容纳所有数据，大数据就必须通过网络搬运数据。搬运的链路延迟要更低、带宽要更大、网络的可扩展能力要更强，还需要一定的可重构能力。

首先，要发挥内存计算的威力，需要更快的跨节点内存数据传输能力，否则传输延

㊀ How Google Backs Up the Internet.https://www.youtube.com/watch?v=eNliOm9NtCM

迟将抵消内存访问的好处。比如，RAMCloud 期望理想的点对点传输延迟低于 5 微秒。RDMA 相关的技术，包括 RoCE（RDMA over Converged Ethernet）、iWarp（internet Wide Area RDMA Protocol）等，对降低网络延迟非常重要。RDMA（Remote Direct Memory Access）即“远程直接内存存取”，是一种软硬结合的技术。RDMA 让计算机可以直接存取其他计算机的内存，而不需要经过处理器耗时的传输。通过消除外部存储器复制和网络栈操作，因而能腾出总线空间和 CPU 周期用于改进应用系统性能，从而减少对带宽和处理器开销的需要，显著降低了时延。目前标准的 RDMA 实现脱胎于高性能计算，其扩展性未必适用于大数据更大规模的集群。这必将催生更轻量级、可扩展的 RDMA 实现。

等到 NVRAM 技术普及之日，RDMA 技术可以把所有 NVRAM 连接成池，并轻易地达到 PB 或者更大的级别，实现更简洁的内存计算，从而把开发者从繁重的程序优化过程中解放出来。把 RDMA 技术应用到大数据环境目前还存在一些局限，主要是可扩展性需要增强，以满足成千上万节点之间任意快速互通的要求。

其次，大数据需要大带宽。超高速光互联是解决这一问题的理想方案，Intel 和 Facebook 在 2013 年初已展示基于硅光（Silicon Photonics）的机柜方案，当时是 100Gbps，这一两年将上升到 800Gbps，并且因为其成本优势全面商用化。2015 年上市的至强融核处理器（Knights Landing）其 Omni Scale 互联也将基于硅光。硅光不仅可以带来大的带宽，其功耗也相对大大降低，同时提供更加绿色的大数据基础通信设施。在 2014 年旧金山 IDF 上，Intel 对外宣布了传输距离达到 300 米的硅光通信模块，这极大地增加了硅光技术的实用价值。未来硬件方面的发展可以进一步降低延迟和提高带宽。RAMCloud 要求点对点传输延迟小于 5 微秒，因此曾经希望把 NIC 集成到 CPU。未来硅光能够实现芯片级的链接，有望实现 RAMCloud 期待的 1 微秒延迟。

再次，大数据要求大的数据吞吐量，并且在突发时要求更迫切，比如在 MapReduce 的 Shuffle 阶段。具有弹性的网络拓扑可以在需要时灵活的增加网络吞吐量，从而更好地适应大数据的应用。Intel 的 Rack Scale Architecture 就是这种新型高弹性网络架构的一个代表。数据压缩、加密解密等硬件加速技术也为网络提供另一方面的可扩展能力。

最后，可重构的网络助力大数据。SDN 技术在多流并发以及优先级管理上可以为大数据应用提供更灵活的网络传输能力。SeaMicro 的 Freedom Fabrics 允许对外部存储进行重构，数据即可以像 SAN 或 NAS 那样共享，又可以像 DAS 那样分配给每个计算节点，构造像 HDFS 那样的场景。

3.2.4　协同设计

软硬件协同设计是大势所趋，具体表现为如下技术取向。

单个节点层面，把硬件暴露给软件栈，以解决计算资源未充分利用（underutilization）的问题。现在的主流大数据栈都是基于 Java 虚拟机，与底层硬件隔离，无法充分了解硬件运行情况、更无从对其进行并行 / 向量优化，优先内存操作以减少 IO，缓存 / 内存局部性优化，拓扑感知优化以避免通信等。Intel 最早采用 NativeTask 对 Hadoop MapReduce 进行优化把某些 Java 代码（如 sort）放到原生代码中，进行缓存优化，对 MapReduce 应用的整体性能提升非常明显（达到 50％）。MIT 的 TupleWare 是借鉴 Hadoop 理念构造的大数据存储系统，但充分利用了底层架构的向量化支持（SSE4）、超线程、分支预测和局部性优化。Spark 最近也开始采用原生代码优化或对缓存进行特殊设计[㊀]。在大数据学习系统（Big Learning System）中，非常强调机器学习软件与底层系统的协调设计[㊁]。

同时，针对硬件特点重新设计软件栈。例如 NVRAM 可以取消检查点机制，全闪存阵列去除文件系统栈，直接提供应用层控制等。

在企业市场，一体机（appliances）被广泛接受。一体机预装、预优化的软件，容易配置，硬件根据软件需求做特定设计，软件最大可能地发挥硬件能力。典型案例有 SAP HANA，甲骨文的 Exadata 和 Exalytics。还有专门针对视频处理的一体机，针对具体垂直领域的一体机（如智能交通一体机）等。在大尺度范围水平扩展（scale out）得到普遍认同的同时，在企业应用场景内以一体机为代表的垂直扩展（scale up）仍有其价值。

机柜层面，推动硬件重构，以国外 Open Compute 的 Open Rack 和国内“天蝎”为代表，而 Intel Rack Scale Architecture（RSA）是重要的支撑平台。RSA 的理念是：机柜内计算、内存、存储和网络资源分离（disaggregation），分别形成资源池，通过硅光（或 PCIe、40G/100G 以太网）实现低延迟、高带宽的互联，通过硬件层 API（如 redfish）和软件实现可重构（recomposable）的架构，针对大数据和云计算的不同负载场景和应用组合实现灵活性和资源的高利用率。

基础设施层面，软件定义兴起。从本质上讲，软件定义是希望把原来一体化的硬件设施拆散，变成若干个部件，为这些基础的部件建立一个虚拟化的软件层。软件层对整个硬件系统进行更为灵活、开放和智能的管理与控制，实现硬件软件化、专业化和定制化。同时提供统一、完备的 API，暴露硬件的可操控部分，实现硬件的按需管理。软

㊀ https://databricks.com/blog/2014/10/10/spark-petabyte-sort.html

㊁ http://petuum.github.io/

件定义基础设施主要包括硬件的三个层次网络、存储和计算。软件定义网络强调控制平面和数据平面的分离，在软件层面支持了比传统硬件更强的控制转发能力，实现数据中心内或跨数据中心链路的高效利用。软件定义存储同样将存储系统的数据层和控制层分开，能够在多存储介质、多租户存储环境中实现最佳的服务质量。软件定义计算将负载信息从硬件抽象到软件层，在异构数据中心的 IT 设备集合中实现资源共享和自适应的优化计算。

3.3 大数据存储与资源管理

在存储方面，重点关注分布式文件系统和分布式数据库的动态进展，尤其对图数据的存储和查询。值得一提的是，内存存储是反复出现的模式。就资源管理而言，本节论述了资源管理的现状以及几大重要问题。

3.3.1 分布式文件系统

当前大数据领域，分布式文件系统仍然以 HDFS 为主。经过几年的发展，对分布式文件系统的需求已经有了很大的变化。一方面，数据中心的资源开始统一管理调度，存储系统从多个分离的单一目的的小集群慢慢转换成统一的大集群，对存储系统的容量上限、存储的空间效率、访问控制和数据安全有了更高的要求。另一方面，用户对存储系统的使用模式发生了很大的变化。之前是周期性的批式应用，现在是交互性的查询和实时的流式应用。之前是单引擎的处理，现在是多引擎综合的交叉分析，需要更高性能的数据共享。以下甄选分布式文件系统的几个最新发展进行介绍。

首先是分布式文件系统缓存 HDFS Cache。现在内存计算大行其道，用户希望能够交互式地完成数据的采集、探索、分析和可视化全过程，对数据访问延时的容忍度极低。HDFS 新实现的 HDFS Cache 功能能够允许用户把某些常用数据块 pin 在堆外内存中，一方面可以增加数据带宽，减少延时；另一方面，可以用于不同应用间的高速数据共享。现在一般推荐单机内存配置 128GB 以上，很大一部分原因是我们希望能把内存当成缓存使用。

Hadoop 栈中大部分应用都是运行在 Java 虚拟机上，有些应用非常占内存，比如 Hbase RegionServer，一个进程要用几十 GB 的内存空间。实践中发现，在运行 HBase 等服务时，JVM 有比较长的 GC 停顿，影响服务质量。现在在整个 Hadoop 生态圈里，有一个现象，越来越多应用开始使用堆外空间来保存数据。比如 HDFS Cache 功能使用堆

外空间保存文件缓存，Spark 生态中的存储层 Tachyon 把数据存在堆外的 RAMDisk 里面，MapReduce 的 NativeTask 通过 JNI 在 JVM 堆外的内存缓冲区做 Map 输出数据的排序，HBase 社区正在讨论把 BlockCache、Memstore 等功能挪到 Java 虚拟机的外面以提高效率。

与之相关的是内存文件系统。在 Spark 数据栈内存文件系统的典型代表是 Tachyon，Tachyon 是 BDAS（Berkeley data analytics stack）的内存式文件系统，特别适合迭代式的计算需求以及多应用共享数据。

其次，支持分层的存储设备。现在，数据中心一般都有内存、SSD 和硬盘等存储设备，新型 NVM 也呼之欲出，在其之上有各类存储系统如 SAN、NAS、netfs 和 Lustre。一方面，这些设备和系统的带宽、延时和存储容量各不相同；另一方面，应用对存储的 QoS 要求也各不相同。比如 HBase 要求延迟极低，Spark 要求带宽大，MapReduce 要求容量大。因此 HDFS 推出新功能 heterogeneous storages（HDFS-2832) 以支持异构的存储设备，适用不同应用的存储需求。在性能方面，MIT 的 TupleWare 借鉴 HDFS 理念构造的大数据存储系统，充分利用了底层架构的向量化支持（SSE4）、超线程、分支预测和局部性优化。

再次，加密文件系统。现在的典型部署是一个大集群容纳所有用户，由此带来的问题就是数据安全。HDFS 有一个新功能是加密式文件系统（HADOOP-10150），使用 AES-CTR 加密算法，能够透明地对 HDFS 上的文件块加密解密，对应用没有侵入性，并且只有很小的性能损失，单核 CPU 加密速度能到 600MB/s，解密速度能到 1600MB/s。

考虑到数据持久性，HDFS 的典型配置是每个文件块需要存 3 份，空间利用率只有 33%。Erasure coding 被认为是一种更好的选择：每一个文件块，只需要 1.4 份（典型配置下）拷贝，空间利用率达 71%。GFS2 采用了基于 Reed Solomon code 的 Erasure coding 实现。开源社区也有诸多版本，包括 HDFS-503、Facebook 的 HDFS-RAID 等。Erasure coding 的计算要求较高，因此多用于冷数据。另外，它会导致较多的额外网络流量，目前也有一些方法来缓解这个问题，如微软 2012 年 USENIX 的最佳论文提出通过 Local Reconstruction Codes 来减少带宽要求。

最后，存储需要更好地支持上层处理。比如列式存储把（半）结构化的表，按照各个列分别存储。谷歌的 Dremel 就是一个很好的案例，说明在大数据 OLAP 领域，列式存储可以减少磁盘读取，加快分析查询速度，提高存储压缩率。在开源领域，以 ORCFile 和 Parquet 应用最为知名，一开始它们只被用于 Impala 和 Hive 等有限项目，

2013 年以来，它们的影响力扩展到了整个 Hadoop 生态，越来越多项目开始使用它们，比如 Crunch、Drill、MapReduce、Pig、Spark、Tajo 和 Cascading 等。另一个案例是 Apache UIMA（Unstructured Information Management Architecture）。它是一个用于分析非结构化内容（比如文本、视频和音频）的组件架构和软件框架实现，允许插入定制的分析组件，并将它们与其他组件交互。UIMA 应用程序不需要知道分析组件共同合作生成结果的细节，集成和组织多个分析组件是 UIMA 框架的工作。IBM 沃森采用了 UIMA 架构。

3.3.2 分布式数据库

NoSQL 数据库摒弃了关系模型的约束、弱化了一致性的要求，从而获得水平扩展能力，支持更大规模的数据。其模式自由（schema free），不再坚持 SQL 查询语言，因此催生了多种多样的数据库类型，目前广泛接受的是类表结构数据库、文档数据库、图数据库和键 – 值存储。

类表结构数据库是最早出现、在模式上也是最接近于传统数据库的 NoSQL 数据库，但多采用列存储。其源头是谷歌的 BigTable，并且在此之上发展出 HBase、Hypertable、Cassandra 和着重安全的 Accumulo。

文档数据库的数据保存载体是 XML 或 JSON 文件，从而能够支持灵活丰富的数据模型。一般文档数据库可以通过键值或内容进行查询。MongoDB 是典型的文档数据库，也是 DB Engines 数据库排行榜中排名第一的 NoSQL 数据库（前十当中只有两个 NoSQL 数据库，另一个是 Cassandra）。

图数据库主要关注的是数据之间的相关性以及用户需要如何执行计算任务。图数据库按照图的概念存储数据，把数据保存为图中的节点以及节点之间的关系，在处理复杂的网络数据时，重点解决了传统关系数据库在查询时出现的性能衰退问题。图数据库以事务性方式执行关联性操作，这一点在关系型数据库领域只能通过批量处理来完成。在实际应用方面，虽然 Facebook 是典型代表，图数据的应用领域已经远远超过社交网络领域，在地理空间计算、搜索与推荐、网络 / 云分析以及生物信息学等各种大数据应用的热门领域，都已经具有广泛的应用，商业价值日益凸显。

Neo4j 是一个基于 Java 的、高性能的、开源的图数据库。2014 年 Neo4j 2.0 发布。新版本增强了数据模型，增加了标签的概念以及可选的模式信息。最值得关注的就是全新的浏览器 UI，用户可以通过图或是表格形式来查看和编辑数据。此外，进一步完善了

Cypher 的语法，使它的语义更加清晰。AllegroGraph 是一个为管理和查询资源描述框架（RDF）而开发的图形数据库。目前新版本的 AllegroGraph 针对 SPARQL 查询优化器、性能提升和内存管理等诸多方面进行了改进和优化。InfiniteGraph 是由 Objectivity 公司推出的商用、分布式、可伸缩的图形数据库。InfiniteGraph 基于 RAM 和磁盘，通过灵活配置算法实现更快速的查询，拥有高级高速缓存的高性能查询服务器可自动分配导航工作量，工作负载可扩展至 PB 级数据。2014 年 10 月，全球领军超级计算机公司 Cray 公司将超级计算技术与 Hadoop 和 Spark 生态系统的技术结合在一起，推出了 Cray Urika-XA 系统，而原先拥有的图数据分析系统 YarcData Urika 设备已改名为 Cray Urika-GD 系统。这意味着 Cray 能够同时向客户提供易用的图数据分析一体机（Appliance）以及支持各种高级分析应用的更加灵活开放的大数据开源平台，这一举措非常适合现阶段各行业对大数据分析的迫切需求（例如金融、电信、媒体、体育、政府、生命科学 / 医药），同时迎合了追求“快数据”分析性能的企业对未来大数据架构扩展性的较高要求。

键 – 值存储因其易用性和普适性形成了 NoSQL 家族中最大的一个分支。键 – 值是最简单的一种数据模型，在此之上可以实现更丰富的数据模型。目前，基于不同一致性和存储介质（内存、SSD 或硬盘）形成了很多的选择。比如，亚马逊 Dynamo 以最终一致性为主，而 Berkeley DB 则保证串行一致性；Memcached 和 redis 是基于主内存的，而 BigTable 一族则是基于磁盘的。

内存键 – 值存储是发展较快的一支，比如 Facebook 历时多年提升 Memcached 的可扩展性。MICA 是一种新的针对多核架构的内存键 – 值存储，在数据划分、并行优化、轻量网络栈和数据结构设计方面做了整体的优化，最高达到每秒 7690 万次访问操作。

在诸多内存键 – 值存储系统中，RAMCloud 特立独行，由斯坦福 John Ousterhout 领导的团队提出。它是由成千上万台普通服务器的主存所组成的大规模存储系统，所有信息在任何时候都存储在这些快速的 DRAM 中，内存取代了传统系统中的硬盘，而硬盘只作为备份使用。其目标是同时实现大规模（100 ~ 1000TB）和低延迟（同一数据中心应用程序访问少量内存云数据只需 5 ~ 10ms，比目前系统快 100 ~ 1000 倍）。

Stonebreaker 指出，传统的关系数据库因为其总体架构的原因，在日志、加锁、缓冲区管理等方面都有很大的开销，如果 RDBMS 能够在设计时想办法避免掉这些开销，性能会大大提升。同时还指出，实际上 NoSQL 正是因为避免了上述开销，才获得了性能上的提升，那么关系型数据库同样也可以通过上述方式来获得性能上的提升。

此外，人们对 NoSQL 运动的理论支持之一的 CAP 理论也有了新的认识，Eric

Brewer 发表了一篇名为《 CAP 理论十二周年回顾：“规则”变了》的文章，他指出 CAP 理论中的“三选二”结论太过简单粗暴，在现实生活中的情况要复杂得多。首先，在同一个数据中心，分区的情况很少出现，意味着在系统不存在分区的情况下不一定要牺牲 C 或者 A ；其次，C 和 A 之间的取舍可以在同一系统内以非常细小的颗粒度反复发生，其取决于特定的操作，特定的数据乃至特定的用户；再次，这三种性质都不是非黑即白的，每个属性都有多种度量。这个前提下，CAP 理论的应用会更加复杂，Eric 提出，CAP 要在大部分时候来允许完美的 C 和 A ；当分区存在或者可以感知时，需要定义一种策略来去探知其存在并根据 CAP 理论的指导来对其进行处理。换句话说，创建一个 CAP 全都有的系统是可能的。

在 NoSQL 提出的 4 年之后，来自 451 Group 的 Matthew Aslett 在 2011 年提出了 NewSQL 数据库的概念。NewSQL 特指市场新出现的一批既能提供近似于 NoSQL 的性能和可扩展性，又能提供类似于传统关系数据库那样的关系模型、事务和 SQL 语言接口的新系统。从架构或者实现角度，NewSQL 系统可以分成三大类：

使用全新的架构。这又可以分成两类，第一类系统一般使用 shared-nothing 架构，所有的节点都具有处理事务的能力，系统具有近似线性的扩展能力。其可以是通用的数据库，比如 Google Spanner，或者为某种特定场合设计的数据库，比如 VoltDB。第二类系统则使用主从架构，有专门的节点来进行事务处理。这种设计使得系统的扩展能力会受到一定的限制。

各种 MySQL 存储引擎。MySQL 是个高度可扩展的架构，人们可以根据特定的应用场景来为 MySQL 编写各种存储引擎，这里比较出名和成熟的有 TokuDB、MemSQL、ScaleDB 等。实际上，最新版本的 MySQL 6.5 既支持传统的关系数据模型，又支持键 – 值对数据模型，此外还支持 Memcached 的访问协议。

透明数据分区技术。这和我们前面提到的 Cobar 很相似，能够自动地对数据进行分区并进行分布式事务管理，这方面的例子有 dbShards、Scalearc、ScaleBase 等。

作为 NewSQL 的一种主流，内存数据库凭借其优越的性能成为新宠。它的快速不仅在于内存读写比磁盘读写快，更重要的是，从根本上抛弃了磁盘数据管理的许多传统方式，设计了全新体系架构，并且在数据缓存、快速算法、并行操作方面进行了相应改进，从而使数据处理速度一般比传统数据库的数据处理速度快很多，一般都在 10 倍以上，理想情况甚至可以达到 1000 倍。

内存数据库主要有两类：一类是传统数据库加上内存选项，如 Oracle 12c（包括

Exalytics 和 Exadata）、IBM DB2 带 BLU 加速，以及微软 SQL Server 2014 等；另一类是完全另起炉灶设计的新型数据库，包括 Altibase、MemSQL、VoltDB、EXASOL、H20 和 SAP HANA 等。

分布式内存数据库在极限事务处理上已经成为一种关键性技术。Gartner 将极限事务处理定义为一种为事务型应用的开发、部署、管理和维护提供支持的应用模式，特点是对性能、可扩展性、可用性、可管理性等方面的极限需求。例如铁路售票系统：百万级的并发用户访问、每秒数以千计的并发事务处理、灵活的弹性与可伸缩性、低延时及 $7 \times 24 \times 365$ 可用性。

在数据库的优化上，如何更好地利用架构的能力仍是个重要问题。SAP HANA 采用了很多底层优化技术，比如 SIMD 算法实现在压缩数据的直接扫描、缓存优化的 B+ 树和 B– 树等。MIT 数据库组在研究如何将数据库（H-Store）扩展到 1000 核的架构上，在硬件上对并发控制、数据传输和时钟同步等方面提出了新的要求。另外，NVRAM 的出现也将对内存数据库架构产生革命性的变化。

传统上，科学计算和数据库管理系统是两条不同演进路线，而如今也有融合的趋势，主要体现为并行数组 DBMS（array DBMS）。目前得到最多关注的是 SciDB[㊀]。SciDB 是一个开源的科学领域的数据库，由 Stonebraker 的研究团队开发出来的。它的设计初衷旨在解决科学研究中数据量大、数据世袭等科学问题。SciDB 使用数组数据模型。数组是大多数科学数据的自然数据对象。SciDB 支持两种查询语言（AFL 和 AQL）和 3 类连接操作（等值维度连接操作、不等值维度连接操作和属性值连接操作）。SciDB 支持用户定义数据类型、用户定义函数、用户定义聚合操作和用户定义数组操作。SciDB 的性能比 RDBMS 快两个数量级。另一个相关工作是 TileDB，作为一个针对数组数据做优化分块（tiling）策略的存储管理器，也将发展成为完整的分布式 DBMS。它针对物理世界数据的高度时钟偏移和稀疏性，实现了非规则分块的策略，从而达到更高效的存储和负载均衡。

谷歌仍然推动着超大规模广域数据库研究的前沿，其代表作是 Spanner，一个可扩展的、全球分布式的数据库。在最高抽象层面，Spanner 就是一个数据库，把数据分片存储在许多 Paxos 状态机上，这些机器位于遍布全球的数据中心。复制技术可以用来服务于全球可用性和地理局部性。客户端会自动在副本之间进行失败恢复。随着数据的变化和服务器的变化，Spanner 会自动把数据进行重新分片，从而有效应对负载变化和处

㊀ http://www.scidb.org/

理失败。Spanner 被设计成可以扩展到几百万个机器节点，跨越成百上千个数据中心，具备几万亿数据库行的规模。应用可以借助于 Spanner 来实现高可用性，通过在一个洲的内部和跨越不同的洲之间复制数据，保证即使面对大范围的自然灾害时数据依然可用。

最后，值得一提的是在上述的 NoSQL/NewSQL 数据库上衍生出很多针对特定用途的数据库。如 OpenTSDB 和 KairosDB 是基于 HBase 和 Cassandra 的时间序列数据库。

3.3.3 资源管理

资源的本质是竞争性的，资源管理的本质是困难的情况下，在一系列约束条件下，寻找可行解。不同类型资源的应用一起部署可以提高总体资源利用率。以 Hadoop 为例，其资源管理目前主要通过两个层次：一是通过虚拟化，二是基于 YARN 或 Mesos 这样的更上层的资源管理层。

第一个层次是虚拟化。灵活的虚拟机操作使得用户可以动态地根据数据中心资源创建、扩展自己的 Hadoop 集群，如有需要也可以缩小当前集群、释放资源支持其他应用。通过与虚拟化架构提供的 HA、FT 集成，避免了传统 Hadoop 集群中的单点失败，再加之 Hadoop 本身的数据可靠性，为企业大数据应用提供了可靠保证。

但同时虚拟化对 Hadoop 栈的效率造成了损失，一方面 IO 密集负载的虚拟化效率较低，另一方面因为虚拟机会迁移，Hadoop 强调的“本地性”有可能被破坏。但这些问题在 VMWare 贡献的 Hadoop Virtual Extensions（HVE）中都有所解决。另外，随着 Docker 的流行，Hadoop 也开始运行于 Docker。

第二个层次是资源管理层。从业界使用分布式系统的变化趋势和 Hadoop 框架的长远发展来看，MapReduce 的 JobTracker/TaskTracker 机制需要大规模的调整来修复它在可扩展性、内存消耗、线程模型、可靠性和性能上的缺陷。在过去的几年中，Hadoop 开发团队修复了一些漏洞，但是最近这些修复的成本越来越高，这表明对原框架做出改变的难度越来越大。

为从根本上解决旧 MapReduce 框架的性能瓶颈，促进 Hadoop 框架的更长远发展，Hadoop 的框架完全重构，独立出一个资源管理器，命名为 YARN。其根本的思想是将 JobTracker 两个主要的功能分离成单独的组件，这两个功能是资源管理和任务调度 / 监控。新的资源管理器 YARN 全局管理所有应用程序计算资源的分配，每一个应用的 ApplicationMaster 负责相应的调度和协调。一个应用程序无非是一个单独的传统的

MapReduce 任务或者是一个 DAG（有向无环图）任务。ResourceManager 和每一台机器的节点管理服务器能够管理用户在那台机器上的进程并能对计算进行组织。Mesos 采取了类似的分层资源管理方法。

目前来看，还不存在一个足够好的调度器能解决下面所有问题，资源管理和调度面临着很多的挑战。

1. 公平问题

首先是如何避免饥饿。饥饿是指在竞争的环境下，一个作业永远得不到资源调度，申请不到资源。如果一个任务的需要 2GB 内存，但机器每次只能释放 1GB 内存，而且很快被别的任务占用，这就会导致当前任务饥饿。YARN 的实现思路是通过预留资源的方式。当 1GB 内存被释放，会被预留的任务占位，继续等待下一个 1GB 的资源，等齐后再执行任务。但这个方法还是很难解决活锁的问题，如果两个应用都申请了部分资源，等待其他资源才能启动，互不相让怎么办？ YARN 能够检测申请了但未使用的容器，在其超时后进行回收，但超时会影响资源使用的效率，而且无法在理论上避免活锁。还有一种解决办法是把这类应用分配在不同的优先级队列中，隔离起来。

其次是异构环境下的调度。数据中心服务器升级比较快，不同服务器的 CPU 的世代可能不同，计算能力也不同，对任务的调度应该考虑到这层因素，有所区别。现在 YARN 的调度逻辑中，并没有 CPU 的差异一说。当前的临时方案是要求用户在节点上配置一个“ vcore”数，如果该节点计算能力强，就多配一点“ vcore”数。这种方法缺点比较大，因为会造成对物理 core 的竞争，影响 CPU Cache 的有效性。

再次是数据倾斜。不同任务对计算资源的容量要求可能相同，但由于数据倾斜的存在，对资源的占用时间并不相同。另外，由于硬件资源虽然有 cgroup 隔离，但在网络、cache、disk IO 方面仍然是共享的。因此，理想情况下，调度器除了根据静态的 CPU 和内存调度外，最好还能感知节点的动态负载情况，包括网络、磁盘、CPU 利用率、实际可用内存等，避免把新的任务调度到繁忙的节点。

AMPLAB 的 Sparrow 调度器有另外一个解决的思路，思想是把一个大的作业尽量打碎成微小的任务，这样虽然微小任务间仍然有数据倾斜的问题，但因为任务数多，每个工作节点的负载还是大致均衡的，不增加整个作业的延时。

最后是基于费效比的调度。调度不仅仅要考虑怎样快，还要考虑怎样省钱。比如弹性增加计算节点时，启动一个计算节点的收益是不是高于支出？比如当前机器已经负载很高了，再调度性任务会不会导致功耗大增？比如调度怎样和数据中心的物理布局配

合，让整个数据中心更节能？比如如何在跨数据中心上调度，在峰时利用其他中心谷时的计算资源？

2. 灵活性问题

要让分布式应用能运行在 YARN 上，需要写一个 AppMaster，现在的问题是写 AppMaster 过于复杂，每个 AppMaster 都需要实现相似的功能，比如节点通信、资源弹性申请、错误检测和处理等。为解决这个问题，思路是实现一层通用的 AppMaster 层，其他应用的 AppMaster 在这层通用的 AppMasster 之上实现。实现这种思路的项目包括 Linkedin 的 Helix、Apache 的 Hoya，以及微软的 REEF。现在这三个项目都已经贡献到 Apache 基金会名下。

3. 调度延时问题

YARN 的资源管理器是单中心的调度器，所有工作节点（Node manager）通过心跳和资源管理器通信，工作节点相互间没有通信，当集群达到一定规模的时候，资源管理器中心的压力会很大，调度延时会变大，作业启动的开销是 O（N）（N 是节点数量）。SLURM 是 HPC 领域的经典调度器，Intel 的 Ralph Castain 2013 年把它移植到 Hadoop 上，它采用了不同的通信模型，Slurm 的原理是允许工作节点间可以相互通信，这样调度决策能够从中心节点瀑布式地传遍集群，作业启动开销只有 Log（N），在 1000 个节点的 Hadoop 集群中，SLURM 可以在 20 毫秒成功启动作业，而 YARN 则需要数分钟。

AMPLAB Sparrow 的调度器采用了另外一种策略，它采用了多中心节点的架构，中心节点间互相独立，通过 Batch-Sampling 来感知工作节点的繁忙程度，也能实现很好的伸缩性。根据其 SOSP’13 论文表示，Sparrow 能在毫秒级延时下，一秒钟能够调度数百万任务。当 task 运行时间小于 3 秒时，Sparrow 已经优于现在的 Spark 调度器；当小于 1.5 秒时，则远为优胜。当然，作为一个研究项目，目前其尚未出现在 Spark 的路线图。

以上是调度器层面的解决办法，在 YARN 的框架之上，为了减少延时对服务质量的影响，业界也有一些解决方案。有两种思路，一种思路是通过预先分配资源的方式以 Daemon 运行，这样任务没有启动开销，Storm 在 Yarn 上的运行就采用这种模式。另一种，是在 YARN 基础上另外加一层调度器，新的调度器只需要管理有限的资源和作业，这样回避了 YARN 资源管理器中心节点的性能瓶颈，Cloudera 的 LLAMA 调度器就采用这种策略。

4. 隔离性问题

为了简化管理，增加资源利用率，我们通常会把多个线上服务部署在同一个 Hadoop 集群上，但这样的话资源是共享的，怎么保证应用之间不互相影响呢？举个例子，HBase 服务的 Region Server 很依赖于操作系统的系统缓存来缓存频繁使用的 HFile，如果系统缓存被清除，HBase 将不得不从物理磁盘读取，延时是数万倍，HBase 的服务质量将大大降低。现有的 YARN 调度器是通过 cgroup 的方式隔离应用，但现在只能隔离 CPU 和内存，对文件系统缓存以及 CPU 的内部缓存并没有很好的办法，而恰恰这两个缓存非常影响性能。2013 年的 Spark Summit 上，Yahoo 的工程师 Tim Tully 给出了 Yahoo 的解决方案，方法是在主集群之外设置一个卫星集群，资源和主集群物理上隔离，虽然简单粗暴，但非常有效。

应用隔离还有一个挑战是用户的安全隔离的问题。现在 Hadoop 的服务基本是采用 Kerberos 做访问控制，但 Kerberos 的 Provisioning 过程非常复杂，Intel 的 Hadoop Rhino 项目提出了 TokenAuth，用轻量级的 Token 在服务间做认证和授权。

YARN 用 cgroup 来隔离应用，功能有限，现在只能隔离 CPU 和内存，还不能实现本地文件系统的完全隔离。社区里面（YARN-1964）正在开发让 docker 运行在 YARN 下面，每个容器都能跑在一个完全隔离的虚拟操作系统里面。有两个好处，一是同时实现了网络、进程号、文件系统等隔离，第二个好处是每个虚拟操作系统都能拥有自己的软件环境，互相之间不会冲突，这样服务的部署会更加灵活安全。

3.4　大数据计算框架与范式

本节主要讨论大数据计算的常用框架或范式，涉及流计算、图计算以及以 Spark 为代表的计算框架最新进展，然后讨论不同计算类型的融合，最后讨论编程模型。

3.4.1　计算范式

托马斯·库恩指出科学革命的结构起始于前科学，或前范式科学，而范式的诞生使其进入常规科学。范式是一系列公认模型的集合体。大数据经过多年的积累和发展，现在已存在一些具有代表意义和普遍性的模式、抽象和范型。2013 版发展报告已经对此有所阐释，这里我们对这些范式作一概述。

目前，以数据流为代表的范式应用广泛。其中包括数据动、计算不动的流式计算，以及数据不动、计算动的批量计算。对于后者，又分为二阶段的 MapReduce，三阶段的

BSP 和 DOT，更复杂的 DAG 和多迭代计算。Hive 查询分析被翻译成逻辑计划时，就是一个有很多 MapReduce 作业组成的任务树。机器学习因为其迭代性在 MapReduce 上面临更严重的问题。Tez 提供了基于 DAG 的表述，从而极大简化了 Hive 任务。Spark 也同样基于 DAG，在多迭代之间利用内存交换数据，与 MapReduce 相比有极大的提升。

另一个维度是并行性，数据并行对大体量数据而言是最容易、应用最广的并行，而流式计算是任务并行，更准确说是流水线并行。流式计算又分单记录和小批量，其中小批量又引入了数据并行。

数据流表现为计算图，节点之间有数据依赖，同节点的数据间独立运算；而图计算工作于具有图结构的数据之上，数据间有局部的计算依赖，同时多迭代又呈现各个计算步骤之间的数据依赖。大数据为图计算带来的机会是图并行：因为图结构的稀疏性，多数顶点只有较小的邻域，这样，在一个迭代中整个图可以划分为很多并行处理的邻域。图并行中的计算依赖必须有一致性保证，如 GraphLab 的可变一致性模型。

图计算并不是同时具有数据和计算依赖的唯一场景。对任何结构化数据模型，只要一个计算步中模型内部不同数据间存在计算依赖就适用，如滑动窗口的卷积。对此，分布式计算中最大的挑战是计算跨服务器的划分，具体来说：一是划分尽量均匀，应用散列划分和图的切割时注意数据偏移；二要通过镜像、最小切割等机制减少跨服务器依赖和通信；三是通过轻量级的任务调度和通信，如 U.T.Austin 的 Galois 可高效实现图计算。

另外，计算模型的不同颗粒度和同步 / 异步模型也会影响性能。如 GraphLab 的异步、细粒度显著加速多迭代机器学习算法的收敛，内存消耗也较小。异步、细粒度的单记录流式计算的延迟也低于同步、粗粒度的小批量流式计算。最近细粒度任务作为一种底层实现机制颇受青睐，例如伯克利的 Sparrow/Tiny Tasks 和上述 Galois。

3.4.2 流处理

低延迟或实时处理已成为大数据平台的差异化竞争优势。事件流处理（Event Stream Processing）、复杂事件处理（Complex Event Processing）和消息代理（Message Broker）是伏笔，以 Apache Flume 和 Facebook Scribe 为代表的日志数据收集、汇聚和处理是开场白；随后 Storm 和 S4 等典型流计算系统的次第兴起；时至今日，谷歌的 MillWheel、LinkedIn 的 Samza，以及流式 / 批量计算一体化的 Spark Streaming 和 Twitter Summingbird 逐一涌现。

Stonebraker 在大数据时代来临的前夜，发表了一个重要见解：一刀切（one size fits

all）的数据处理架构将寿终正寝，在流处理、数据仓库、数据库和科学数据库等方面会出现专用化的引擎，比传统通用数据库快一到两个数量级以上。他身体力行地实践了这个理念，分别创建了 StreamBase、Vertica、VoltDB 和 SciDB。他特别提出了实时流处理的八大要求：

- 流内（In-Stream）计算：数据长流不息，无须持久存储即可计算；
- 支持高阶“StreamSQL”语言，以 SQL 为主，内置可扩展的流原语和算子；
- 容忍流的破缺，妥善处理数据的延迟、乱序和丢失；
- 输出结果可预期、可重复；
- 整合已沉淀数据（持久存储数据）和在流数据；
- 保障数据安全性和可用性；
- 可划分、可水平扩展；
- 实时处理、实时响应。

接下来，从五个方面对流处理进行讨论：编程、容错、效率、可视化以及应用。

1. 编程

第一个方面是流的抽象。所有流处理系统都是一种数据流模型，计算各就各位，数据恒动，流过各计算节点，形成计算图。流处理编程抽象就是如何去描述数据流、计算以及图。Storm 通过 topology（拓扑）描述图，贯穿于其中的数据流就是 stream。谷歌流系统 MillWheel 允许动态拓扑。

多数流系统都是 record-at-a-time 式的，即对节点的执行函数来说，一次从流中取出一个记录进行计算。相比之下，Spark Streaming、Summingbird 和 Storm 的 Trident 抽象是 small-batch-at-a-time，容错机制较为简单，容易与批量计算结合。

第二个方面是编程语言。Stonebraker 提出要用 SQL 扩展，并且领导开发了 StreamSQL。据 Strata 2013 年的 Data Science Salary Survey 的报告称，SQL 是数据科学家最常用的数据工具。大数据的分析工作最终还是要回归到数据科学家的手上来，因此流平台对 StreamSQL 的支持是必需的。对 StreamSQL 的支持包含两部分内容，第一部分是指把一个 SQL 语句翻译成流处理的计算拓扑，第二部分是支持流计算领域特有的 windows、select、join 等。现在在这几个主流平台上，都没有对 Stream SQL 的完备支持。Intel 在 2014 年的 Spark Summit 上介绍了在 StreamSQL 的验证性工作，能把一个 SQL 翻译成一个 spark streaming 的作业，但现阶段还不支持 windows 计算等。

在实际应用中，SQL 语言作为一种声明式（declarative）语言，在表达复杂算法

时力有不逮。因此，主要的流系统都支持命令式（imperative）语言。Storm 支持多种语言，Summingbird、Samza 和 Spark Streaming 都以 Scala 为主，同时支持可扩展到其他语言。

第三个方面是通过 DSL 实现易用性。易用性是流平台以及大数据总体平台的一个重要指标，应用开发者只需要关心对数据是如何变换的，而不用关心平台的实现细节。Twitter 推出的流 / 批量计算一体化系统 Summingbird 采用了统一的编程语言（Scala）和统一的编程抽象，下层可以支持 Storm、Spark、Scalding 等多种流和批处理平台，对上层应用层抽象出了一套公共算子，包含 flatmap、group 等运算，应用开发者可以用这套算子描述数据流和操作，以类型安全、平台无关的方式编写应用程序。除了 Summingbird 外，还有 Storm Trident、Spark streaming 等 DSL。Spark Streaming 的独特地方在于它的 DSL 算子是建立在 RDD 之上的，RDD 不仅能表达运算，还能代表数据。通过 RDD 这一层可以无缝地和 Spark 平台上的批处理引擎、图计算引擎、SQL 引擎等打通，实现数据共享，免除数据 ETL 在不同引擎间转换的烦琐过程。

在 DSL 这一层，流引擎还有很大的改进空间。第一，表达能力有限，还不能像图一样灵活地表达运算。比如 Spark streaming 的 RDD，受 API 限制，流可以有多个输入的 RDD，但只能有一个输出的 RDD。第二，优化能力有限，对 DSL 化简能力很有限。现在 Spark streaming 和 Summingbird DSL 的优化只能做相邻算子的合并等优化，还做不到 filter-pushdown，基于代价的优化等。第三，不能智能使用内存资源。比如现在 Spark streaming 需要用户手工控制是否将 RDD 缓存到内存。

2. 容错

容错首先是要能妥善地处理流的破缺。网络的抖动、系统内在的并发性，都可能导致流数据的延迟、乱序和丢失。这需要一些通用的机制。

容错机制一是世系（lineage）。数据流经计算图的每一个节点，在发生变化或产生新的数据流的同时会记录下沿途各节点的信息，这就形成了一个世系。当某一个节点出错时，对世系进行重放（replay）实现可靠的流内计算（无需存储）。最典型的代表就是 Storm。其他手段还包括 replication、logging 和 checkpointing。replication 硬件代价高，logging 恢复时间长，checkpointing 在流系统中也得到了广泛应用。MillWheel 就采用了这种机制，但它需要持久性存储。

重放的一个前提是每个节点的计算是幂等（idempotent）的：出现错误时重新执行计

算，执行两次、多次的效果跟执行一次必须是一样的。这个前提在某些系统里需要程序员的配合。

容错机制二是时钟语义（Timer Semantics）。在流系统中，有两类时钟。一类是流系统本身的时钟，通常叫 wall clock，它的一个重要功能是检查系统是否出错（超出 timeout）。

另一类时钟是数据的逻辑时钟。在流系统里常常需要做基于时间窗口的聚合，这个时间是记录产生的时间，而不是流系统收到记录时的系统时间。如 MillWheel 的外部数据源可能是遍布全球的数据中心，当它希望统计每秒钟搜索关键词的频率时，这里的“每秒”指的是搜索发生的时间，而 MillWheel 的统计计算必须在搜索记录到达后发生，这到达时间与搜索时间可能相隔很远。流系统不可能无限期地等待确实发生在那一秒、却被大大延迟的数据，因此它采用了低水位（low watermark）时间戳机制。MillWheel 只处理那些时间戳比低水位更大（即更新）的记录，当小于低水位的记录姗姗来迟时，系统会直接将其丢弃。这些丢弃记录仅占所有记录的 0.001%左右，可以忽略不计。

综上，如果计算需要每隔一段时间做些事情，就用 wall 时钟；而如果需要根据记录产生的时间窗口做聚合，那就是用低水位时钟。

最后一个容错机制是记录的投递保证（delivery guarantee）。不同的流系统可能要求一个记录被投递至少（at least）N 次，或正好（exactly）N 次，或最多（at most）N 次。绝大多数情况下，N 是 1。而且，多数系统要求“至少一次”或“正好一次”。

Storm 通过 tuple DAG 和 ack 操作来实现“至少一次”的投递，但不能实现“正好一次”的投递。另外，在计算图执行到一半时，可能会产生一些状态，当某些节点失败时，这些状态也丢失了。Storm 的 Trident 高层抽象通过事务（transaction）解决了这两个问题。

与 Storm 相比，MillWheel 实现“正好一次”投递的方法不同，系统保证每一个记录都有唯一的 ID，而持久状态写入时也会带有该 ID。当一个新记录到达时，首先检查该记录的 ID 是否重复。MillWheel 以 BigTable 或 Spanner 作为持久存储。

record-at-a-time 的流系统在容错机制上比较复杂。上面所说的机制基本上都是 upstream backup，即由起始节点维护数据缓存并在出错时重发数据。也有一些系统采用 replication 的方式，但这样会引入更大的硬件成本。另外，record-at-a-time 天生会导致流的破缺。相比之下，Spark Streaming 的容错机制是最为简单的。它不用像 Storm 那样建立一种额外的机制来跟踪 lineage，因为 Spark 天然支持 lineage，而且可以利用 Spark 运

行时的并行恢复能力快速恢复。

3. 效率

流系统最重要的卖点就是实时或低延迟。因为流数据随来随算、没有 barrier 的特性，基本上所有的流系统都能够获得较低的延迟。加上数据量相对小，计算可以在内存中更快地完成。而且流系统的调度都采用了较为轻量级的做法，不像 MapReduce 会引入至少秒级的调度开销。

延时低意味着需要尽量避免中间环节，数据避免落地。Storm 采用了更近似流内计算、不引入中间存储开销的做法。比如 Storm 0.9 版本之前是通过 Storm 外部的 ZeroMQ 消息队列来传递消息，一进一出，增加了 IO 和延时，0.9 以后开始用 Netty 直接在 Storm 节点间传数据，性能提高到 2 倍。

除了避免中间通信环节外，还需要很好地平衡包的大小和延时的关系。包由多条消息组成，包越大，网络效率越高，吞吐量越大，但生成包需要等待更长的时间，传送时间也越长，延时反而更高。一个好的流系统应该能平衡两者，根据网络状况，自适应的调整包的大小。Storm 0.9.2 根据这个思想重新实现了传输层，性能较 0.9.0 增加了 150%，延时下降到只有几毫秒。

MillWheel 的一个重要设计要求就是水平扩展性，即在更多机器加入时延迟保持恒定。Spark Streaming 的扩展性主要源于 Spark。

持久信息的写入对于水平扩展性很重要。应用中 Storm 常常搭配 Cassandra、HBase 等快速的数据库系统一并使用。MillWheel 则采用 Spanner 或 Bigtable。

扩展性的基础是划分，尤其是对流的聚合操作来说，大时间间隔或大数据划分的计算可以通过合并小时间间隔或小数据划分的计算结果获得，这叫做可加性（Additivity）。在多数情况下，流系统满足可加性，如简单的计数或求最值。某些情况下，可加性的获得需要一些额外工作。如求平均值时，不能简单地平均更小划分上获得的均值，这时需要保存更多的信息，如这个例子需要保存每个小划分的数据单位的个数。还有最复杂的情况，如计算 UV（Unique visits），上一个时间间隔的 UV 数与这个间隔的 UV 数不能简单相加，这里需要为每个时间间隔维护用户 ID 表，对不同间隔的表做交集或联合。Count-Min Sketch 是一个很好的工具，能够把不具可加性的数据转换为具可加性，在 UV 计算例中它把用户 ID 表转换为高度紧凑的统计计数器作为近似，并带来可加性。

基于流的近似计算前途光明。Summingbird 子项目 Algebird 实现了大量的近似计算 Monoid，比如 HyperLoglog、CountMinSketch、Top 等，用户可以以极少的内存对十亿级记录做统计。初创公司 Streamdrill 提供了很好的近似计算解决方案。

对线上应用，我们通常希望能随着负载变化，零挂机时间，增加弹性或减少资源。Google 的 Millwheel 在计算节点间采用 HBase 风格的 Key-range 的哈希方案。当负载变化时，可以切分或合并 Key-range，实现资源的弹性分配增长，并且不影响应用，不增加延时。

4. 可视化

在流处理系统里面，消息恒流不止，没有一个固定的状态，如果出现异常，通常很难分析具体原因。所以，对流平台而言，可视化尤为重要。可视化的目标包括两方面，第一方面是平台系统的可视化，包括网络拓扑、节点状态、网络压力等。第二方面是用户应用的可视化，包括计算图的拓扑结构、消息的流速变化、端到端的延时分布等。Google 在 2014 年的 Google IO 上，演示了 Cloud Dataflow 的流式可视化方案，能显示出用户应用的拓扑结构图、流速大小和消息延时等。Storm 在刚发布的 0.9.2 版本也新增了一个类似的功能。不过，现在的可视化实现功能还相当有限，指标很少。

5. 应用

流系统在物联网中得到了广泛应用。到 2020 年，将有 260 亿的物联网设备。现在的网比以前更大，“物”上的数据要能够融入到流平台中来，实现一体化管理。Azure 在今年 10 月宣布了 IOT 一体化流处理解决方案，打通了物联网数据流的收集，流分析处理，可视化的链条。计算还要能跨地域、跨平台。计算跨平台领域，计算要能在物联网设备上运行，就像在云端一样。计算能够按需动态部署到设备上，设备上的计算能像云端一样，以相同的抽象运行，实现计算的位置透明，消除“物”和“网”之间的鸿沟，让数据真正“联”起来。

另一方面，流处理与在线机器学习相得益彰。比如微博的垃圾评论，每天都在快速变化，而且量很大，相信不少博主深受其扰。怎么确定这些评论是不是垃圾评论？如何跟上垃圾评论变异的速度？这就是一个复杂的流处理问题，需要做实时分类，一边更新模型，一边分类。Yahoo 2013 年的新项目 SAMOA 就是这么一个流处理上的机器学习平台，现在已经支持一些分类和聚类算法。SAMOA 在下层可以支持 Storm、S4 等多种平台。

3.4.3 图计算

现实生活中的很多现象表现为图（如社交网络、路网、病毒传播）。很多科学问题的数据结构是图结构，算法为图算法，机器学习概莫能外。图计算框架简化了大规模分布式机器学习的实现。

MapReduce 的出现一度被认为是大规模机器学习的福音。但作为单输入、两阶段、粗粒度数据并行的简单数据流抽象，MapReduce 在表达复杂多迭代算法、稀疏结构（图）和细粒度数据 / 计算依赖（而非数据流中的粗粒度数据依赖）时，往往力不从心。如 PageRank，首先是多迭代直到收敛；其次顶点和边组成了不规则的图结构，边密度稀疏且分布不均；最后顶点之间的边是数据关系（邻接），也是细粒度的计算依赖（顶点的计算结果向邻近节点传播）。由于图结构的稀疏性（多数顶点只有较小的邻域），整个图可以划分为很多可阶段性并行处理的子图，这类问题也称为图并行问题，是 Amorphous Data Parallelism 的一种表现形式。

图计算的技术发展主要在编程模型和图计算引擎两方面。

在编程模型上，最早使用 MapReduce，但由于迭代间通过磁盘交换数据，形成 all-to-all 通信，非常低效。Pregel 连同其后的 HAMA 和 Giraph，采用了 BSP（Bulk Synchronous Parallel）计算模型。BSP 实现了超步（Superstep）级的同步和串行推进。每个超步有三阶段：首先不同计算单元上的本地计算并行发生，随后的全局通信阶段把每一个计算单元上的计算结果传播到其他单元，最后一个 barrier 等待所有通信完成，并结束当前的超步。这种模型编程简单，而且迭代间通过内存交换数据，数据局部性保证最少通信，与 MapReduce 相比有很大提升。但性能还是差强人意，因为“最慢那部分计算的诅咒”（与木桶理论中短板决定盛水量一样）。在图计算中，因为顶点 / 边密度的不均衡、网络、计算硬件的差异性和多租户条件下的波动性，被“诅咒”是常态。

Pregel 最早提出“think like a vertex”的顶点函数模型，所有顶点不加区分同时执行，计算基于上次迭代的状态，收敛较慢，类似 Jacobi 方法。这时候异步计算被提上日程。异步不仅拿掉了 barrier，没有同步开销，而且顶点执行有优先次序，计算基于最新状态（而非上一迭代的状态），在迭代计算中加速收敛，类似 Gauss-Seidel 方法。GraphLab 采用异步的顶点函数模型，一开始是 update 函数，后来在 PowerGraph 中演化成 GAS 模型：从邻接顶点获取数据（Data）、应用（Apply）和扩散（Scatter）。它的优势是简洁明了，各部分语义清晰，而且把整体行为拆散成各个步骤更易于做分布式调度。但缺点是：比起 Pregel 或 GraphX 把计算一步步清晰地描述在一个函数中，GAS 把计

算分散在不同函数的做法使开发者不易掌握函数之间的逻辑关系、产生整体的理解。因此，GraphLab3 WarpGraph 试图能够兼顾两者的优势。

基于 Spark 的 GraphX 重回数据并行，巧妙地在 RDD 数组结构上实现了图描述。相比 GraphLab，GraphX 在编程模型上有一定优势：一是能够描述整个图计算的流水线，从图构建、计算到后处理；二是能够很容易地与表结构进行关系操作；三是能够描述更复杂的数据、计算依赖，比如与邻居的邻居的关系。GraphX 能够灵活地实现 Pregel 和 PowerGraph 的 API，但短期内不会有异步支持。

在图计算引擎上，有以下几个主要的问题。

首先是调度。图计算希望有细粒度的调度器（如 Galois），而且能针对不同编程模型提供基于拓扑的静态调度和基于数据的动态调度。静态调度有 BSP 模型（类似 Jacobi 方法），如 GraphLab 的异步调度和 Galois 的同步调度，以及异步模型（类似 Gauss-Seidel 方法），如 GraphLab 的同质调度和轮流调度，或 Galois 的自治调度。而异步模型还可以支持动态调度，以加速收敛，包括用户自定义的顺序和应用程序规定的顺序（user-specified order（当前 vertex program 显式往 worklist 加入待计算顶点）和 application-specific priorities）等。

其次是一致性。GraphLab 允许三种不同强度的一致性，包括：全一致性、边一致性和点一致性，供应用选择。PowerGraph 进一步提供了三种模式：同步执行，异步执行但不保证一致性，异步执行且保证串行一致性。降低一致性可提升吞吐量，但可能导致收敛变慢。

再次是分布式引擎，包括分布式锁实现、容错和划分。比如分布式 GraphLab 采用了基于回调函数的非堵塞的 readers-writer 锁和基于 Chandy-Misra 的并行锁，基于 Chandy-Lamport 的完全异步的快照机制实现容错。对于划分，它通过过度划分来实现可扩展，ghost 缓存达到最小通信，vertex-cut 尽量最小化数据偏移（尤其是符合幂律分布的图）。陈海波教授在 PowerLyra 中采用 Hybrid-Cut 方法，对 low-degree 顶点采用 Edge-cut 以最小化计算、通信和同步开销，而对 high-degree 顶点采用 Vertex-cut 以最小化不平衡和竞态。

相比分布式引擎，另一种尝试是小型化，如 GraphChi 通过改变图的数据结构，对磁盘顺序、大块访问，加上内存计算，单机能够实现 Hadoop 大集群的计算能力。同时，新的数据模型还能支持图的演化。另一种单机引擎 X-Stream 采用从边的视角出发的编程模型，边无需排序因此没有预处理代价，对内存的需求更小，同样能避免硬盘的随机访问。

从总体趋势上来说，图计算有如下发展目标。

- 易编程性，如 GraphLab3 WarpGraph 所做的努力。另外，通过新的高层语言和语言接口（如 Python）以及新的工具（如 CloudCV）来进一步解决编程问题。
- 提高更好的图引擎，如更好、更快的 cut 方法。
- 支持在线 / 流式的机器学习，这需要图引擎支持图结构的演化。
- 与交互式机器学习和可视化结合，支持用户反馈和主动学习。

3.4.4 Spark 新动向

以 Hadoop 生态圈为代表的众多开源大数据系统覆盖了各种计算范式和应用场景。在一套完整的大数据流水线中，用户往往需要综合多种范式，因此必须同时部署和对接多套特化系统。然而这些特化系统在存储、调度、计算等方面的优化一般都局限于自身擅长的计算范式和场景，因此各个特化系统间只能依赖以 HDFS 为代表的分布式文件系统进行数据交换。即便各特化系统自身的吞吐和延迟相对理想，系统间数据共享造成的 IO 瓶颈仍然大大拖累了数据流水线的整体效率。针对这一问题，UC Berkeley AMPLab 抽象出了 RDD（Resilient Distributed Dataset，弹性分布式数据集）这一数据结构，并围绕 RDD 开发了 Apache Spark，为打造一体化大数据流水线打下了坚实的基础。

Spark 的秘诀在于：

通用性。围绕 RDD 定义的各种运算是 MapReduce 的超集，可以完成 MapReduce 所能完成的一切运算。

内存计算。RDD 可以在兼顾数据分布局部性的同时充分利用集群内存，通过将常用数据集缓存在内存中，达到加速以机器学习算法为代表的迭代型计算的目的；相对于 Hadoop 上的同类系统（如 Mahout），其加速比往往可以达到一到两个数量级。

线程级并行。这使得 Spark 的任务调度延迟得以降至亚秒级，为 Spark Streaming 这样的以 micro batching 为基础的流计算奠定了良好的基础。

DAG 流水线优化。与 Dryad 等 DAG 计算系统类似，RDD 丰富的运算集可以轻松表达复杂的 DAG 计算，不再需要像 MapReduce 那样为每一步操作调度一个单独的作业。再加上作业中每个 stage 内部辅以流水线优化，即便不启用内存缓存，Spark 的执行效率往往也数倍于 Hadoop。

容错性。RDD 的不可变性和世系特性使得 Spark 可以以数据分片为粒度追踪数据的历史。当集群中的节点宕机时，只需追踪故障节点负责的 RDD 分片的世系，便可重新

计算出丢失的分片，而且整个错误恢复过程可以并行执行。数据冗余在数据恢复过程中只起到加速作用。

数据共享抽象。RDD 较好地解决了大数据分析流水线中各环节的数据共享问题，避免了频繁的分布式文件系统 IO。

多范式支持。由于底层框架提供了优秀的通用性和效率保障，Spark 得以在应用层同时实现批处理、流处理、关系查询、即席查询、以机器学习为代表的迭代型计算和图计算等多种范式。而且实现各范式的组件只需聚焦于各自的问题域，无须重复解决底层框架中的分布式、容错、数据共享等共性问题，从而实现了一体化大数据流水线的愿景。

在计算层面 Spark 可以完全取代 MapReduce；但在数据存储层面，Spark 则采取了全面兼容现有 Hadoop 生态的策略。这也是 Cloudera、Hortonworks、MapR 以及 Pivotal 等老牌 Hadoop 厂商乐于拥抱 Spark 的重要原因。2014 年以来，由 Spark 创始团队组建的大数据创业公司 Databricks 和 Spark 社区一起针对 Spark 进行了诸多改进，大大改善了 Spark 在大规模应用下的稳定性和可伸缩性，主要包括：

可插拔 shuffle 接口及基于排序的 shuffle（SPARK-2045）。Spark 默认的基于散列的 shuffle 相对高效，但每个 mapper 要为每一个 reducer 输出一份文件，这些文件不仅导致了大量 IO，它们占用的缓存也会随数据量的增加而快速膨胀，从而导致内存溢出。从 1.1.0 起，Spark 引入了基于排序的 shuffle，通过将单个 map 任务的输出按 key 的分片序号进行排序，使得每个 mapper 仅需输出一份文件，大大减小了处理超大数据量作业的内存消耗。

基于 Netty 的新网络模块（SPARK-2468）。新的网络模块实现了零拷贝消息发送，并且绕过了 Java GC 自行管理内存，减少了对 GC 的影响。

外部 shuffle 服务（SPARK-3796）。该服务建立在上述的 Netty 网络模块之上，并将 shuffle 数据供给从执行中剥离了出来。这样一来，即便 mapper 端的执行陷入 GC 停顿，reducer 仍然能够读取所需的数据。

在 2014 年度的 Daytona GreySort 竞赛中，Databricks 参赛。在上述改进的帮助下，Spark 创始人及核心团队的 Matei Zaharia、Reynold Xin 等人用 Spark 赢得了并列第一。在该项赛事中，参赛队伍需要在一系列苛刻约束下对 100TB 随机数据进行排序。此前，赛事记录由 Hadoop 保持，而 Spark 在不启用内存缓存的前提下以十分之一的机器数将排序时间缩短至原记录的 1/3。

除却稳定性和可伸缩性，Spark 也在性能和功能上下了很多工夫。

在内核层面，Spark 对现有的块管理进行了较大程度的重构，并增加了对 Tachyon 的原生支持。

集群管理层面，YARN 的集成更为紧密，新增的资源分配动态调整能力可以更好地满足多租户需求。

相对于批处理，流处理的挑战更加严峻。Spark Streaming 在 HA 方面做出了多项改进，支持了更多的数据源种类，并增强了对 Flume 的支持。

Spark SQL 是 Spark 最大的 alpha 组件，自 2014 年 3 月发布以来发展迅速。Spark SQL 已经正式取代 Shark 成为新的 SQL on Spark 引擎。相较于 Shark，除了兼容 Hive，Spark SQL 可以借助 SchemaRDD 无缝对接所有 Spark 组件。SchemaRDD 成为了 Spark 结构化数据处理的瓶颈（下一步 MLlib 也将基于 SchemaRDD 之上）。新增的外部数据源 API 使得外界储存系统和数据库可以非常容易地连接到 Spark SQL 的 SchemaRDD，并且在查询时候优化器甚至可以直接把一些过滤操作发送到实现这个接口的数据库里面，最大限度地利用这些数据库自身的过滤功能减少网络传输。

在机器学习方面，MLlib 不仅新增了一个用于完成抽样、相关性、估计、测试等任务的统计库，还酝酿着一套新的 pipeline API，可以更加便捷地搭建机器学习相关的全套流水线。其中还包括了以 Spark SQL SchemaRDD 为基础的数据集 API。

自 OSDI14 亮相以来，Spark 的图计算组件 GraphX 近期的更新主要集中在性能提升方面。淘宝自 2014 年 5 月 Spark 1.0 发布时起，已经将 GraphX 用于线上生产环境。

3.4.5 范式的融合

在多种范式各行其道的同时，组合、融合和统一也成为主流。流式、批量和迭代的组合和融合，可以在不同层面发生：如 Twitter Summingbird 在编程接口层面融合，Lambda 架构在应用框架层面通过增量计算和批量缓存实现融合，Spark 在实现框架层面完成融合，而微软利用资源管理层通过 REEF 来支持多计算模型。

在诸多融合的努力中，基于 Spark 的 Big Data Analytics Stack（BDAS）风头最劲。它在同一平台上实现交互式查询（SparkSQL）、流式处理（Spark Streaming）和更复杂的运算（图计算 GraphX 和机器学习 MLlib）。在其之上可以形成不同的使用场景：

流查询。Spark Streaming 与 Spark SQL 融合成 StreamSQL，或者形成实时系统与数据仓库的结合；

实时 + 批量（类似 Lambda 结构）。融合实时洞察和历史洞察形成全时洞察，采用 Spark Streaming+Spark Core；

流处理 + 机器学习。实时获得更复杂的洞察，采用 Spark Streaming+MLlib；

图流水线（图创建 /ETL+ 图计算 + 后处理）。原来是用 Hadoop+ 特殊图引擎，现在可以用 Spark Core/Spark SQL+GraphX；

即席查询 + 机器学习。采用 Spark SQL+MLlib。

Databricks 在 2014 年 Spark 技术峰会上演示完美演绎了这一理念——利用多种计算范式完成了一个 Twitter 实时概念搜索：先是 MLlib（机器学习库）基于维基百科数据建立 TF-IDF 词模型，随即据此模型用 Scala 语言编写了一个不同词之间的相似函数，进而将此函数注册到 Spark SQL（成为它的 UDF）。准备就绪后，Spark Streaming 从 Twitter firehose 生成一个真实的 tweet 流，最后 Spark SQL 交互地过滤出与用户给出搜索词相关的 tweets，比如搜索足球会显示世界杯的 tweets。

BDAS 融合的基础是 Spark 的数据并行模型。微软也提出 Naiad，在数据并行的数据流模型之上加入时间因素、异步消息流和轻量级协同机制，从而高效地实现流处理、批量、迭代（图计算和机器学习）等计算模式。

在多范式融合的基础上，最终趋向于统一编程界面。早在 2005 年，Stonebraker 在《“ One Size Fits All”： An Idea Whose Time Has Come and Gone 》即已指出，过去 25 年追求的全能引擎概念已不可能存在，反之会存在很多独立的、针对不同应用场景的数据处理引擎。但他又指出，可以通过统一的前端解析器把这些引擎整合起来。

这对 BDAS 不难，因为所有子系统都基于 Spark，前端除了 Scala 之外，还可以通过 SQL、Python、Java、R 和 Cascading 等语言。

对于异构的子系统，需要提出更好的抽象。比如 Summingbird 通过 API 级的抽象整合 Storm 和 Hadoop。MIT-Intel 大数据科研中心有个宏伟的计划，整合其所有大数据子系统（H-Store、S-Store、SciDB 等）成为一个技术栈 Big Dawg，采用一种叫做 BQL（ Big Dawg Query Language）的统一编程语言作为“细腰”——这称为大数据的 LLVM。

这种融合的趋势一直从计算范式延伸到宏观层面。

比如，数据管理与数据分析的融合。前大数据时代两者是分开的，事务型 RDBMS（OLTP）与分析型 RDMBS（OLAP）各行其道，打通两边还需要 ETL。大数据时代融合成为趋势，SAP HANA 和 NewSQL 为代表的内存数据库平台采用列存储，在一套数据上

实现 OLTP 和 OLAP。所谓的 Big SQL 或 SQL-on-Hadoop 集大数据管理和分析能力于一身，Strata 上做的调查一致认为 SQL 是数据科学家的必备技能。还有一种思路是把分析嵌入到数据库管理引擎中，这是 MIT 数据库派的主要做法，如 S-Store 把流式处理的支持嵌入到内存数据库 H-Store。

又如，以内存数据网格（in-memory data grids）的名义，MPP 的设计理念与内存键 – 值存储和传统数据库产生融合。

还有，MPP 与 Hadoop 的融合。基于 MPP 数据库和 Hadoop 生态系统的大数据查询解决方案各有千秋：多年的发展让 MPP 数据库性能很好，但是其在可扩展性和容错性方面还存在着不足；基于 MapReduce 各种大数据查询方案足够鲁棒，扩展性也非常好，但是在性能方面还是无法和传统的 MPP 数据仓库相提并论，尤其是在数据集相对较小的情况下更是无法比较。于是，EMC 的 Pivotal HD，将大规模并行数据库 Greenplum 与 Apache Hadoop 框架集成，HP 的 HAVEn 平台整合了 Hadoop、Autonomy 语义处理引擎和 Vertica 列存数据库，而国内南大通用也实现了 GBase 8a 与 Hadoop 的集成。上述平台以 MPP 数据库及计算框架为核心，将 MPP 运算调度引擎完全融入非关系型运算调度框架，实现可以同时调度关系运算和非关系运算的调度引擎，构建统一的结构化信息提取和数据类型转换框架，将非 / 半结构化数据映射为关系模型，实现面向关系模型的全数据统一视图，从而平滑地实现 MPP 数据库和 Hadoop 的统一调度和处理。

最后，高性能计算（HPC）/ 超算（Supercomputing）与大数据也在融合之中。在第四范式的话语体系里，传统认为前者是以模拟为主，后者以数据探索分析为主。前者计算密集，后者数据密集。前者是高大上（以 scale up 为主），后者平民化（以 scale out 为主）。前者多采用 MPI、C++、CUDA/OpenCL 等编程模型，而后者采用 Hadoop（Java）、Spark（Scala）等。近年来，两者呈现融合的趋势。首先，HPC 应用也出现很多数据分析，有人提出新的子领域 High Performance Data Analysis。其次，很多 HPC 应用也是数据密集的，被称为 Data Intensive Supercomputing，另一方面，不少大数据应用也是计算密集的（如深度学习）。再次，HPC 的高端架构在下移，如光互联、RDMA、至强融核，越来越多应用在大数据场景中，另一方面，大数据和云计算的架构也在影响 HPC，我们看到越来越多 HPC 计算在云服务和大数据平台上完成。最后，就编程模型而言，MPI 和 Hadoop 的混合平台在很多互联网公司应用，2015 年的新至强融核芯片（Knights Landing）将支持完整的 X86 指令集，这意味着超级计算机也将能够运行 Java 虚拟机及其之上的 Hadoop 和 Spark 软件栈。

3.4.6　编程模型

编程模型是计算模型的实现机制。理想的编程模型不但表达力强、性能好，还要不容易犯错。在大数据技术栈中，比较广为接受的编程模型是数据并行、函数式语义、共享内存或消息机制等。MapReduce 是数据并行，并且通过上层框架实现函数式语义。Spark 也是数据并行，在语言层（Scala）和框架层都利用了函数式语义。考虑千台节点集群未来将是百万硬件线程的规模，上述编程模型还将是主流。

Java 虚拟机的丰富生态环境促进了类似 Clojure、Haskell、Groovy 这样小众语言在特定领域的勃发，这些语言多提供了函数式语义和并行 / 并发编程的支持。其中一个值得称道的语言是 Clojure，不但享受了函数式语言不易出错的优点，而且表达力强，支持动态编程、元编程和同像（homoiconicity）。有一些实验揭示，实现同样的功能其代码量为 Java 的 1/10，Scala 的 1/5。

还有一些值得一提的语言和编程模型。老牌语言 Erlang 凭借其基于 Actor 的编程模型仍然在大数据、尤其是消息驱动的架构中大有用武之地，但是基于 Scala 的 Akka 有后来居上之势。Go 是谷歌力推的语言，在代码复杂性、并行 / 并发支持和代码效率上有比较好的折中。

本节将讨论一些新的挑战和趋势。

首先，数据并行编程模型在大数据中应用最广，在查询分析和简单分析中驾轻就熟。虽然其对非规则数据结构表达力不强，对于复杂的稀疏数据结构或非规则表，往往通过数据预处理后仍能发挥作用（类似于 Dremel 的 array of structure 转 structure of array）。但是，对于非规则计算（每个点的计算行为不一样），数据并行的威力受到一定的限制。比如，在机器学习中有大量的非规则计算，这使得 Spark 在实现这些计算的时候必须有所取舍。一种方法是把数据并行蜕变为任务并行，在宏观层面各个任务仍是数据并行，微观层面把 RDD 拆开成为元素、分别计算。这是类似于多线程的做法，导致的后果是实现不简洁，代码量大为增加。Spark MLlib 的解决办法是尽量把非规则计算转化为 GraphX 的数据并行，把算法建构在 GraphX 之上，仍然维持算法的简洁性。但数据并行的问题是无法获得本迭代的其他顶点的最新更新值，可能导致收敛变慢。

相比之下，针对表达为图并行的非规则计算，以 PowerGraph 代表的 GAS 编程模型在表达能力和效率上有优势。它把整体行为拆散成各个步骤更易于做分布式调度。但缺点是：比起 Pregel 或 GraphX 把计算一步步清晰地描述在一个函数中，GAS 把计算分散在不同函数的做法使开发者不易掌握函数之间的逻辑关系、产生整体的理解。

PowerGraph2 试图能够兼顾两者的优势。

另外，编程模型要支持不同数据结构之间的关系操作，最典型的就是图结构与表结构的交互，比如，要对一张社交网络图和一张用户表做交互。在这一点上，GraphX 有得天独厚的优势，因为其图结构和表结构都是基于 RDD 的。原来 GraphLab 不支持，但近期版本加入了 SFrame 来支持表结构。

学术界在试图提出真正普适的编程模型。微软最先提出 LINQ（Language-Integrated Query），在 C# 和 Visual Basic 基础上加入查询功能，但只能支持关系代数，不能支持复杂数据模型、迭代和并行计算。MIT-Intel 大数据科研中心在华盛顿大学的 MyriaQ 的基础上设计了一个新的通用编程模型，在核心支持关系和线性代数、复杂数据模型、迭代计算和并行计算。

在实际工作中还有一个重要的选择：框架（framework）还是库（library）。业界有个框架 v.s 库的玩笑。框架是“好莱坞”范儿：你按它的规矩（API）做事，别 call 它，它会 call 你。MapReduce 就是这样。框架带来方便的同时，也限制了你自由发挥的空间，更因其粗粒度让你背上厚重的“框架”。若想来去自由（用户程序随意调用）、腾挪自如（库功能可自如组合），库是正选。库是“职业烟民”范儿：自己卷自制的烟丝来抽。

最后介绍近年来一种新的范式，reactive 范式。10 年前，企业应用只需要有限几台服务器，处理 GB 级别的数据，可以接受数小时的挂机维护；而现在的数据应用横跨各种移动设备和数据中心，动辄需要处理 PB 级别的数据，动用成百上千台服务器，毫秒级的响应速度，并且要 24 小时在线。比如，对一个手机 App 的点击操作，就可能触发后台数百台应用服务器协同工作。在变化面前，传统的命令式的、过程式编程模型工作得不好。

在新需求面前，我们需要一个天生就能伸缩、能容错的架构。Reactive 就是这么一种架构，它是反应式的，而非过程式的。它以消息驱动为核心，组件之间完全隔离，没有共享的状态，完全异步，只有异步消息一种信息通道。错误即消息，错误异常等不是通过调用栈隐式传递，而是封装成显式定义的消息传递。Reactive 架构就像我们人类社会的架构一样，人人相互独立承担自己的工作，通过消息驱动方式来分派工作，实现协同合作。互相之间依赖降到最低，独立容错。一方面，每个人都可以失败，地球不因某一个人就不转；另一方面，地球上每一个人都在运转，几十亿人分布式工作，忙碌且和谐。

Reactive 本身而言不是一个新的技术。1973 年，Carl Hewitt 在“A universal modular ACTOR formalism for artificial intelligence”一文中就对以消息驱动为核心的架构模型做过细致的理论分析。但在大数据的时代，这是一个新的命题，重新认识 Reactive 具有时代意义。一方面，大数据的分布式特点正是 Reactive 架构善于解决的问题，另一方面，在大数据领域按照 Reactive 思想建构的软件还太少，只能说是刚刚开始。第三方面，要想在终端用户的边界实现 Reactive 的属性，需要全服务栈的上下游都按照 Reactive 的方式互联，这需要整个产业链开始建立共识。

历史上，Reactive 架构有两次大发展。第一次是在电信领域，20 世纪 80 年代，Ericsson 的 Joe Armstrong 发明了 Erlang，实现了 Reactive 架构的 Actor 模型，用 170 万行代码实现 ATM 网核心交换机 AXD301，达到了 99.9999999％的在线率，成为行业的标杆。

第二次大发展是在 Web 领域，发展于上世纪末的互联网泡沫时期。从 1993 年开始的 10 年期间，互联网用户数增长了 54 倍。大型网站普遍遇到了高并发性的瓶颈。为了解决 C10K 问题，支持更多的并发连接，网站架构逐渐切换到以事件驱动为核心。这里面最负盛名的是 Nginx。Nginx 采用了事件驱动的异步非阻塞架构，成功解决了 C10k 问题，根据 W3Techs 2014 年的数据，在 Top1000 的网站中，Nginx 是最流行的网站服务器，占有率比 Apache 和 IIS 加起来还多，达到 42.1％。

Reactive 在电信和 Web 行业率先落地，原因在于这两个行业对高可用性和高并发性都有非常高的要求，更早遇到传统架构的瓶颈。

现在进入大数据和物联网的新时代，根据 IDC 的报告，全球每年的增量数据将以 140％的速度指数增长，根据 Gartner 的报告，到 2020 年，将有 260 亿的物联网设备，大规模和分布式将成为一种新常态，Reactive 架构的应用将超越电信和网站系统的范围，影响更多的垂直行业，深入生活每一处。正如 Scala 语言发明者 Martin Odersky 所说，“Reactive 编程无所不在”。这正在成为 Reactive 架构发展的第三波浪潮。

2013 年，Reactive 宣言[㊀]从定义上明确了 Reactive 架构的四个特点：

- Message driven：系统由消息推动，系统组件间建立了一个安全的边界，只通过消息交换数据，实现并发和隔离。
- Elastic：应用能充分利用计算资源 Scale up 和 Scale out。能根据负载变化弹性增加或减少资源分配。

㊀ http://www.reactivemanifesto.org/

- Resilient：应用能够容忍错误发生，仍然可以正常响应用户请求。
- Responsive ：应用在过负载和软硬件错误的情况下仍然能保持实时的、交互式的响应。

过去 5 年，Reactive 开发平台中要属 Google 的 Golang 和 Typesafe 的 Akka 最具光彩，它们都是在 2009 年前后公开，他们都是基于消息驱动的，都侧重于解决高并发问题。Golang 是基于 1978 年 Hoare 提出的 CSP（Communicating Sequential Processes）模型，思想是把应用打碎成细粒度的 Goroutine。Goroutine 是一个轻量级的过程，对外隔离自己的状态和操作。没有依赖的 Goroutines 可以独立运行，有依赖关系的 Goroutine 通过 Channel 交换数据，pipeline 化运行，提高整个应用的并发度。Akka 是基于 1973 年 Carl Hewitt 提出的 Actor 模型，Actor 和 CSP 很像，但不同的是 Actor 使用异步非阻塞的方式传递消息，而非同步的 Channel。这一设计让 Actor 模型能实现位置透明，即不论是在单机还是在分布式环境中，Actor 之间都能以一致性的方式交互，和 Actor 所在的位置无关。这让 Actor 模型能天生 Scale out。

另外，即使有些大数据系统不是用上述开发平台实现的，仍有可能采用了 Reactive 的思想，如 Storm。

2014 年以来，产业界有一些新的进展。为了实现端到端的全数据栈的 Reactive，让不同供应商的服务能够以 Reactive 的方式串接起来，来自 TypeSafe、Twitter、Oracle 等的工程师提出了 Reactive Stream 接口规范。实现这些接口的 Reactive 服务是兼容的，消息能够自由且安全地在服务间流动。

3.5 大数据分析

基本数据处理包括数据 ETL 和查询，以及简单的统计分析。其在大数据任务中最简单也最常用，MapReduce 能很好实现。Hive、Impala 和 Spark SQL 等框架的出现也解放了 ad-hoc query 的生产力。3.5.1 节介绍了大规模统计查询分析的进展。大数据给基于机器学习的复杂数据分析带来新的挑战。3.5.2 节介绍大数据影响下机器学习算法本身的变迁，即从算法的角度去更好地利用大数据。3.5.3 节涵盖了如何将大数据化小，通过精炼数据，使得数据可以适应复杂的分析方法。3.5.4 节介绍机器学习优化算法分布式执行的一般模式，使分析方法适应大数据的体量。3.5.5 节对大规模数据学习系统的现状做总结，并对其未来的发展进行思考。最后 3.5.6 节介绍了多种大规模数据分析中的实用性问题，这些问题在学术著作中提及较少，但在工业应用中却至关重要。

3.5.1　大数据的统计查询

随着 Hadoop 在各种大数据数据应用场景的广泛使用，基于 Hive 的大数据查询分析得到广泛应用。但其性能无法达到交互式使用的要求。所谓交互式使用，是指发出查询请求以后，能在数秒到一分钟以内得到结果。

要达到上述要求，在 MapReduce 计算模式之上的批处理模式显然不能再胜任了。业界开始考虑借鉴大规模并行处理数据库的很多技术，在 2010 年的 VLDB 大会上，Google 发布了著名的关于 Dremel 的论文，从实践上确认上述思路的可行性。开源社区迅速跟进，到现在已经有四个不同的实现，他们分别是来自 Cloudera 的 Impala，来自 Hortonworks 的 Stinger Initiative，来自 MapR 的 Drill 以及来自韩国的 Tajo 项目。这四个项目各有特色，而且都处于其发展的初期，到底谁会脱颖而出现在还难以断言。

在 Spark 之上，也有查询分析的实现，最早是 Shark，后被替换为 Spark SQL。Spark SQL 采用了令人耳目一新的前端设计，其核心是一套名为 Catalyst 的关系查询库。在 Catalyst 中，表达式、逻辑执行计划、物理执行计划都被抽象为不可变的函数式的树结构，查询计划的优化策略则被抽象为规则，并实现为针对上述树结构的模式匹配和函数式变换。由于 Spark 的实现语言 Scala 对函数式范式的优良支持，基于这两种抽象来表达各种查询计划优化策略十分清晰简洁。同一优化策略，Spark SQL 实现的代码量往往只有 Hive 的十分之一到五分之一左右。这一点从根本上促成了 Spark SQL 的快速演进。

从查询分析的语言来看，基本上是从类 SQL 语言（如 HiveQL）向完整的 SQL 语言（如 SQL92）发展，进而对其扩展，如 StreamSQL。一些研究团队，如 MIT-Intel 大数据科研中心，也在试图设计新的查询语言，能够把关系语义和数组语义进行融合。

另外还有大规模的图数据查询分析。Cypher 是一个由 Neo4J 采用的描述性的图形查询语言，用户无需编写遍历代码就可以对图进行富有表现力和效率的查询。Cypher 的设计原则是可读性和内部一致性。它被设计成一个便于开发者，尤其是即席查询的操作者来使用的一门语言，尽量做到简单化。Cyper 灵感来源于一系列不同的方法和实践，许多关键字如 like 和 order by 是受 SQL 的启发。模式匹配的表达式来自于 SPARQL。正则表达式匹配借鉴了 Scala 语言。作为一个描述式的查询语言，Cypher 关注的是从图中检索什么，而不是怎么去检索，这和命令式语言如 Java 以及脚本语言 Gremlin、JRuby 等明显不同，这样，用户无需了解实现细节，也不用关心查询优化等问题。

SparQL（Simple Protocol and RDF Query Language）是为 RDF 开发的一种查询语言和数据获取协议，它是为 W3C 所开发的 RDF 数据模型所定义，但是可以用于任何

可以用 RDF 来表示的信息资源。SparQL 协议和 RDF 查询语言（SparQL）于 2008 年 1 月 15 日正式成为一项 W3C 推荐标准。SparQL 构建在以前的 RDF 查询语言（例如 rdfDB、RDQL 和 SeRQL）之上，拥有一些有价值的新特性。而且，SparQL 将 Web2.0 和 Semantic web 两种新的 Web 技术联系起来了，很有可能成为将来的主流网络数据库的查询语言和数据获取标准。

1. 查询系统中的数据抽样

大数据是否还需要抽样？这个问题富有争论。以畅销书《大数据时代》为代表的观点是，既然已经有了总体数据、不再需要抽样。但许多人并不同意，认为大数据的存储、检索、甚至采集等各环节上的抽样，不仅依然需要、而且非常实用。以 Twitter 为例，它为非商业研究者提供免费抓取实时数据，但为了确保其网站畅通，故在其 API Stream 中只提供 1% 的抽样数据。经第三方实验研究，这个 1% 样本与其总体（俗称 firehose）之间存在很大差别。以此推测，Twitter 的抽样技术不行、或者没有认真地去做。这个例子反映了一个普遍存在的问题，即很多人轻视或甚至无视大数据抽样的难度。事实上，大数据抽样并不简单，即存在工程技术问题、更涉及科学理论问题。但至今相关研究甚少，本节仅简略讨论三种常见的大数据抽样任务。

第一是用于数据的原始拥有者（包括受其限权者）提取已存的大数据。这时，数据已经存储并毫无权限问题，所以是最容易的任务。小数据抽样的经典随机方法，如均匀（uniform）、分层（stratified）等，在这里一一适用。主要的挑战在于如何做到最快速度和最小误差之间的平衡，所以基本上是工程问题。以 UCBerkeley 的 BlinkDB 项目为例（blinkdb.org），他们在普通 SQL 查询系统中加入多维分层抽样（multidimensional stratified sampling），并让用户按各自需要而在反应时间与抽样误差之间选择最优方案。其结果比无抽样（即总体数据）快 10 至 200 倍，而其精确度当然比无抽样低、但比只用简单随机（即均匀）抽样的结果却要高很多。

第二是用于第三方抓取公开但访问次数受限的大数据。这时，抽样往往是最好的选择。该任务也不算太难，主要的挑战在于获得随机抽样所必需的抽样框（sampling frame）。抽样框可以是总体成员的所有 ID，也可以是这些 ID 的理论分布区间。例如，社会科学家常做电话调查来收集数据，他们如果幸运的话，能够从电信公司得到调查地区所有居民的电话号码，否则根据该地区电话号码的分布规律（如从 8000000 到 8999999），然后从中抽样。当然，用后者抽样，会抽中空号、非住户号等，效率较低。社会科学家为此做了大量实验，从中发展出一系列优化方法（统称 Random Digit

Dialing，RDD）。很多社会化媒体分配给用户的 UID，既有相似的分布规律、而且也被合法探测。参照 RDD，Random Digit Search（RDs）方法先后用于新浪博客、新浪微博、Twitter、YouTube 等网站的抽样抓取，即使只抽几万用户，便能对总体作出既准（unbiased）又精（with small random errors）的推断。

第三与第二的条件相同，但数据是网络结构，如社交网数据等。这是最困难的抽样任务，因为经典的概率论及随机抽样方法均以数据个体之间互相独立为前提，而网络结构数据恰恰以个体之间非独立性为特征。近年来先后有过若干实验研究，测试网络数据的各种抽样方法，其中以 Leskovec 等人的研究最为系统。他们将分别属于随机抽节点（random node selection）、随机抽连边（random edge selection）、随机游走（random walks）三大类的十种方法，用于五个真实网络的数据，发现随机游走和森林灭火（forest fire）方法表现优异。然而，这些方法尚不能在其他真实或模拟数据中得到复制。严格说来，网络数据的抽样目前还是一个 NP 问题。

2. 查询系统中的近似计算

大数据统计经常需要做如下的查询和统计：

- Cardinality Estimation，用于统计集合中不同的元素；
- Frequency Estimation，用于统计独立元素出现的次数；
- Top-k Elements，用于计算出现次数前 k 多的元素；
- Range Query，用于统计出现在某个区间中的元素；
- Affiliation Query，用于统计集合是否包含某个元素。

在小数据量的统计情况下，这些计算都可以采取合适的数据结构（如散列表）来做精确的统计。但是在大数据量，大基数的情况下，传统的精确统计已经不再合适。这里出现了很多优秀的算法，都是通过一些概率计算在保证一定的概率上的正确性的情况下降低大数据统计的空间消耗。包括 Cardinality Estimation 中的 Linear Count、Loglog Count、Hyper Loglog Count、Adaptivity Count，Frequency Estimation 中的 Count-Min Sketch、Count-Mean-Min Sketch，Top-k Elements 中的 Count-Min Sketch、Stream-Summary，Range Query 中的多 Count-Min Sketch 组合，以及 Affiliation Query 中的 Bloom Filter。这些算法通常都可以让计算的空间复杂度下降几个数量级，而精确度降低不多。

近似算法的并行计算也是近年的研究重点。Locality Sensitive Hashing（LSH）常用来在大规模数据集中查找相似的元素，而不用进行次方时间复杂度的比较。思路是用多个散列函数将数据划分到散列表中，相似的元素分到相同的桶中概率更大。VLDB 2014

发表的 Parallel LSH 可以在多机多核（Xeon core）上实现高速并行的 LSH 算法，可以对数十亿的推文进行流式的相似性查询。

除了这些单个的算法，BlinkDB 作为大规模数据上的限时 / 限误差的统计查询。BlinkDB 依据系统存储的性能差异从数据库中离线地创建均匀的、层次的采样，并通过 MILP（Mixed Integer Linear Program）优化选取 GROUP BY 与 WHERE 子句中常用的列。线上统计查询的时候通过 ELP（Error Latency Profile) 分析找到保证满足给定的时间限制或者准确度限制的最佳样本。每次查询除了返回统计结果，还会返回合适的误差以及置信区间。

3.5.2 大数据的数据模型和算法

1. 算法对长尾的支持

大数据带来很多有趣的现象，其中之一就是长尾。长尾也是让经典统计学失效的典范。传统把推荐算法当做黑箱的做法是用一些机器学习算法来优化目标函数，也就是将多种推荐算法转换成数学问题求解。只要是数学问题，就存在假设。然而有很多过去看似正确的假设不能更好地应用在大数据中。机器学习算法中最重视的一种分布叫作指数分布族，几乎所有的经典算法都是建立在指数分布族假设上的，而指数分布族中的经典包括高斯分布、二项分布等。这里牵扯了很大一部分算法包括 PCA、LDA、LSI 等。

这些算法在 benchmark 数据上都有不俗的表现，但是应用到实际系统中却又差强人意。问题就出在指数分布族的假设上。这些分布“ fat head”在计算模型的时候把大量的精力放在高频部分上，从而忽略了低频部分的长尾。传统机器学习中的“噪声”，去噪”也表明这些长尾是多么不受重视。然而长尾部分正是“人生百态”，是个性化方法的必备数据。

长尾问题难解来源于两个方面，一是没有经典的分布去拟合这些数据，二是这些数据本身稀疏的特性也阻碍精确求解的可能。这里面牵扯最深的就是个性化推荐算法。推荐算法存在两个极端，针对每个用户来说，数据过于稀少的时候存在冷启动的问题，这种情况下只能使用最热门的推荐等方法，阻碍了用户进入推荐系统的意愿。而数据过多又导致传统算法失效，每个用户的兴趣收集的过于集中，用户比较发散的兴趣难以收集学习，导致无法产生新奇的推荐效应。

那么如何改进算法，合理地利用长尾呢？可以从两个角度去考虑这个问题。一是依照模型组合的思路，把数据按照频率拆分成不同的等级，每个等级的数据单独用模型训练。或者直接使用模型组合的方法，让模型自己去主动学习那些长尾数据。二是按照谷

歌 Rephil 的思路，舍弃指数分布族等不合理的假设，采用更高级的概率图模型或者神经网络模型去拟合数据。

2. 有参、无参以及混合模型

在小数据量的时代，通常倾向于对变量之间的关系做深度的人工挖掘，研究变量之间的因果关系并用于指导实践。针对某个领域数据的深度挖掘成就了很多经典的模型，这些经典的模型大多数都是参数化的模型，即“有参”模型。有参模型是在用相对固定的复杂函数通过改变其特定的参数去拟合数据，这方面有很多经典的代表如逻辑回归、支持向量机、隐语义模型等。

然而这类“精确刻画某种领域数据”的模型很难随着数据规模的增长改变自身，其模型的表达能力是有一定限度的。相比之下，“无参”模型并不对模型本身做固定的假设，随着数据的变化，模型的参数和模型本身都会发生改变。最简单的无参模型就是统计直方图，以及最近邻算法等。通常我们不会严格限定无参模型的函数空间，因此它能更好地去拟合现有数据。

小数据的时代我们通过固定一些模式，之后用这些特定的模型来找到最佳刻画当前数据的参数。大数据的时代我们更倾向于让数据本身反映其本质属性。我们可以看到小数据和大数据的一些对比。小数据模型是一种一般性的规律总结，大数据模型则可以发现一些特殊性的规律。同时，小数据基于逻辑和推理并且更关心因果性，而大数据则更关心关联性。

大数据模型，即“无参”模型，与小数据“有参”模型各有千秋。为了兼顾表达力和大数据学习效率，出现了多种混合（hybrid）模型。其中模型组合（ensemble）就是混合模型的一种，其前段是多种不同的数据（重采样等）以及不同的模式学到不同的模型，后段是通过某种方法（通常是有参的方法，如线性加权）将前段学习的结果组合到一起。类似的还有支持向量机，其前段相当于使用某种核函数做无参学习，后段用有参学习将其融合；还有基于密度的逻辑回归，前段用无参的核密度估计处理特征，找到数据真实分布，后段用逻辑回归这种有参学习进行组合。甚至人工神经网络也是一种后段对前段的层层混合。混合的方法可以是有参模型加上无参模型，判别模型结合生成模型，非线性模型加上线性模型等。

3. 在线、流式机器学习

流式机器学习并不等同于普通的流式数据处理。或者说，流式机器学习要比真正的

流式数据处理简单一些。关键点来自于几个方面：

- 机器学习算法更多的是 batch 和 mini-batch 的处理方式；
- 机器学习算法对少量的数据丢失有较好的容忍性；
- （大多数）机器学习算法对数据到达顺序没有严格的要求。

因此，流式机器学习并不用像传统的流失数据平台那样对数据可靠性要求如此严格，并提供如 at-most-once、at-least-once、exactly-once 等语义，也不用真正做到每个数据都真正“流动”，只需要做到简单的 mini-batch 就可以了。

流式机器学习的主要需求在于生产系统。不同于实验室的机器学习处理环境，在线系统会有大量的实时数据导入到训练系统。因此，生产系统要处理好训练系统与实时数据的关系。在传统的生产系统中，我们可以采取简单的 batch 策略处理这些数据，即以天为间隔，将每天的数据存储到如 HDFS 等环境中，通过 crontab 等逻辑每隔 24 小时执行一次模型训练，并更新线上模型。在生产系统的长时间运行看来，这种以天为单位的批量更新策略也可以看做某种程度上的流式学习。

当我们把“天”这个重复训练的粒度不断缩小，并且把数据的规模、流量不断增大后，才会出现一些必须解决的问题：

- 增量训练模型的可行性；
- 快速模型部署的可行性；
- 模型异步更新的可行性；
- 快速可靠的更新逻辑。

模型必须做到增量训练，否则每次训练模型要过大量的数据，可能会导致模型训练的时间大于模型更新的设置时间，产生大量不必要的计算。快速部署模型最简单的方法就是做大在线的模型查询，这方面可以参考 Jubatus 的工作，以及参考一些参数服务器的设计。模型异步更新可以参考参数服务器的设计，当模型的参数较大时，可能每次只需要更新参数的一部分。面向程序设计者，更重要的是如何提供良好的可编程框架，让业务逻辑快速完善和部署，这部分可以参考 Lambda 架构的设计，比较好的有 Cloudera 实现的 Oryx 框架。

因此，相较之下 Spark Streaming 比起 Storm 更适合用来实现流式机器学习，其 mini-batch 处理策略和简洁的容错逻辑都更适合机器学习的场景，Strom 则更适合用在很多高可用的场景。模型的更新可以参考 PMML 的设计，但是 PMML 过于庞大，不太适合在生产环境中作为模型更新的策略，这方面更方便的是参数服务器，或者说其他的全

pipeline 数据内存共享工具如 Tachyon。另外可以用 Akka 等框架快速实现流式机器学习的框架。

3.5.3　大数据的降维压缩

1. 降维：数据的压缩与探索

降维是机器学习中无处不在的操作。试着看一下这些为人熟知的名字：NMF、LSI、LDA、PCA、SVD、auto-encoder、Isomap、LLE 等。降维几乎是高维数据探索、数据理解，乃至数据分类中最常用的步骤。几乎所有特征提取操作都是一种降维操作。

降维的基本假设也是一种“可学习”的基本假设。其要求数据本质存在于高维空间中镶嵌的低维流形中。至于这个低维流形的形状和维度都无从得知。近些年来，越来越多的线性、非线性降维方法层出不穷，但是在人工数据上表现较好的非线性降维方法鲜有能在实际应用中取得很好的效果。因此，本节从最常用也是最重要的线性降维方法 PCA 开始说起，之后再分别介绍基于全局方法计算的非线性降维方法和基于局部方法计算的非线性降维方法。

PCA 是一种线性降维方法。线性的含义是假设高维的数据本质上镶嵌在低维的平面中。PCA 的主要操作是收集数据中那些方差尽可能大的维度，而方差很小的维度可以基本视为噪声因此滤掉。这里其实和一些有监督的学习方法很像，如决策树和 Linear Discriminative Analysis，只不过后者的方差最大是用来分离不同的类别。PCA 的优化目标函数可以直接得到解析解，最终转变为零均值数据协方差矩阵的特征向量求解的问题。PCA 的缺点是计算量与数据量成正比，因此很难做到大规模的计算。不过有很多迭代的算法可以让 PCA（或者其近似值）在大规模数据的情况下可解，设置得到概率解。比较常见的方法有 Lanczos 算法、ALS 算法等。

之后是非线性的降维算法。主要有三种类型，一是保持数据全局性质的降维方法；二是保持数据局部性质的降维方法；三是使用混合的线性模型逼近数据的方法。保持全局性质的降维方法有 MDS、Isomap、MVU、Kernel PCA、diffusion map、multilayer auto-encoder 等。多维缩放（MDS）代表了一类降维操作，它们在将高维数据映射到低维的过程中尽可能保证两者之间的距离，因此，其目标函数就是最小化降维前后两两数据点之间的距离差异。为了保证目标函数可以同时对“远距离”和“近距离”的点都能保证相同的缩放，使用 Sammon 目标函数进行正则化。将其中的欧氏距离函数换成别的测

度，就可以进行不同的降维操作。MDS 导出了一系列的降维操作如 SPE、CCA、SNE 等。MDS 模型表达简单，重要之处就是找到合适的优化方法优化目标函数。

MDS 的假设是数据之间的距离使用欧式距离计算，然而，这在实际中是未必满足的。如降维中常用的 swiss roll 就是一种蜷曲的空间。Isomap 不同于 MDS，其计算测地距离而非欧氏距离。测地距离通过构造邻居网络并求两个数据点之间的最短路径来计算，而非之前的欧氏距离。最短路径可以使用 Dijkstra 或者 Floyd 算法求得。Isomap 有一些缺点，例如对于短路图谱、流形中的洞，以及非凸流形都难以处理。

最大方差展开（MVU）类似于 Isomap，不同之处在于 MVU 显示展开了 Isomap 中蜷曲的流形，将其摊开至欧式空间。其方法是最大化摊开空间之间的距离的平方和，而限定其中的数据节点的距离与之前的测地距离一致。

2. 对稀疏的更好支持和简化压缩

面向稀疏数据良好地设计数据结构和算法，可以归约数据量，降低计算复杂度。其一，通过发掘和支持自然数据中的稀疏属性，算作自然资源的挖掘；其二，通过对现有数据的聚集归约，视为人工数据聚合。这两种方法在 Spark MLlib 中均得到了很好的体现。

稀疏性是自然世界的本质属性。在大数据时代的数据几乎是稀疏数据的天下。稀疏性来源于两个方面，一是数据“基”非常大，即数据空间之大；二是数据“点”非常少，即可观测的性征少。N-Gram 是前者，数据空间大小以全体文字数目为幂次，而世界上能出现的 N 个文字的组合却远少于这个值。点评数据是后者，由于长尾因素，大多数用户能点评的数据非常有限，因此仅填充了“用户——物品”矩阵很少的一部分。

从机器学习人员的角度来看，现实世界的数据等于稀疏数据（或者低秩数据）加上噪声。而这正是机器学习算法能够成功的一个重要基础。统计学家和机器学习人员的一个重要工作就是在貌似繁杂的数据中找到那些简单的构造因素。有点类似于牛顿三定律，不论世界多么复杂，物质作用多么繁复，只要是在经典力学的范围内，就要遵循三定律。三定律就是世界的模型，由此构成稀疏的世界。一幅图像是稠密的，其内部充斥着 RGB 三原色的组合，并且大都不为零。一个神经信号是稠密的，在时间线上连续存在。对于这种数据，我们可以通过小波变换或者傅里叶变换找到它稀疏的一面。

如果把数据点除以数据空间作为数据密度，那么 Netflix 竞赛数据的密度只有 1.17%，rcv1 数据集的数据密度仅 0.15%，近些年来所用到的欺诈检测的数据平均密度仅有 10%，而且这些数据为了提取更多的特征，通常都增大了欺诈数据的条目，即是

有偏的！现在估算一下你手头数据的数据密度，是否要考虑稀疏性了呢？那么应该怎么做？核心就是要善于发现和利用这种属性。

有了稀疏向量只是一个基础，说明我们有能力发掘现实世界的稀疏性了。然而下一步更重要的是，如何在机器学习算法中更好地利用这种性质？稀疏性无处不在，但又非常脆弱，很多操作就能破坏它。例如向量加法，两个稀疏向量相加的结果会比之前的向量稠密一些，多个向量相加的结果就完全是稠密矩阵了。因此，善于利用利于保持稀疏性的线性代数运算是其中的关键。

3.5.4　大数据的分布式并行学习

基于机器学习方法的数据分析和基于深度学习的数据理解，前者计算复杂，迭代次数多，一般情况下参数较少；后者不仅计算复杂，迭代次数多，参数、模型还非常庞大。新旧计算框架在与机器学习算法的适配性上仍存在不同程度的差距。

随着大数据的不断深化，“简单模型 + 大数据”的数据关联寻找方式逐步代替了“复杂模型 + 小数据”的数据因果寻找方式。而大多数的复杂模型在真实场景中并没有得到很好的表现。Google 在《 Lessons learned developing a practical large scale machine learning system 》一文中也充分强调了这一点。不仅仅是简单模型，其对精确性的要求也放松了很多。对于一个能够上大数据的实际应用系统来说，易用性和系统稳定性是更值得考虑的目标。

为了更好地实现分布式（并行化）的机器学习程序，最重要的是打破多数机器学习程序优化算法的串行性要求，这里面有很多关键技术，从下面几个方向来考虑：一是从优化算法本身来考虑，如选择可分裂计算的优化算法如 ADMM，而非串行的 SGD。也可以考虑一些数据规约的简化方法，如 MLlib 中实现决策树算法就用到了一定的数据采样压缩的方法，而适当地放松精确性的要求，使用“ Stale Synchronous”的思想，以及使用线性代数库加速等，也是不同的方向。

1. 舍弃精确性与参数服务器

本质上绝大多数机器学习算法是缺乏“并行基因”的，迭代内和迭代间存在空间和时间的强依赖，这在与计算模型适配过程中制约了效率。所以分布式机器学习的关键是制造基因突变，削弱依赖，使其真正能水平扩展。

最常见的基因突变是放松数据同步中的一致性，这可能导致训练过程中丢失数据，从而降低精度。然而，大数据可以通过丰富的数据资源来弥补单个数据精度的丢失，同

时允许在大规模集群上水平扩展。典型的例子是流式 / 在线分布式机器学习计算框架 Jubatus。它采用了一种叫做 mix 的数据同步算法，这种算法使得集群里的服务器在并行训练的同时，可以通过较细的粒度（每过一段时间或每完成一定数据的训练）周期性地在服务器间同步训练的中间结果。服务器更新各自的模型后继续训练，同时因为已经有了不断改进的模型，可以接受推理或预测的请求。

以随机梯度下降为例。随机梯度下降算法的本质是，每当我们看到一个样本，就拿来更新一次模型参数。当然，如果觉得一个样本会引入太大的噪声，可以使用数个或数十个样本的统计均值来进行一定程度上的降噪。传统的随机梯度下降算法缺乏并行基因，模型参数每次更新都来自一个样本计算得出的梯度，而每次梯度的计算都需要当前最新的模型参数。这种计算依赖性导致并行几乎不可能。DistBelief 采用了一种异步变体。首先将数据切片分成子集，然后在不同的子集上执行一个模型副本。模型副本之间通过参数服务器这块“黑板”来交换参数。为实现分布式并行，每台机器执行自己的“同步 SGD”，迭代一定步数之后，再去更新黑板上的参数，或者从黑板上取回最新的参数。但由于每次模型副本更新不一定会获取最新值，造成准确度有所损失。初始时刻模型副本都一样，而以后参数服务器中的参数分片之间则是独立更新的。举例来说，参数服务器有 10 个分片，每个分片都掌握着整个模型 1/10 的参数。并行体现在两个层次：模型副本之间独立运行，而参数切片之间也是互不影响。比之于 MapReduce 的 SGD 实现，DistBelief 完成了从“小规模参数”的机器学习算法到“大规模参数”深度学习的递进。全新的框架设计保证了高效迭代。参数服务器以及不均等延迟同步的思想传递了全新的理念，由此引出了 SSP 模型。

分布式机器学习的研究早有意图从上到下突破大数据和数据 / 模型依赖性的重围，但不适用于大规模集群。2013 年，CMU 推广 BSP 模型为 SSP 模型，使用了类似 DistBelief 参数服务器的做法。SSP（Stale Synchronous Parallel）模型在理论和实践上都是可圈可点的，可看作是对 DistBelief 方法的推广和扩充。

2. 线性代数库的加速：大数据分析与高性能计算的交汇

在大数据机器学习里，矩阵计算的重要性不言而喻。线性代数计算最初源自 HPC 领域。对于单机多核的线性代数运算优化来说，Intel 的 MKL 是最快的计算库。然而，面向大规模的分布式矩阵计算却有所止步。这里主要分为三个角度来看线性代数库加速的问题。一是单机线性代数库的多版本对比，包括 native 语言和 JVM 上的对比；二是分布

式线性代数库的现状；三是分布式机器学习与分布式矩阵的关系以及映射。快速和易用一直都是线性代数库两难的抉择。

先来看矩阵计算的 API 设计问题。从 Kernel lib 来说，一般都设计成 BLAS 接口。BLAS 接口按照“向量 – 向量”、“向量 – 矩阵”、“矩阵 – 矩阵”三组不同的操作分成 BLAS-1、BLAS-2 和 BLAS-3 三种接口。但这些还不够，因为这并没有包含一些高级的矩阵操作，如矩阵求逆、特征向量分解以及多种矩阵分解方法。这些高级的操作由 LAPACK 等来提供。除此之外，还有数学运算包和统计运算包，如求解 sin/cos 值，或者按照分布采样等。这类的矩阵运算包有 MKL、CBlas、BLAS/LAPACK、atlas、suitesparse 等。

这些计算库在运算速度上都是当之无愧的高速，但是在 API 上却并不友好。其接口没有矩阵计算教科书上那么自然，没有经过特殊的训练并不能很好驾驭。因此出现了一系列方便易用的接口和语言，如 Matlab、Scipy 等。

随着 GPU 的快速发展及其在深度学习中的大量应用，GPU 上的线性代数运算包越来越受到重视。这一类的机器学习包比起 CPU 上的线性代数包要丰富很多，例如 cuBlas、clBlas、MAGMA、CULA 等。

仅有 native lib 是远远不够的，因为现在大量的大数据计算的平台都建立在 JVM 平台上，要想直接在 JVM 的平台上调用 native lib 同时又考虑接口的设计以及高速的数据交换不是直观的。因此，出现了一系列的 JVM 之上的矩阵计算库。这些矩阵计算库的设计理念基本分为两种，一是直接在 JVM 上从头构建一种矩阵计算程序包，二是通过 JNI 调用 native lib 并做大量的通信优化。对于前者而言，比较流行的有 JAMA、COLT、EJML、Commons Math 等，后者有 Netlib-java、JBlas、MTJ、Breeze 等。能够直接调用 GPU 的 JVM 平台上的矩阵运算库有 scalalab。总体而言，从头实现的矩阵计算库通常不能包含所有的 BLAS 操作，而且在大数据量 / 多核的计算上不能很好地利用底层架构，JIT 对其的加速也是比较有限，因此速度有限。这些不同的 JVM 上的矩阵运算库实现的接口也各不相同，如 Netlib-java 提供类似 blas/lapack 的原生接口，而 Breeze，COLT 等提供了方便用户是用的高阶接口。

然而，矩阵计算的运算量通常很大，单机计算有时会受限。因此，从分布式的角度考虑，又出现了很多分布式的矩阵计算库。Native 语言主要由 MPI 实现，如 SUMMA。这里主要考察 JVM 上的分布式矩阵计算库有 Apache Mahout 上的 Mahout-math，不过这个主要是为了 Mahout 的矩阵计算而设计的，因此包含的高级操作很少，仅有一些基本

操作。另外比较成熟、高效的有 BIDMat，其底层调用了 Intel MKL 和 cuBlas，高层使用 scala 语言针对不同的矩阵模式封装了相同的接口，很方便地做到 CPU、GPU 的切换。最近也有在 Spark 上实现的矩阵计算库 Marlin。

而设计一个矩阵计算库需要考虑很多内容，如浮点数的类型，是否支持双精度浮点型；是否同时支持稠密矩阵和稀疏矩阵计算，并能自动在两种格式之间转换，如 MLI；是否同时支持 CPU 和 GPU，并根据 cost-model 做自动切换；是否同时支持 vector 和 matrix；是否支持所有的 BLAS 接口还是实现一些高层的接口等。针对分布式的矩阵库还要考虑分布式的计算部分和单机多核计算的部分之间的协同，哪种控制方式效率更高。

回到分布式的机器学习角度来看待分布式矩阵计算的问题，实际上，很多时候机器学习的优化算法本身不需要太复杂的矩阵计算，因为很多时候机器学习的优化算法中复杂的矩阵计算如矩阵乘法、矩阵求逆、QR 分解等都是在拆解后的小矩阵上做的，这些小矩阵通常都能很好地放入单机内存。而直接在优化算法中使用大规模的复杂矩阵计算也是不明智的，通常而言有很多替换选择。

3. 机器学习系统并行的一般策略

在传统 MapReduce 批量处理的运行时模型，MLlib 上的算法表现为经典的块同步处理（BSP）模式。说白了就是“一放、一算、一收”。下面通过看似最简单的向量参数模型来说明机器学习算法与 BSP 模式简单映射。

“一放”代表着模型分配到各个计算节点上。对于广义线性模型来说，这一步尤其简单，因为广义线性模型的参数基本上都是向量，其维度等于数据特征的维度。一般而言，我们不认为在单机上存储一个向量是个难事儿。MLlib 现在的做法都是将参数数组直接序列化出去，即把模型参数放到运算的闭包（closure）里面就可以了。因为模型比较小，而且每个计算节点都会用到所有维度（坐标梯度下降类除外），也不用作模型划分。

“一算”就是每个计算节点根据本节点的数据独自更新自己的模型。这部分是分布式程序的关键，要在确保正确的基础上才能对机器学习任务进行分割，否则空有水平扩展，没有正确的结果也是没有意义的。市面上各种机器学习并行算法背后一般都有比较复杂的数学证明。一般而言，切分到每个计算节点上的部分都成为一个独立的本地机器学习算法，或者说本地机器学习算法的一部分。这部分计算的正确性由水平扩展部分保证，而性能提升靠垂直扩展保证。

“一收”就是各计算节点结果的回收整合。这里的结果可能是一个更新量，比如说梯度，也可能是模型本身，或者模型的一部分。具体拼接过程视不同的模型不同，需要具体分析。广义线性模型的通常做法就是把梯度收回来做平均，然后在 Driver 上本地更新模型。但也不尽然，有些广义线性模型可以直接把各个计算节点得到的模型拿回来做更复杂的平均，比如说做模型组合（这时模型之间的区别来自不同的训练数据），会得到更好的结果。

4. AllReduce 的参数收集

源自于 MPI（Message Passing Interface）的 AllReduce 模式在分布式机器学习的参数收集过程中独领风骚。凯文·凯利所极力推崇的去中心化思想在这里也得到充分的体现。由于机器学习算法中“参数”这一独特的性质，在分布式机器学习中看似是无可避免的无法逃离中心化的命运。试想一下，不论是批量 / 流式数据处理，还是 SQL 查询，基本上不会出现大量的集中化参数的影响，因此可以精巧的躲过中心化的厄运，得到极强的分布式逻辑。事实上，以 MapReduce 为代表的分布式市场从来都对机器学习算法产生一些“不适”。

最近火热的参数服务器更是把中心化的思想表达到极致，因此不得已会用到“ Stale Synchronous”和模型分布式的观点去减轻中心化带来的压力，但并未改变其本质。

回头来看 AllReduce，却是对去中心化的分布式机器学习参数收集一次较好的尝试。如果说在现实世界中找个类比，可以参考比特币与 ATM 等刷卡设备了。比特币也是去中心化的思路，比起各种信用卡消费，前者更像是“电子现金”。本质上就是“电子现金”的交换不用再经过银行这一中心节点。而不论是刷卡消费还是 ATM 取现，都必须要从银行的数据库走一遭，做出“付款 – 扣钱 – 收款 – 加钱”的事物操作。然而普通的纸币交易是无中心化，人们共同信任纸币体系。比特币就是将“电子现金”真正做到了去中心化。

AllReduce 就是这种情况，但又稍有不同。之前是每个 Executor 产生的结果汇报 Driver 做累计，更新参数后再由 Driver 向各个 Executor 序列化地回送新的参数。AllReduce 抹去了 Driver 的参与，参数收集更新由 Executor 们共同完成。这样，打破了“一放、一算、一收”这个流程。而 AllReduce 和真正去中心化的不同之处在于，最后会有一次全局的广播让各个节点得到最新的参数。

5. 向量模型的思考

Pedro Domingos 教授的“ A Few Useful Things to Know about Machine Learning”一

文把机器学习总结为模型表示、评测标准、优化方法三要素。前者决定模型复杂度（如粗细粒度和相互依赖），而后两者又决定了计算复杂度（如迭代次数和串行相关性）。当把它们映射到分布式系统上时，计算模型与上述两种复杂度的适配决定了最终的效率。粗粒度数据表示的计算模型，有适于少迭代的两阶段或三阶段数据流，如 Hadoop；也有对多迭代更优的 DAG 数据流（如 Spark 和 Tez）。很多机器学习算法基于细粒度且存在相互依赖的数据表示，又有多迭代，GraphLab 的 GAS（Gather/Apply/Scatter）就比较适合。不同计算模型的选择决定了时间维度上迭代间的串行 / 并行性，也决定了空间维度上节点间的并行性以及通信 / 同步的效率。从评测标准和优化方法出发，使模型表示适配到计算模型，是实现分布式机器学习的难点和重点。

对于并行机器学习算法而言，不像那些本地程序，必须有所规则决定哪些 Worker 使用哪些参数，更新后的参数交给谁。由于参数一般是全局共享的，所以 Driver 可以整体接管所有参数分发、更新、收集动作。（在这一点上，Driver 所起的作用非常像参数服务器。）广义线性模型的运算方法符合 Andrew Ng 等人于 2006 年发表在 NIPS 上的“ Map-Reduce for Machine Learning on Multicore”一文中描述的‘统计查询与加和范式’的情形，非常适合使用 MapReduce 实现，因此广义线性模型也是最常见于 Mahout、MLlib 等分布式机器学习平台中的算法。然而，广义线性模型这种向量参数化的模型并不是全部，像 ALS、LDA 等算法有着更复杂的模型表达（矩阵）以及更细粒度的数据访问。

6. 矩阵模型的思考

最简单直接的方法就是跟向量模型等同处理，完全忽视它可能超过单机内存引起存储不足的情况。其实说的可怕，现实中可能没有那么大的问题。以现在单机内存的能力，存储一个超大矩阵也不是难事儿。何况，我们还可以做很多预处理工作来压缩矩阵维度呢。举个例子，使用 ALS 做矩阵分解，从而进行个性化推荐。假设有 m 个用户（user），n 件物品（item），整体来看有 m × n 种映射关系，所以这个 m × n 的矩阵作为输入。当然为了利用输入数据的稀疏性，通常选择稀疏行压缩（RSC）/ 稀疏列压缩（CSC）等经典方式存储矩阵。这个量看起来是很可观的，但对于我们要训练的模型来说就没那么大了。通常情况下，会给定一个秩的值，通常称作 k。ALS 结果得到的模型是 m × k 和 n × k 这两个“瘦长”矩阵，即行数远大于列数，两者不在一个数量级上。因此矩阵的存储量可视作 O（m）或者 O（n）。更何况，对于超大矩阵直接做 ALS 就开始推荐通常意义不大，先做个商品聚类、用户聚类，再在聚类结果之间做 ALS 是个更常见的选择。聚类之间再进行 ALS 操作可以获得更多的有用信息，还能极大的规避一些不重要的长

尾。至于很多情况下，大数据要发挥长尾优势（如一些冷门电影的推荐），可以对算法重新设计调整，如图书推荐系统中所谓的“专家推荐”。

大数据导向的全体应用集体向分布式系统的迁移带来更多挑战，使得原来单机多核的并行计算成为多机多核的分布式计算，访存模式更加不确定。非一致性内存访问（NUMA）的系统结构加上机器学习算法的多变特性，使问题越来越繁杂而有趣。抛开 Spark、Hadoop、GraphLab 等特定的架构，未来的分布式机器学习中要考虑多个方面的平衡：首先是模型的分配与回收，到底是复制还是消息传递；其次是分布式数据分配，到底是复制还是切分；然后是数值优化，硬件效率与统计效率之间如何权衡；最后是访存模式，细粒度数据访问与粗粒度数据访问交相呼应。多方面综合考虑，再加上系统分层带来的设计简化以及数值优化、统计学等不断演进，相信在可见的未来该领域会有长足的发展。

3.5.5　大数据机器学习系统

机器学习在大数据时代日益重要。现有的分布式数据并行计算系统比如 MapReduce，Dryad 和 Apache Spark 等已经可以胜任对超大规模的数据进行灵活的计算、处理以及分析。但是，简单的分析还远没有充分利用大数据的潜能。

Frank Zappa 曾提出：“Information is not knowledge.Knowledge is not wisdom.Wisdom is not truth.Truth is not beauty.Beauty is not love.Love is not music.Music is the best…”。简单的信息提取和分析还谈不上洞察和知识，更远谈不上智识和真相。而机器学习被普遍认为是从大数据通向人工智能（从历史数据中学习对未来的预测模型）的主要技术手段。并且，大数据和机器学习互相依赖：没有大数据，机器学习也学不到有普遍意义的智能，而没有机器学习，大数据也失去了重要的意义。

不同于基于单机和偏小规模数据的机器学习系统，设计一个高效的、可扩展的、易用的大规模分布式机器学习系统面临诸多的挑战，目前仍处于探索和研究阶段。这将是过去几年和未来几年的热点研究领域，工业界和学术界都持续的投入相当多的资源，各种不同的机器学习系统也不断涌现，比如 VW、GraphLab、DistBlief、Project Adam、Petuum 等。

任何的技术趋势大体上都会如图 3-4 所示，在人们意识到该技术的必要性和重要性的早期，会涌现出大量五花八门的系统（“百花齐放”）。但是，随着时间的推移，最终只会有有限的几种系统胜出（“适者生存”），这样的系统会在性能和易用性上达到最佳的平衡，拥有丰富的生态，不断持续地进化。机器学习系统的趋势也不会例外。

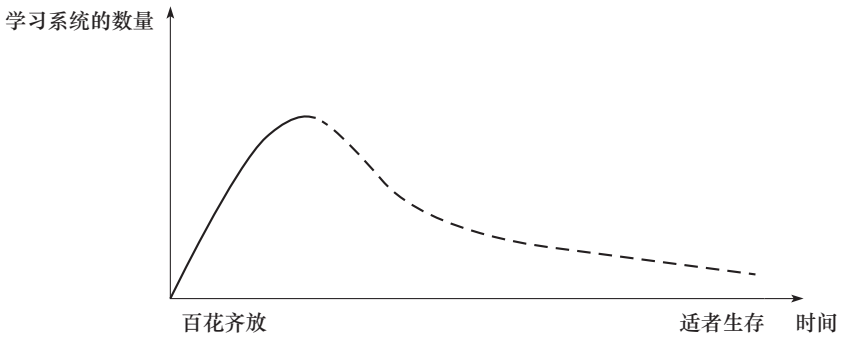


图 3-4 技术趋势的发展、进化、成熟和统一

目前看来，机器学习系统的设计仍然处于“百花齐放”阶段。当前的系统设计大多是针对特定的机器学习模型和特定的计算模式，其设计方法基本上按照针对特定的应用，选择特定的模型，从串行原型、并行原型到分布式原型进行持续的优化，包括使用细粒度的锁、特定的数据结构和特定的分布式协议等等。尽管这可以最大化利用系统资源达到最佳的性能，但是这种比拼底层实现，将学习和系统混杂在一起的设计，难于调试和维护。这是由大规模机器学习系统本身的复杂性决定的，机器学习系统既包括机器学习，又包括大规模系统。而机器学习又包括模型、训练和精度要求等；大规模系统也需要考虑并行度、局部性、调度、容错、掉队者（straggler）和网络等因素。如图 3-5 所示，这些因素之间互相影响，交织在一起，极大地增加了系统设计的复杂性。

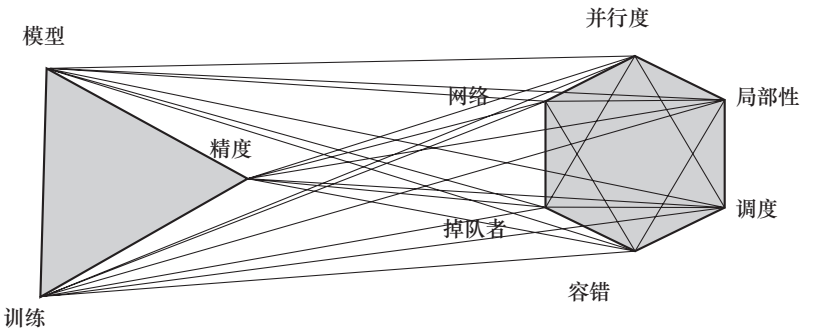


图 3-5 机器学习系统设计的复杂性

秉承这样的方法论，将涌现各式各样的复杂学习系统（“百花齐放”）。假设有 n 个研究机构，和 m 个模型，可能会有 n × m 种不同的系统。这些系统往往都可能在重复同样的设计和实现，解决同样的系统挑战等等，使得抽象这些重复的系统设计变为可能。这还将大大增加开发者和用户的学习成本。更严重的是，尽管这些特有的学习系统可能

在模型训练阶段会有更好的性能，但是如果从整个机器学习生命周期的角度看来，如果训练数据的提取和生成、模型的训练、模型的查询和服务等由不同的系统实现，除了大大降低用户的生产效率之外，由于数据将在不同的系统之间传输和保存，机器学习端到端的整体性能也未见得是最佳的。

那么什么样的机器学习系统会最终胜出呢？胜出的系统需要具备怎样的条件呢？第一，机器学习系统应该从整个学习的生命周期来考虑，包括训练数据的提取、分布式训练、模型的查询和服务都应在同一个计算平台。第二，该系统提供多种的并行训练模式，支持不同的机器学习模型和算法。第三，该系统提供对底层系统的抽象，可以支持通用的底层计算引擎，并且提供数据科学中常用编程语言的 API。第四，该系统应该拥有开放和丰富的生态，能够广泛应用和快速进化。

另外一种思路是通过系统抽象来降低系统设计的复杂性。如图 3-6 所示，通过定义包含特定接口的系统抽象来将上层机器学习和底层分布式系统解耦开来，将机器学习实现在现有的大数据计算平台之上，比如分布式数据并行计算，而不需要去考虑底层系统层面的因素。而机器学习库只需要考虑常见的机器学习模型和算法，以及各算法和模型所共有的并行计算模式，包括数据并行、图并行以及模型并行等。而这些系统抽象 API 提供这些并行计算模式接口，以及模型和训练数据的表示，底层分布式系统负责提供高效的训练实现。

图 3-6　从机器学习系统到机器学习库

目前看来，Apache Mahout 和 Spark MLlib 基本符合上述的设计思路，即从设计特有机器学习系统到通过解耦机器学习和底层系统，并提供通用机器学习库的转变。以 MLlib 为例，其在通用的 Apache Spark 执行引擎上提供和实现机器学习的库，提供多种模型、算法以及基础数据结构和评测实现。整个 Spark 生态可以提供数据并行、图并行（GraphX）。当然，Spark 生态需要进一步进化，提供包括模型并行、参数服务器、在线学习和多模型学习等至关重要的功能，这其中蕴含着潜在的研究机会和工程挑战。

3.5.6　其他实用性问题

1. 特征工程、超参选择与多模型训练

特征工程与超参选择是机器学习书本中经常忽略但是实践中却又很重要的话题。实现一个有效的机器学习训练系统并能支持线上应用，远比书本中一个模型的原理和实现

复杂得多。特征工程牵扯到数据的基本 ETL、归一化、零值处理、空值处理、异常点处理、特征选择、特征融合等多项要素，而且现阶段除了通过 validate set 来进行随机选取外鲜见有较智能的做法。有些工作通过启发式的特征搜索降低特征组合爆炸引起的过度计算，还有些工作将特征工程的工作融入到训练过程中，因此不至于做过多的试探计算。超参选择基本上也是通过 grid search 等方法进行随机尝试。除了考虑计算量之外，设计良好的分布式计算系统还要考虑对于用户而言的易用性，系统是否可以提供简洁的特征工程和超参选择方法。

针对合适的场景，多模型训练比单模型训练能够更好地利用一些并行的特性，如数据级别并行、指令级别并行等。数据并行如 MapReduce，每个 Map 获取部分数据进行训练，之后在 Reduce 中进行模型组合。指令级别并行有矩阵计算上的加速以及 SIMD 加速的可能。在矩阵计算加速的场景中，可以将单模型训练的向量乘法变成多模型训练中的矩阵乘法，如多模型的逻辑回归，原始情况是每个数据与单模型向量进行点积，现在是多模型的参数矩阵与 mini-batch 的数据矩阵之间的矩阵乘积，因此可以用到 MKL 等矩阵加速计算。而 SIMD 加速层面也很可观，如随机森林算法中多棵决策树训练可以通过 SIMD 并行来做，其改变的仅仅是参数。由此可见，多模型的组合解决的不仅仅是比之于单模型的精确度提升，还能做到尽量让多模型的训练时间不要超过单模型训练过多，避免性能弥补精确度而引起“得不偿失”。

2. 无监督特征学习

随着 deep learning 越来越火，无监督特征学习的方法也是不绝于耳。不同于传统的暴力的特征融合、组合方法，无监督特征学习倾向于先学习到一些正交的、互补的、本质性的特征。现在常说的无监督特征学习经常用来特指 auto-encoder、RBM 等非线性特征学习方法，实际上，几十年前 PCA、NMF、LSA 等就是无监督特征学习的起点了。不同之处在于之前尝试的多种方法都是线性的学习方法，近年来多用一些可学习的非线性方法来代替，因为非线性方法通常可以学到更丰富的特征。如果说线性的无监督特征学习是一些线性因式分解方法，那么非线性特征学习就是非线性因式分解的组合。

然而，在具体的使用中需要依数据的本质而定。像图形、声音、文本等复杂数据，可以用非线性的特征学习来覆盖其可能的特征空间。但是对于一些交易数据、事务数据等不存在很复杂的特征空间就不适合用复杂的非线性模型，因为线性的特征抽取足以覆

盖其特征空间，过于复杂的非线性空间反而会引起过拟合。

3. 全 pipeline 框架设计

平台做出来是要给人用的，因此，无论实现了多么复杂尖端的算法，如果不能留下良好的接口就会损失很多用户。而一个缺少用户的平台就缺少实地测试以及错误反馈，是没有生命力的。然而这方面却很少提及。不过不论是在学术界还是工业界，都出现了一些设计良好、思路清晰的典范。

DataFrame 是目前业界常用的机器学习算法表示数据集的方法。最早使用 DataFrame 概念的是 Python 的 scikit-learn 系列产品，它们在设计机器学习 API 的数据结构和接口方面独树一帜。其 API 设计理念包括：

- 一致性：所有模型对象共享相同的接口设计（针对多模型工作模式）
- 可检查：参数与超参作为 public 元素可以直接访问
- 可组合：支持模型 ensemble 以及 pipeline
- 初始值：模型应该有有意义的初始值（参数与超参）

该概念随后被 GraphLab 借鉴，用来设计 GraphLab 3.0 系统，并成为 SFrame。R 语言中也使用了 DataFrame 的概念。现在的追随者是 Spark 上的 ML 系统。ML 系统作为独立于 MLlib 外的全新接口，主要面向全 pipeline 的数据分析和机器学习算法设计。通过借用 Spark SQL 中的 SchemaRDD，采用将数据的 schema 与数据本身分离的思想，从比 job 更高的层次来控制数据分析的全过程。之所以使用 SchemaRDD 还有一个原因，就是其列式计算。这样就保证了 DataFrame 的每个列在计算的时候，只需要必要的列就可以了，不用调用整个 DataFrame 的各个列。

在 Pipeline 设计方面，scikit-learn 抽象出几个重要接口，包括：

- Estimator：所有有监督 / 无监督的模型训练算法都要提供该接口，特征提取、特征选择以及降维均要实现该接口。
- Predictor：基于 Estimator 得到的对象，调用相关接口进行预测。
- Score：通过某种评价规则得到测评指标。
- Transformer：用于数据训练前的预处理、特征选择、特征提起、降维等操作。
- Meta-estimator ：很多模型都是许多简单模型组合起来的“元模型”。如 Ensemble 模型、多类别分类模型、多模型推荐等。最终模型的结果是许多简单模型的一种组合。
- Pipeline ：数据在模型中训练通常是一个流水的过程，分别经过不同的组件如预处

理、训练、测试，而训练的时候流水可能会过多次循环。模型选择可能是针对整个流水，并非是针对单个模块。

- Model selection：针对每一种模型，都存在一些超参（hyper-parameter）需要通过搜索确定，如 Latent feature 抽取时的隐变量维度，或者聚类时社团的个数等。模型选择有两种常用的方法，即暴力搜索和启发式搜索。

除了 ML 这种数据分析的 pipeline 设计，还有一些针对应用设计的应用全 pipeline 设计。这些 API 不再面向于单纯的数据分析，因此不提供 DataFrame 这种级别的抽象，反而提供一些更加简单的应用级别抽象，如 Persistentable 的模型抽象，简单的数据抽象，json 格式的文件抽象等。这些抽象不再关注如何高效地进行数据处理和模型训练，而是更加关注具体应用的数据流，对数据本身、模型本身等提供简单的封装即可。

4. 缺乏标注数据的有监督学习

大数据时代最不缺的就是数据。不过光有数据还是不够的，分类工作基本上都需要数据的标注。那么，缺乏标注的时候分类工作怎么开展呢？这里主要讨论三种方法，即 semi-supervised learning、transfer learning、self-taught learning。

semi-supervised learning 存在的时间比较久远了，它的基本假设是少量的标注数据和大量的未标注数据是独立同分布的。这是一种较为严格的假设，可以把它看做用聚类去辅助分类的做法。transfer learning 时间要短一些，它不要求两类数据来自于同分布，但是要求辅助数据集也是有标注的，这是一种通过其他分布的标注数据来辅助目标分布的少量标注数据决策的方法。最后一种 self-taught Learning 出现得最晚，其对辅助数据集的要求也最低。它只需要来自于其他分布的未标注数据集即可。self-taught learning 本质上是一种特征学习。

不论是三种方法中的哪一种，本质上都是去寻找数据底层共享的基底（basis），或者说一种“知识”。而不同之处在于共享基底所处的层次。就 transfer learning 学习形式而言，有如下三种：

- 归纳法（Inductive）的迁移学习；
- 直推式（Transductive）的迁移学习；
- 无监督（Unsupervised）的迁移学习。

其中，归纳法的迁移学习需要目标领域的标注数据，但不对辅助领域的数据标注与否进行限定；直推式的迁移学习需要辅助领域的标注数据，但是不需要目标领域的数据

进行标注；无监督的迁移学习两个领域的数据都不用标注。

就 transfer learning 学习方法而言，分为如下几种：

- 单例（Instance）数据迁移；
- 特征维度（Feature representation）迁移；
- 模型（Model）迁移；
- 相关知识（Relational knowledge）迁移。

单例数据迁移主要是将辅助领域的数据进行加权（通常是降低辅助领域数据的权重），利用到目标领域的学习中。这种情况下共享的基底是数据本身。特征维度迁移通常用来同时训练两个相关领域的模型，并假设两个相关领域共享部分隐式变量（特征）。这种情况下共享的基底是特征。模型迁移指的是相关的学习任务会共享一些模型参数，共享基底是模型本身。

5. 基于隐私保护的机器学习

隐私是大数据带来的另一大问题。大数据的最终源头本质上来源于人的活动，只要是人的活动就存在隐私的问题。在隐私保护备受关注的今天，出现了越来越多的系统架构以及算法设计在保证用户隐私的情况下进行知识提取。本节不考虑隐私保护的系统如 CryptoDB 以及隐私保护的常见手法如加噪、去识别信息、微分隐私等，仅从机器学习算法的设计上来考虑隐私保护的问题。

机器学习算法本质上是一种知识提炼，因此本身就有隐私保护的属性。有的隐私保护的机器学习算法都基于特征变换，尤其是在任务指导的情况下的特征变化。为什么着重强调“在任务的指导下”呢？因为不同的任务 / 算法需要从数据中抽取不同的特征，现有的手段很难保证随机抽取大量的特征可以用来给每种算法进行训练，因此特征抽取、隐私变换、任务学习通常都是耦合在一起的。

但是现有的方法都不存在可扩展的能力，针对不同的算法基本上要重新开发新的算法来解决隐私特征变换的问题。例如 IBM Watson 实验室开发的 SVM 下的隐私特征变化为 low-rank feature preserved learning，而 clustering 方面的是 distance-preserved feature transformation，决策树算法的是 distribution-preserved feature transformation。

3.6　大数据可视化

可视化通过交互式视觉表现的方式来帮助人们探索和理解复杂的数据。面对大数

据的挑战，可视化与可视分析能够迅速和有效地简化与提炼数据流，帮助用户交互筛选大量的数据，有助于使用者更快更好地从复杂数据中得到新的发现，成为人了解复杂数据、开展深入分析不可或缺的手段。

总体来说，大数据的可视化主要面临两个挑战，感知（percetual）和交互（interactive）。体量庞大的数据超越了人感知和认知的能力，而相对有限的显示器分辨率限制了人们对大数据的进一步认知和发现；此外，对大数据的查询往往导致较长的延迟，无法满足交互实时性的要求。这些问题日益积累，具体化为十大挑战，其中技术层面的挑战涉及：

- 原位交互分析（In Situ Interactive Analysis）：数据仍在内存中时就会做尽可能多的分析。
- 交互与用户界面：突破数据规模增长对人的认知能力的挑战。
- 数据的表示：能够对大规模数据做抽象可视化（如投影和降维）。涉及多级层次架构（Multilevel Hierarchy）时，随着数据规模变大，架构的深度和复杂性也变大，在遍历和选择分辨率时带来了挑战。
- 不确定性的量化：大规模数据分析常常带来不确定性（如为获得实时性进行了采样，因而丢失了异常点），这需要一种直观的视图让用户理解和控制不确定性。
- 面向领域和大众的可视化工具库。

3.6.1 实时可视化

随着超级计算机计算能力的迅速发展，科学家可以采用更大规模、更高精度、更长时间的计算来模拟更复杂的物理化学现象和工程设计问题。由此产生的数据也呈爆炸式增长，单次模拟的计算量已到达 TB/PB 量级数据。而超级计算机的 IO 速度却远远落后于计算速度的增长，超大规模数据量也超出了超级计算机的存储能力。传统的数据预处理和可视化基本上是后处理模式，超级计算机进行数值模拟后输出的海量数据结果保存在磁盘中，当进行可视化处理时从磁盘读取数据。数据传输和 IO 瓶颈的阻塞问题就增加了数据处理和可视化的难度，降低了整个数据模拟研究的效率。同时，为了降低数据处理通量，常采用时间上的采样方法，每几百个时间步输出一个时刻，只保存部分数据。但是，这个方法会造成 90％以上结果数据的丢失，很可能丢失包含特征现象的重要时刻的数据，而时序上的特征有时比空间上的特征更为重要，这就违背了需要大规模高精度数值模拟的初衷，无法充分利用高效能计算机的优势，造成资源浪费。

原位可视化是指对计算过程中产生的数据不经过存储而直接在计算模拟的同时进行实时可视化分析的过程，它将模拟计算和可视化处理紧密结合，计算出来的数据在原位被缩减和处理，结果数据量将大幅减少，需要保存和传输的数据也将大幅减少，从而提高了可视化与分析效率。原位可视化与分析被认为将是解决千万亿次规模计算数据分析的最有效途径。

人们在原位的数据压缩与组织、原位的分析算法及并行绘制、原位分析的系统结构等方面进行深入的研究，取得了一系列引人瞩目地进展。Abbasi 等人更进一步扩展了原位可视化，提出移位可视化分析，把可视化与分析算法的分类，把不适合在计算模拟同一节点的分析算法移位到不同的节点，而与计算共享节点的分析算法，并充分利用节点硬件的特点优化算法，减少对计算节点的影响；最近 Landge 等人提出来一种基于合并树分割的原位特征提取方法对大规模燃烧模拟取得了较好的效果；ParaView、Visit、Danaris、Datespace 等可视化软件也都开发了相应的原位可视化模块。但原位可视化真正全面走向应用还有很多问题需要进一步深入研究，如适合原位处理的数据特征提取与跟踪算法、原位数据压缩与组织算法、高可扩的并行可视化分析算法、高效的数据传输方法等，而易用的原位可视化中间件，是原位可视化走向应用的关键。

内存数据立方体（In-memory Data Cubes）是实时、交互可视化常用的一种数据表示。数据立方体是数据库 / 仓库中加速查询的一种数据结构，但并不适合主内存处理。Dwarf cube 对其进行了改进，采用多层次树（hierarchical tree）结构，去除了空单元，极大地减少了内存需求，更适合于可视化，但不擅长时空维度多分辨率下的平滑缩放和平移。Nanocube 在 dwarf cube 的基础上对时空数据的查询进行了优化：它采用了分层（layering）策略，所有层次分为空间、类别和时间三类，并且按照这样的顺序遍历——先是最多 25 层不同分辨率的空间层，然后是类别层（如 Twitter 数据包括设备、应用和语言三种类别），最后是时间层（同样具有不同的分辨率）。查询可以优化为按照遍历顺序在不同层之间跳跃串成的路径。加上其他的优化，Nanocube 可在一台 16GB 内存的机器上对 32TB Twitter 数据进行流畅、交互式的可视化。MIT 的 ImMens 结合了内存数据立方体、客户端缓存、Brushing 和 Linking 技术，尤其是利用 WebGL 来获取 GPU 的能力。

MIT 的另一项工作 Massively Parallel Database（MapD）同样采用了 GPU 加速，并且在 GPU 前端实现了一个 SQL 列存储引擎。基于这个思路，MIT 进一步构建了 ScalaR。ScalaR 是基于 DBMS、用于支持可视化的数据存储，该存储利用分块技术实现“按需分辨率”，同时基于用户的交互日志对用户查询可能使用的数据进行预取和缓存。

3.6.2 不同数据类型的可视化

1. 文本可视化技术

在纷繁复杂的信息世界里，文本是最为重要的信息传递及交流沟通手段。随着计算机网络和信息技术的广泛应用，文本信息越来越海量化、多样化和即时化，帮助人们迅速分析海量文本信息，获取洞察，已成为当前十分关键也颇具挑战性的问题。在过去的几十年中，研究人员在文本挖掘领域提出了一系列基于文本主题的分析技术，可以在一定程度上帮助用户理解这些复杂的文本信息。然而主题分析技术的结果通常非常复杂，一般的用户难以理解。另一方面，不同的用户有不同的需求，一个主题模型很难满足不同用户的信息需求。为了解决这一问题，研究人员将文本分析技术和交互式可视化技术结合在一起，设计了多种文本可视分析技术。这些技术充分利用人类与生俱来的对图形的迅速辨识及分析能力，将文本挖掘结果及相应的文本数据转换成直观的、可交互的展现形式，使人们可以通过视觉迅速获得有用信息，从而达到对大文本数据集进一步分析、推理以及理解的目的。已有的可视分析技术主要包括静态和动态两大类方法：静态可视方法不关心文档的时间属性，着重研究文档以及内容直接的静态关系；而动态方法则研究文档集合中随着时间变化的内容以及相应关系，用于找出一些关键的时刻和事件，并进一步推导相应事件产生的原因。除了对于主题分布等方面的分析，近来的工作逐渐开始关注更高层的语意信息，例如主题之间的对立、竞争等关系。对文本大数据的可视化，出现了对于多层次主题演化的工作。

2. 网络（图）可视化技术

图（网络）数据指包含实体、实体间关联及实体与关联上附加属性 / 类别的一类数据集。与传统的网络研究（Networking）相比，这里定义的网络更强调其作为数据的特征而非作为因特网核心体系架构的角色。在数据分析相关领域，“图”（Graph）与“网络”（Network）几乎可等同视之，因此我们统一使用图数据代表。在席卷全球的大数据浪潮中，研究人员普遍认为图数据难以采用传统的存储、查询及分析方法（如关系型数据库）处理。同样地，图数据可视化也面临大数据背景下的新挑战，以下我们按照大数据的核心特征分类梳理：1）图规模（Volume）：如何解开数据线团？伴随着信息与互联网革命，我们生成并得以获取的图数据体量已非经典的 Zachary’s Karate Club 数据集（34 个节点）可比拟。例如，国内四大微博注册用户超过 8 亿，每日新增微博约 2 亿条，用户与微博之间产生了巨大的关注、评论及转发图。这些大容量图数据的直接可视化在互联

网上极少出现，采用常见算法布局上百万节点需要数小时或更长时间。即使对于仅有数千或几万节点，大部分的可视化结果非常杂乱，常被称为“线团”（Hairball）。如何结合可视化与分析手段，快速高效解开数据线团，是本研究领域面临的首要挑战。2）高维属性（Variety）：如何展示附加数据？大数据的多样性与近期图可视化的研究热点——多维网络可视化（Multivariate Network Visualization）不谋而合。在传统的拓扑数据之外，重点关注节点及连接关系之上的高维属性。著名在线社会网络 Facebook 中，每个用户的档案包含个人信息、活动记录等近百个属性；从大量生物文献中利用文本挖掘技术抽取出的基因与蛋白质间的交互关系分为刺激、抑制、中性等多种类型。传统可视化方法通过节点（边）的大小、形状、颜色等视觉编码可呈现 5 ~ 10 维附加属性，但难以承载维数更高的新型图数据。3）动态特征（Velocity）：如何刻画图变化？不难理解，我们所探讨的图数据大部分处于在快速变化中。人口流动网络依日夜交替、春秋变换、时代演进不断变迁；软件函数间调用关系随程序的执行进程而动态变化。如何利用可视化手段刻画图变化并分析总结其规律，是本领域研究者钻研多年的难题。简单直观的“动画”式展示方法给用户以愉悦的使用体验，但多个重要研究结果表明，该方法难以适用于数据分析任务。用户往往过于关注可视化的动画效果，而忽略图数据的时序变化情况。

当前主流的三类研究方向：1）面向可视化的图数据挖掘方法。传统的图挖掘方法从量大、多变、多样的数据中发掘出深层的知识或者模式。然而，类似方法原始结果的可视化通常难以为用户理解。另一方面，图可视化方法已在众多应用领域取得显著进展，如微博信息传播可视化系统。如能根据用户友好的可视化形式改进数据挖掘方法，即面向可视化的图数据挖掘方法，将极大地提高图数据可视分析方法的有效性；2）自然和谐用户交互。人机交互方法是可视化界面不可或缺的组成部分。典型的图数据可视化交互，如视图放缩、平移、节点拖拽，已得到大量应用并随众多开源图可视化工具包广泛传播。在大数据时代，图数据的可视分析在多方面亟需创新的交互方法设计，如图浏览、图比较、图挖掘算法（参数）调优等。3）多态图可视隐喻。当前学术与工业界绝大多数图数据采用节点 – 边可视隐喻展示。这与节点 – 边形式符合人类对于关联数据的感知规律有关，详细讨论可参阅格式塔理论（Gestalt）。在大数据时代，图数据的体量、动态性及丰富的节点 / 边属性正给这一传统带来挑战。针对动态图数据，基于平行坐标隐喻的平行边分割方法可提供更佳的可扩展性。

图数据蕴含世间万物相联、相生、相克的复杂关系，与我们的生活息息相关。在大数据时代，图数据正走向大型化、动态化、多样化。如何通过可视化与可视分析方

法诠释数据特征，真实展示图结构原貌，最终深度服务用户，是信息可视化、数据挖掘及人机交互领域面对的共同挑战。值得指出的是，大数据时代的图数据可视化问题，学术界及工业界尚无系统、完整的方案与方法论，这需要整个研究领域的持续投入与共同协作。

3. 时空数据可视化技术

在过去的一年中，时空可视化方向得到了稳定的发展。研究者们发表了大量论文工作，涉及交通、电力、市政、情报、博物馆管理和生态环境等众多领域。其中，一部分工作的研究重点在于可视化方法。它们提出了一些新颖的视觉表现形式，以更好的展示数据的特性。另一部分工作的重点在于交互式分析方法。它们通过探索式的可视分析界面，支持用户发现数据的总体规律和异常行为。

可视化设计方面的工作主要采用符号（Glyph）设计来表现数据中的时空模式。交互式分析方法方面的工作主要是基于成熟的多视窗联接视图技术（Multiple linked view）。该技术将数据的不同方面用各自最适合的可视化方式展示在不同的分析窗口中。每个窗口都支持筛选和联动。但是为了解决不同的问题，每个工作都有不同的设计。按照具体的分析内容，这些工作可以分为高维特征分析、事件分析、语义分析、交通状态分析、旅行时间分析、空间边界分析和模拟分析。在高维特征分析方面，研究者们主要利用投影的方法探索异常的时空信号。

4. 多维数据可视化

多维数据可视化领域的研究进展主要集中在可视分析中的应用、方法的评估和局部改进上，新提出的可视化形式和方法并不多见。利用 small multiple 的策略可以并列展现数据属性随地理空间的变化折线图，使用户能够比较和分析各个属性的关联与演变规律。类似的自底向上的多维网络构建方法，并使用并列的数据直方图来展现和比较网络节点 / 边的属性分布。并列视图的方式有利于单个属性层面的比较分析，却并不适合表现规模较大的数据，而并列图符的形式正好弥补了这一缺点。应用上，可以特定的图符来表达预测模型中不同参数的影响，以帮助用户选择重要的模型参数。技术上，新的工作包括利用图符来表征散点图的特征，提出了图符散点图矩阵的新方法，以帮助用户进行简单的多变量关系分析。研究者们全面地考察变换空间，以使得特征的提取与分析更为深入、健壮和多样化。针对更多的数据，密集散点图的概括表示问题也被研究，可以利用多类蓝噪声采样的方式来降低散点图密度，在防止视图混杂的同时保持其数据特征

及表达效力。

3.6.3　交互可视化

交互作为人与计算机之间的媒介，连接着现实与虚拟，具象和抽象的两个世界。在过去的近半个世纪里，从命令行输入，到图形化界面，再到现在的触摸式交互，人机交互技术在快速地发展。作为可视分析中最重要的部分之一，人机交互的设计具有重要意义。

1. 支持可视分析过程的界面隐喻与交互组件

早在发明出计算机之前，人们已经有分析数据的习惯。在此过程中发展出很多的工具或方法，被广泛地应用到数据分析的过程中。在设计可视分析的交互时，常常借用这些人们惯用的分析工具或方式，设计出自然有效的交互方式。其中，支持可视分析过程的界面隐喻来自于人们在进行数据分析时的行为或使用工具。

交互的设计是基于人们熟悉的工具。这其中经常使用的一类界面隐喻是放大镜（Lens）。在现实中，我们可以通过镜头对感兴趣的区域进行观察。在可视分析中，采用镜头的概念作为设计的基础，可以对选择的数据区域进行不同的视觉表达，从而达到探索数据的目的。在实现时，放大镜可以是虚拟的视觉图元，支持双手触控的手指放大镜（FingerGlass）；也可以是物理的交互道具，比如纸镜头（PaperLens）就使用实际的薄片作为镜头，对数据进行探索。除了镜头这种隐喻外，还有许多基于熟悉的工具进行的交互设计。例如使用橡皮、笔的隐喻，作为修改绘图结果。在所见即所得的科学数据体可视化系统中，用户可以直接在体数据渲染的三维结果上进行涂色。

交互的设计是基于对人在分析数据时的行为观察。在可视分析领域，研究人员通过实验观察用户在分析任务时存在的行为。其中的一个例子是借用人们在选择物体时，通过圈定选择的方式，设计了套索选择的交互方式。在许多可视分析工作中，套索选择被作为一种基本的交互形式添加到系统中。

2. 多侧面、多尺度、多焦点交互技术

可视分析中常常会面对一些具有多个属性的数据集，用户一般会根据其感兴趣的问题会选择属性中的一个或者几个进行研究，期望从所研究问题的不同侧面来理解数据。例如在文献集合数据中，一个数据元素可以包含有发表年份、作者、发表期刊会议名称、关键词等信息。在这些数据中，分析者可能从作者关系、研究问题等多个侧面入手

来研究这个数据。新的可视化方法提供多种不同的形式，例如聚类视图、链接视图、网格视图、图视图、树视图等形式来描述数据在不同侧面之间的关系。

而在这些多属性的数据集中，通常会有一些属性是具有一定层次的目录（categorical）数据，对于这部分属性往往可以应用多层次交互技术。Ben Sheinderman最早提出的“Overview first，details on demand”就是多层次交互方法思想的最好诠释。一些最近的可视化工作中会在缩放的操作上加入更多的语义信息，即使用语义缩放（Semantic Zooming）来展示层次目录数据的内涵。在这类数据中，通过对这些属性进行上卷（Roll-up）和钻取（Drill-down）操作，可以查看数据在不同层次的聚集结果。对于大规模的图可视化方法，同样可以采取多层次的方法。不能缩放层次下可以对图进行基于密度的聚类并加以渲染，以避免严重的视觉干扰，并且保留图的主要特征。在图缩放到节点较少的层次时，可以将图的多变量特征使用矩阵的形式表现在节点与边中，从而便于用户探索图中的多变量特征。

用户在对复杂数据进行可视分析中，可能对多个数据子集产生兴趣，并期望研究它们之间的关系，一些可视分析方法则提供了多焦点的交互探索手段。使用者可以设置多组互相独立的查询语句来分别查询感兴趣的数据子集。同时作者也允许查询语句及其否命题之间相交的结果，并通过连边建立不同数据子集间的联系。在针对大数据情况下，众包的形式可以允许用户对图数据中不同的子集进行自定义的排布，然后通过改进的图布局算法，能生成结合了多子集排布结果的更好的综合图布局结果。

3. 面向 Post-WIMP 的自然交互技术

WIMP 是 Windows 窗口、Icons 图标、Menus 菜单和 Pointer 箭头的组合。自 Xerox Star 交互模型开始，WIMP 以极其简单的“点击”交互操作赢得了用户的青睐。Post-WIMP 模式是基于 WIMP，使用如手势、声音等其他输入来代替传统的鼠标。苹果公司所开发的 iPad 系列就是典型的后 WIMP 模型。虽然仍是窗口、图标、菜单和箭头这些操作控件，但以手势作为输入使得交互更为自然。将 Post-WIMP 应用到可视分析中为使用者提供自然的交互技术，有利于提高分析的效率。

基于多点触控屏幕的可视分析工作，其交互设计多是采用 Post-WIMP 的自然交互技术。通过触摸手势作为输入，另一种应用是开发大屏幕上的交互设计。由于大屏幕与常规的显示器不同，它的尺寸大，基于传统的鼠标键盘输入，用户不能有效地进行交互。通过移动的触摸输入面板，对大屏幕进行控制，适用于在大屏幕前移动观察细节的情况。

3.6.4　可视化的可用性

可视化的主要使用者是领域专家和终端用户，而并非可视化的专家。因此，亟需提升可视化的可用性。近年来，在如下领域获得了进展。

首先，是面向领域和大众的可视化库、框架和工具。

其次，是可视化与自动数据计算挖掘的结合。机器自动形成的可视化与人的交互式反馈结合，可以极大地简化可视化的任务。典型的案例是 VizDeck，它对查询进行分析，进而判断给定数据集的所有可能可视化形式以及用户如何与可视化交互，然而显示一个列表对各种可视化模板排序，用户的进一步交互能够帮助它优化可视化的策略。MIT 的 SeeDB 同样对所有可视化可能的空间进行探索，锁定和推荐最优的可视化方案。MIT 的另一个系统 Scorpion 能够自动发现可视化中的一些异常值（outliers），锁定对这些异常值具有贡献的数据记录，并且解释这些记录中导致异常值的共有属性。

客户端可视化与后端数据存储的交互也将变得更为容易。MIT 试图整合 ImMens 和后端的 Myria 多引擎数据库系统，并且通过声明式（declarative）的规格语言描述可视化和分析步骤，最优地在前端和后端之间分配计算。同时，通过对用户感知和交互的建模，能够更好地做负载优化和数据预取。

最后，协作可视化（Collaborative Visualization）协同与众包可视分析技术也是新的发展趋势。

3.7　大数据安全

大数据安全是个很宽泛的领域，可以包括：大数据系统的安全、数据自身的安全以及隐私保护、大数据应用带来的安全和隐私问题，以及大数据技术应用于安全领域。本节主要涉及前三个问题。数据安全随后引出了审计和问责，以及数据定价。

3.7.1　大数据系统安全

以 Hadoop 为代表的大数据系统栈早期主要处理公开领域的 Web 数据，因此并没有在安全上着力，但近年来有了长足的进展，逐步加入了用户和服务鉴权（基于 Kerberos），加入 HDFS 文件权限，对数据块的权限控制，对任务的授权，对网络上流动数据的加密以及 DataNode 内静态数据的加密等。Intel 的 Project Rhino 做了很多有益的尝试。

美国国安局使用的 Accumulo 在类似 BigTable 的架构上加入了新的安全特性，尤其是实现了一种基于 label 的、非常灵活的、细粒度（cell 级别）的访问控制。HBase 也基于 coprocessor 的架构实现了类似的 cell 级别访问控制，同时加入了用户透明的表加密。

在数据高度分布、去中心化的场景中（如物联网），数据的交换是匿名的、无需信任的、无单点控制、在网状网络（mesh network）随时发生，未来安全的责任可能会落到代理层（agent layer），而且可能通过数字货币的安全交易架构实现，如比特币的分布式加密总账技术区块链（blockchain），可以用于可容忍一定延迟的数据交换，而且一些新兴的架构（如 Ripple 和 simcoin）已经可以避免下载所有区块链，将延迟降到一秒以下。

3.7.2 数据自身安全

当整个业界在高谈全量数据的理念时，碰到的最大的问题是没有全量数据。数据的碎片化和孤岛化没有改观，主要有两个原因：一是没有经济上的激励机制，二是数据安全和隐私的忧虑。前者需要研究数据定价、微观经济学的价值交换机制。后者需要法律和技术的双管齐下。法律上，必须厘清数据权利 / 权力，例如数据拥有权（作为资产的所有权）、隐私权（什么不给别人）、许可权（什么给别人）以及许可后的撤销和转移权、审计权（有权知道被许可方怎么使用数据）和分红权（数据的外部性和持续产生价值后拥有者分享价值的权利），本书不作细谈。这里说一说技术上的一些进展。

数据安全首先是静态数据的安全，主要是访问权限控制，已经在 3.7.1 节中述及。

其次是动态数据的安全，主要是加密和动态审计能力。目前动态审计能力主要还是在企业内，表现为数据泄露防护（Data Leakage Prevention）技术，对重要数据进行分级、标识，实现跨平台（端点、移动设备、网络和存储系统）的统一管理。

技术和商业的发展也赋予个人控制自己数据的能力，比如浏览器对 Do Not Track 的支持，手机混淆 MAC 地址的能力。互联网服务提供商（如谷歌）和数据服务提供商（如安客诚）开始允许用户主动删除他们被收集的数据。同时，学术界也在寻找让个人获得更大自主权的能力，其中最典型的项目是麻省理工学院的 OpenPDS（personal data store）。OpenPDS 允许个人收集和存储自己的数据和元数据，并在保护隐私的前提下向第三方提供细粒度的访问控制。其中元数据隐私保护通过问 – 答系统 SafeAnswers 来实现。

最后，在开放数据运动中，需要对数据脱敏，对隐私进行保护。常常采用的方式是去标识化，或匿名化，比如去掉与人名、地址等相关的内容（这些能够精确唯一标识个人的属性被称为标识符）。去标识化不一定能够做得彻底。美国方面做了研究，只要有性

别、出生年月以及邮编这三个数据，就有很大的把握能够把个人的信息还原出来。这些通过组合能够以较大概率标识个人的属性被称为是准标识符（quasi identifiers）。

重新标识化（re-identification）往往通过多数据源的相互匹配。美国有案例把匿名的搜索信息与选举人登记信息匹配，把个人重新标识。Netflix 开放了匿名的评论以及打分的信息，但是当跟国际电影数据库 IMDB 数据进行匹配时，有用户（包括一个有同性恋倾向的人）被识别了出来。

另外一种重新标识的可能性是基于统计。麻省理工学院的研究人员发现，如果获悉一天之中个人在四个不同的时间点所处的不同地点，有 95% 的可能把这个人找出来。基于 Netflix 数据研究揭示，因为数据的稀疏性和不相似性，只要获得某人区区几个评分信息，就能够有较大的概率将其标识出来。

现在已经出现不少隐私保护技术，来满足不同的隐私需求，如避免被确认某人是否在某个数据集中，或某人是否有某个特殊属性，或某记录是否对应于某人。比较典型的技术有 K-anonymity、L-diversity、T-Closeness 等：

- K-anonymity 保证对于每一个准标识符相似组（组内所有记录都具有相似的准标识符值），都至少有 *k* 个记录。但如果这 *k* 个记录的敏感属性（如薪水或病症）都具有同样的值，仍然会导致信息泄露。
- L-diversity 对此进行了改进，要求在每个准标识符等值组（组内所有记录都具有相等的准标识符值）内，敏感属性都至少有 *l* 个不同的值；
- 而 T-Closeness 要求敏感属性在任一准标识符等值组内的分布与其在整个数据集内的分布近似。

上述方法仍然有被攻击的可能性，特别在敏感属性不够多样化，或攻击者具有背景知识时。另一种叫作差分隐私（differential privacy）的匿名化技术会把噪声加入到数据集中，但仍保持它的一些统计属性（如果它们对个体数据的改变不敏感）。这种技术可以很好地支持典型的机器学习方法（如分类、聚类和频繁模式挖掘），并且最近在运营商开放数据中得到了应用[㊀]。

当然，所有这些隐私保护方法都是牺牲原始数据的质量来获得高匿名性。因此，需要在隐私安全性和数据可用性之间做好平衡，当噪声大到一定程度，数据可用性会变差。从实践意义上来说，隐私保护的成本一定要低于数据本身的价值。

㊀ http://www.technologyreview.com/news/514676/how-to-mine-cell-phone-data-without-invading-your-privacy/

3.7.3 数据使用安全

数据开放代表了时代的潮流，但是在很多商业场景中，数据不能无条件向公众开放，而只能做点对点的共享，或者基于某种特殊约束的多边交易。大数据时代，更典型的一种机制是生产资料的租赁制：我不一定拥有整个数据集，但可以基于特定的目的租赁相关的数据进行计算。租赁和使用的过程中要保证数据的权利，数据可以使用，但不可以看见，“可用不可见，相交不相识”。

其实这不是个新问题，姚期智先生在 1982 年提出“millionaires’dilemma”问题：两个百万富翁谁都不愿意说出自己有多少钱，但他们希望比较谁更有钱。这就是典型的“可用但不可见”场景。这个场景有很多的实际应用，比如美国国土安全部有恐怖分子的名单，而航空公司有飞行记录。国土安全部向航空公司索要飞行记录，但遭到拒绝，因为涉及乘客的隐私承诺。而国土安全部也不可能提供恐怖分子名单，因为这是最高机密。在双方都不能看到对方数据的前提下，怎么能够把两份数据放在一起产生价值（如确定恐怖分子的行踪）？这是技术上试图解决的重要问题。

现在的主流技术包括：基于同态加密、支持 SQL 的加密数据库，基于加密协议的多方安全计算，基于可信计算环境的多方安全计算，基于隐私保护的机器学习算法等。

同态加密数据库技术以 CryptDB/Monomi 为代表。它采用了所谓的洋葱加密法（onions of encryption），一层一层地对数据库表的敏感数据列进行加密。而为了支持 SQL 语义，它采用了同态加密算法。比如：对于需要做相等性判断的数据，保证相同的明文产生相同的密文（这样可以在密文上做相等性判断）；对于偏序语义，两个值明文的大小关系与它们密文的大小关系相同；对于相加、查找等操作，以此类推。这样，Monomi 运行 TPC-H 的所有查询在加密数据库上执行。这种设计，一方面防止了现在常常出现的内部人员导致数据泄露问题，另一方面允许数据供他方分析，但又不至于被他方盗取数据。

加密协议是姚先生最早实现多方安全计算的方法。典型的加密协议允许双方或多方基于各方的输入数据联合计算一个函数，同时保证输入数据的私密性。它的缺点是只能处理相对小规模的结构化数据，计算复杂度较高，多用于特定的应用场景和数据使用模式。

最近，另一种实现“可用但不可见”的技术得到了应用，这种多方安全计算技术基于可信计算环境。各方经特定的密钥管理机制获得密钥，对数据加密后送入可信执行环

境，密文在可信虚拟机中被解密，交由经过验证的数据处理逻辑处理。它的优点是允许任何应用和数据类型，支持更大的数据规模。它的缺点是需要各方把数据送入可信执行环境，在广域数据交换场景中，限制了数据的规模。这将可能催生涵盖分布式可信计算环境、分布式计算范例（如 MapReduce）和加密协议的多方计算研究。

当前的可信计算环境主要基于 IntelTrusted Execution Technology/Trusted Platform Module（TPM）和 VT-d，未来将转向更为安全的 Software Guard Extensions（SGX）。对于前者，数据在隔离环境的内存中被解密，而对于后者，数据在内存中仍以密文形式存在，进入 CPU 内部后才被解密。

我们期待这些技术在现实的数据共享和交换场景中得到广泛应用，以滋润新兴的数据经济。案例一：两家电商，一家卖衣服鞋帽，一家卖化妆品，他们对用户的画像都非常片面，如果把两者通过多方安全计算交换数据或帮助数据较少一方实现“冷启动”，将能够对用户全面刻画。案例二：癌症是一个典型的长尾病症。过去 50 年癌症的治愈率只提升了 8%，其中一个原因是每个科研机构拥有的基因样本非常有限，而美国医治保险携带和责任法案（HIPAA）对于数据共享具有严格限制。如果能够通过多方安全计算让不同的科研机构把这些样本聚拢成为更大的数据集，那在癌症治疗技术上面就可能有所突破。

3.7.4　审计和问责

在数据被授权他方使用后，最重要的问题是使用是否产生滥用，使用是否符合法律法规（如健康领域的 HIPPA 或金融领域的 GLBA），使用是否符合双方或各方同意的隐私条款。

在多方计算中，加密协议或可信执行环境能够保障数据传输和计算中人不能看到数据，但分析程序是看得见数据的。谷歌开始在 Gmail 加入广告时，通过隐私条款和自律保证只有机器程序会读取邮件内容（而谷歌的员工不能）。而数据交换涉及多方，光自律是不能保障数据安全的，为了保证分析程序不做坏事，不暗地里偷数据，要对其数据使用进行审计。

因此，需要解决两个问题：一是把数据安全或隐私条款形式化，通过一种规范语言描述；二是根据形式化的规约，对数据的使用进行审计。

关于数据安全或隐私条款，需要通过规范语言将其形式化，确定条款涉及的主体（如数据拥有方和使用方）、数据的安全属性、时间约束、授权目的、动作谓词（如授权

或取消）、事件处理（出现违约时的动作）等。基于上述形式化规则的合规性检查，即需要传统的形式化方法（如模型检测），也需要语义分析的支持，比如，授权目的的描述可能是比较模糊的。

关于使用的审计，可以有不同的实现方式。

一种是系统具有数据引用的检测器（reference monitor），在数据使用中产生日志，在使用后基于日志进行合规性的审计。如果出现违约，通过进一步的分析对使用主体进行问责。

另一种是实时审计，数据违约在发生之前被阻止。目前来说，为了做到实时检查，数据安全或隐私的规约相对比较简单。

这两种方法都需要对信息流进行建模，从而形成所谓的数据血统（data lineage）。数据血统是数据治理（尤其是生命周期管理）中的重要概念，但多工作在粗粒度，而这里需要数据使用的非常细粒度的跟踪和建模。

3.7.5 数据定价

数据定价未必属于数据安全的范畴，但数据安全的基础设施有助于数据作为商品的定价。

数据和信息的一个重要区别就是，信息是具有特定意义的数据，因为特定意义存在，所以信息价格很好估计。但数据在还没使用的时候其价值是不确定的，就像赌玉石，不把它剖开来谁也不知道它的价值。更重要的是，汽油的价值在其燃烧的一瞬间兑现并消失，但数据跟传统物质资源不同，它是可以反复使用的。而且数据具有外部性，当应用于不同领域时可能产生超出其采集时预期的价值。

因此，这种测不准现象决定了数据定价的常见场景：不为买断式的数据产权变更定价，而是为一次使用形成的价值变现定价，先使用后定价。

在多方安全计算中，可以根据数据血统反推数据使用中各方数据的贡献，基于此来定价。即使是企业的内部数据资产，也可以根据使用来估值。某些数据被用得多，它也就估值更高。

数据作为一种商品，必然具有符合商品的价值规律——即效用和稀缺性。效用方面，我们前面提到在多方计算中根据数据血统反推贡献进行估值。此外，还需要考虑稀缺性（类似供求关系）。比如在多方计算中数据 A 的贡献度虽然不太高，但它非常稀缺，具有不可替代性，我们就需要调高它的估值。反之，我们需要调低某些贡献高的数据的估

值。解决这个问题，可以给每个数据加上一个权值来标识其稀缺性。这个权值可以由数据拥有者初始化，并在数据交易中由市场来调节。

数据的稀缺性也某种程度上反映了数据的价值密度。当一类数据很稀缺时，高估值会激励公司和个人来收集和产生类似或同质的替代数据，从而由市场进行价格的自动调节。

数据共享和交易中的一个重要挑战是要防止劣质数据。多方数据相遇时，如果一方混入了劣质数据，将影响最终结果的价值。

可见，在数据审计、问责和定价的基础设施达成后，数据经济将会勃发，数据交易将极大繁荣。比如数据市场上，较之今天以买卖数据集的产权为主的方式，未来租用和付款一站式的交易将把产值提升几个数量级。而数据交易所能进一步基于市场进行价值发现和定价，像股票交易所那样，每天产生亿万计的小批量、高频率的数据交易。

第 4 章　大数据 IT 产业链和生态环境

马云曾经说过：“人类正从 IT 时代走向 DT 时代。”随着大数据、云计算、移动互联网技术不断发展，我们已经从信息社会进入大数据（Big Data）时代。

大数据是继互联网、云计算、移动互联网之后的又一个新热点，是新一代通信技术的代表。大数据在推动技术创新、改善公共服务等方面有着重大应用前景，各国政府纷纷布局大数据战略，推动大数据产业的发展。

随着我国信息产业宏观环境不断改善，各地方政府纷纷出台大数据发展计划，各高校和科研组织开始培养专业的大数据人才，而以 BAT 为代表的科技企业也开始涉足大数据产业。目前通过各方努力，我国已具备加快大数据产业发展的基础，大数据产业链正在加速形成。

然而，我国的大数据产业发展还处于起步阶段，大数据产业还存在一些问题，如大数据相关的法律法规有待进一步完善，对于信息安全保护还需提高重视等。因此，我国大数据产业还需要在国家宏观指导下进一步发展。

4.1　国内外大数据产业链现状

4.1.1　大数据产业链全景图

大数据产业包括与大数据的生产与集聚、组织与管理、分析与发现、应用与服务相关的所有活动。数据产业链按照数据价值实现流程，包括生产与集聚层、组织与管理层、分析与发现层、应用与服务层四大层级，每一层都包含相应的 IT 技术设施、软件与信息服务，如图 4-1 所示。

全球互联网公司纷纷进入大数据产业，希望瓜分这一市场。据电信与媒体市场调研公司 Informa Telecoms & Media 在 2013 年的调查结果显示，全球 120 家运营商中约 48% 的运营商正在实施大数据业务。该调研公司表示，大数据业务成本平均占到运营商总 IT 预算的 10%，并且在未来 5 年内将升至 23% 左右。可见大数据巨大的产业价值已经逐渐被企业重视。

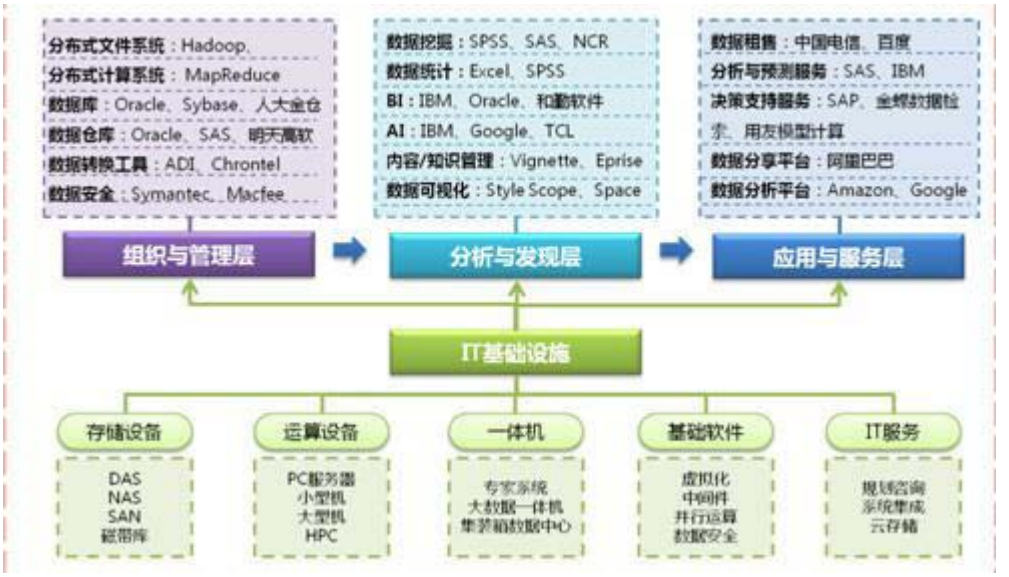


图 4-1　大数据产业链全景图㊀

而根据数据从产生到应用，继而产生新数据的过程，大数据产业形成了一个环形产业链（见图 4-2）。从数据产生到应用，参与企业逐渐增多，数据价值逐级增加。概括起来这个环形产业链主要包括以下几个方面：以云计算、物联网、移动互联网等新一代信息技术而不断生产交易数据、交互数据与传感数据的大数据生产活动；以搭建大数据平台、支撑大数据组织与管理的服务器、存储设备、网络设备、数据中心附属设备等 IT 基础设施硬件的销售与租赁活动；大数据平台的运维与管理服务，系统集成、数据安全、云存储等解决方案与相关咨询服务；支撑数据分析与发现的嵌入式芯片、服务器、高性能计算设备等 IT 基础设施硬件销售与租赁；与大数据应用相关的数据出售与租赁服务、分析与预测服务、决策支持服务、数据共享平台、数据分析平台等。

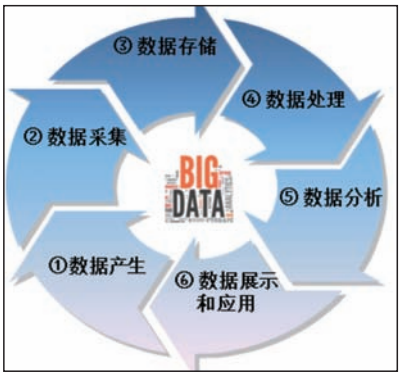


图 4-2　大数据环形产业链㊁

随着每次数据产生到数据价值实现的循环过程，数据规模不断扩大，数据复杂度不断加深，数据创造的价值不断加大，同时，也加速大数据技术创新与产业升级。

㊀ 图片来源：赛迪顾问产业数据库，2012，12。
㊁ 图形来源：赛迪顾问产业数据库，2013。

4.1.2 产业链上中下游

随着现代技术不断发展，从最初的“在互联网上，没有人知道你是一条狗”，到现在的“不仅知道你是一条狗，还知道是一条什么样的狗”。随着大数据时代来临，围绕大数据的生产与集聚、组织与管理、分析与发现、应用与服务的各级活动正在加速构建。

信息数据产生作为大数据产业链的第一个环节显得尤为重要。随着新技术不断发展，数据产生的方式也越来越多样，例如，人们每天使用的互联网和无线通信，即时通信、微信、微博、手机电话、短信、彩信甚至是每一次在互联网上点击（通过点击习惯可以分析经常浏览某类网站，喜欢某类商品，以及上网时间等使用习惯）都会留下记录，带来爆炸性的数据增长。大数据产业的上游是一批能够掌握大数据标准、入口、汇集和整合过程的公司，它们在大数据存储、使用和分析的基础上推出个性化、精准化和智能化的机制，跨网站、跨产品、跨终端、跨平台，让人与人、人与物、物与物之间实现高效撮合与匹配，从而建立起崭新的商业模式，这些企业包括 Google、百度等。

随着信息数据的爆炸式增长，对数据存储提出了更高的要求。虽然存储设备是整个产业链中技术含量最少的，发展空间也较窄，但却可能是增长最稳定的子行业。不仅数据采集大多来自半结构化或非结构化数据，具有零散性，而且许多数据的产生是散乱和随机的，因此数据整理显得尤为重要。大数据产业的中游是在某些垂直领域或者某些特定区域能够掌握大数据入口、汇集和整合的一批公司，它们掌握全部网络用户的部分网络行为，或者是部分网络用户的全部网络行为。这些公司需要大规模地采集信息数据，并有效地剔除噪音数据，可见未来这些公司有机会在这些垂直领域或特定区域成为规则制定者和商业模式创新者，例如 IBM、EMC 等。

大数据产业链的最末端是信息数据的展示和应用环节，也是最具技术含量和产业价值的行业。大数据产业的下游由网络公司组成，它们基本上扮演的角色是大数据生态圈里的数据提供者、特色服务运营者和产品分销商，主要通过开放平台和搜索引擎获取用户，处于产业的边缘地带。任何数据不经过分析这一环节，都无法落实到实际应用。而且，在同样的数据面前、谁分析出的结果最有效，谁才是真正的“大数据”智能产业领跑者，例如 SAP 等。

4.1.3 大数据产业链发展趋势

1. Gartner 对大数据趋势的预测

在 2014 年著名咨询机构 Gartner 发布的技术成熟度报告中，大数据技术作为最热的新技术已经被物联网取代，并且大数据技术过渡到“幻觉破灭阶段”。这并不是一件坏事，并且

表明大数据市场开始变得成熟，变得更贴合实际应用，大数据产业将变得更加商业化。

在过去两年里，围绕着大数据的技术已经被过度炒作。以 Hadoop 为核心的生态环境、NoSQL 数据库、内存数据库和其他大数据技术，已经能够使公司或组织存储、处理和分析大量数据。而且，投资公司在大数据领域投入了数十亿美元，但 Hadoop 和 NoSQL 公司几乎不盈利。为了获得规模和主导地位，大数据公司都在不断扩展，未来几年可能会见效益。对于 Gartner 的报告，我们更多地应该从科学技术的角度去观看，但不应过度解读。

Gartner 总结的技术成熟曲线分为五大阶段：技术萌芽期（Technology Trigger）、期望膨胀期（Peak of Inflated Expectations）、泡沫化的谷底期（Trough of Disillusionment）、稳步爬升的光亮期（Slope of Enlightenment）和最终的产品成熟期（Plateau of productivity）。从 2011 年到 2014 年，大数据技术经历技术萌芽期、进入峰顶并开始快速下滑。大数据技术滑向谷底进入复苏期，最后走向成熟，这可能需要很多年。

关于大数据技术本身，Gartner 在 2014 年 7 月发布的技术成熟度曲线如图 4-3 所示。从图中可以看到，Gartner 预测内存计算技术作为一项提升大数据处理效率的技术手段，将会在两年内达到成熟期；物联网作为大数据技术的应用方向之一，目前处于关注的顶峰；基于 SQL 的 Hadoop 查询接口也是关注的热点，预期在 2 ~ 5 年内发展成熟；虽然数据开放在学术界的呼声很高，但预期在 5 ~ 10 年内才会进入成熟期。

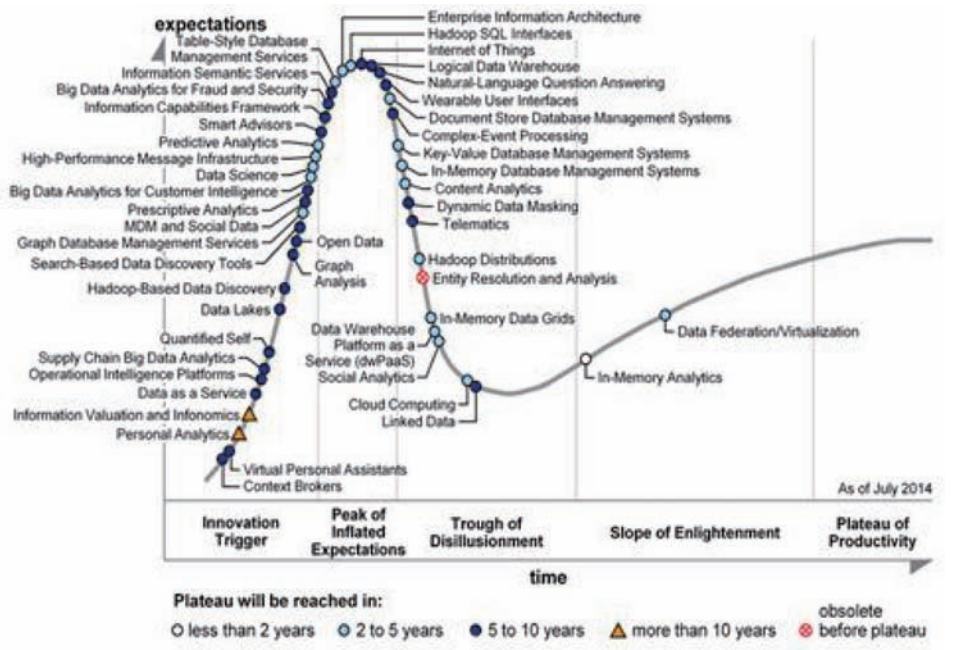


图 4-3 （2014）Gartner 大数据技术成熟度曲线

大数据中的创新将持续发展，并且创新的催发机制已经很充足，大数据在经过泡沫期后的谷底期的历程将十分迅速。这种持续的增长将所有的供应商进行洗牌，留下那些紧跟潮流的供应商。这种趋势可以促进很多初始的并且处于创业初期的项目的发展，大数据将实现从系统创新到差异化的目标导向的发展。大数据将会被纳入市场现有的解决方案中，同时将会取代现有方案的一些功能。总而言之，大数据市场将会逐步进入一个更加合理的发展方向，新的技术和实践将会嵌入既有的解决方案。

2. 大数据产业链发展趋势全景图

2012 年，FirstMark 资本的 Matt Turck 绘制了大数据产业链趋势全景图 2.0 版，如图 4-4 所示。该图将数百个大数据创业公司和 IT 厂商根据产品和商业模式划分为基础设施、分析、应用、数据源、跨基础设施分析、开源解决方案 6 大类，共计 38 项。2014 年，Matt Turck 从一个风险投资者的角度对两年来大数据市场的最新发展进行了深入研判，对未来趋势进行解读，并绘制了大数据产业链发展趋势全景图 3.0 版，如图 4-5 所示。

图 4-4　大数据产业链发展趋势全景图 2.0 版

图 4-5　大数据产业链发展趋势全景图 3.0 版

在大数据产业链发展趋势全景图 3.0 版中，一共有 358 家公司，和 2.0 版对比只有 16 家公司退出（被收购或者上市），只占 4.5% 的份额。其中，退出最多的是数据分析（7.5%）以及基础设施（4%），而数据源和前端分析类均无企业退出。

在大数据产业链发展趋势全景图 3.0 版中，最热门的分类分别是数据分析、基础设施和应用 3 个大类，如图 4-6 所示。

从大数据产业链发展趋势全景图 2.0 版和 3.0 版的对比分析中可以看出，大数据产业链有几个

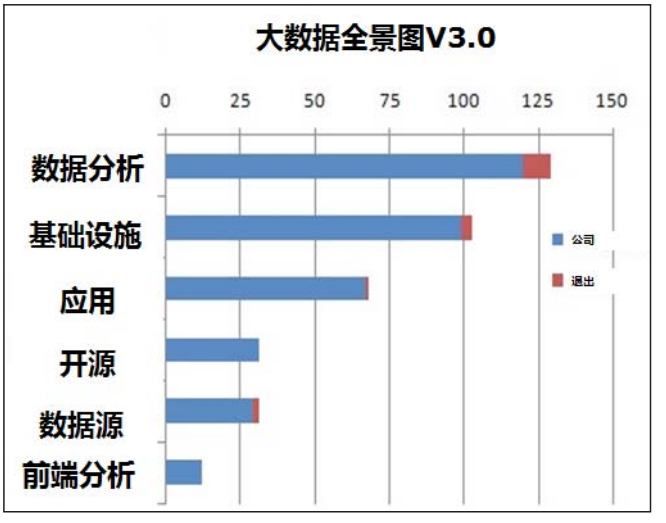


图 4-6　大数据产业链分类

关键的变化趋势，下面分别介绍。

（1）竞争加剧

创业者纷纷涌入大数据市场，尾随的风险投资商也是挥金如土，导致大数据创业市场目前已经非常拥挤。一些创业项目类别，例如数据库（无论是 NoSQL 还是 NewSQL）或者社交媒体分析，正面临整合或去泡沫化（随着 Twitter 收购 BlueFin 和 GNIP，社交分析领域的整合已经开始）。虽然大数据创业市场已经人山人海，但是依然有足够的空间给新的创业公司，现阶段大数据基础设施和分析工具领域的创新吸引了大量的资金，当然，这类大数据创业本来就是资金密集型项目。

（2）大数据市场尚处于初期阶段

虽然大数据的概念已经热炒了数年，但该领域依然处于市场的早期阶段，虽然过去几年类似 Drawn 和 Scale 这样的公司失败了，但是相当多的公司已经看到了胜利的曙光，例如 Infochimps、Causata、Streambase、ParAccel、Aspera、GNIP、BlueFinLanbs、BlueKai 等。还有不少大数据创业公司已经形成规模和气候，并且获得了海量融资，例如 MongoDB 已经募集 2.3 亿美元，Plalantir 已经募集 9 亿美元，Cloudera 已经募集 1 亿美元。但是就成功的 IPO 或公司而言，市场尚处于早期阶段（虽然已经有 Splunk、Tableau 等成功 IPO）。此外，目前阶段一些传统 IT 巨头已经展开了收购大战，例如 Oracle 收购 BlueKai，IBM 收购 Cloudant。在很多大数据创业领域，创业公司依然在为市场领袖的地位展开混战。

未来几年是大数据市场竞争的关键时期，企业的大数据应用从概念验证和实验走向生产环境，这意味着大数据厂商的收入将快速增长，市场将继续扩大。根据 ICT Research 关于大数据在世界范围内近几年的市场走势及未来发展趋势的预测，大数据在 2015 年到 2016 年间市场规模会有大幅度增长。

（3）从炒作回归现实

大数据在经历数年的火热后已经归于平淡，但这恰恰是大数据真正落地的开始。未来几年是大数据市场竞争的关键时期，企业的大数据应用将从概念验证走向生产环境，这意味着大数据厂商的收入将快速增长。当然，这也是一个检验大数据是否真的有“大价值”的时期。

（4）大数据基础设施

虽然 Hadoop 已经确立了其作为大数据生态系统基石的地位，但市场上依然有不少 Hadoop 的竞争和替代产品，这些产品还需要时间完善。基于 Hadoop 分布式文件系统

的开源框架 Spark 近来成为人们讨论的热门话题，因为 Spark 能够弥补 Hadoop 的短板，例如提高互动速度和拥有更好的编程界面。而快数据（实时）和内存计算也始终是大数据领域最热门的话题。一些新的热点也在不断涌现，例如数据转换整理工具 Trifacta、Paxata 和 DataTamer 等。

（5）大数据分析工具

就投资者的投资比重来看，大数据分析可谓是大数据市场最活跃的领域。从电子表格到时间线动画再到 3D 可视化，大数据创业公司提供了各种各样的分析工具和界面，有的面向数据科学家，有的选择绕过数据科学家直接面向业务部门，由于不同的企业对分析工具的类型有不同的偏好，因此每个创业公司在自己的细分领域都有机会得到发展。

（6）大数据应用

大数据应用的发展进程相对缓慢，但目前阶段大数据确实已经进入了应用层。从大数据产业链发展趋势全景图 3.0 版中可以看到，一些创业公司开发了大数据通用应用，例如大数据营销工具、CRM 工具或防欺诈解决方案等。还有一些大数据创业公司开发了面向行业用户的垂直应用。金融和广告行业是大数据应用起步最早的行业，甚至在大数据概念出现之前就已经开始了。未来大数据还将在更多行业得到广泛应用，例如医疗、生物科技（尤其是基因组学）和教育等。

4.2　产业链和生态环境的瓶颈和建议

4.2.1　大数据发展产业链和生态环境的瓶颈

1. 数据开放程度低

大数据的拥有者包括政府部门、运营商和互联网大企业等。目前政府部门的数据开放程度还相当低，可以说处在探索阶段。政府大数据“富矿”可供全民开采，但总体来讲，政府部门数据开放和共享在国内还未形成一定规模，不管是开放的数据类型、数据量、开放数据的形式，都还有很大提升空间。李国杰院士总结的阻碍政府信息共享的因素包括：不愿贡献、不敢共享、不能共享；缺乏政府数据共享的理念，对治理现代化认识不足；缺乏政府数据共享机制的责任主体，怕犯泄密错误，宁可不作为；缺乏数据共享的法规和制度，无法可依或者法律法规间相互冲突；缺乏政府数据共享的统一标准和规范；缺乏治理机制设计。而运营商、互联网大企业等，由于存在竞争关系，数据的开放程度也很低。政府部门和运营商之间的数据共享和合作也需要进一步加深。

2. 大数据技术问题

在大数据的采集、预处理、存储、分析等技术方面，互联网企业有一定的技术基础，已经摸索出一套适用的大数据技术体系。目前，在大数据产业发展初期，这些企业在现有大数据存储处理分析技术的实际应用层面上不存在明显的瓶颈。在大数据存储产品方面，国内的一些厂商推出的大数据存储平台在技术方面与国外商用、开源的平台有类似的地方，相对来说技术差距不大。对政府部门和运营商来说，可以通过自主研发、技术外包和购买商业产品的方式来获取大数据存储处理分析技术，相对于其他方面，比如数据共享、数据安全等，即使技术是瓶颈，但也不是首要的因素。然而，从更高层次和长期来看，大数据技术也受制于存储 I/O、网络传输、计算机架构等方面，确实存在瓶颈。在智能分析、处理性能、可视化、安全和隐私保护的研究、应用方面也都需要突破。比如 Hadoop 平台不易学、不易编程等，这在一定程度上妨碍一些企业使用 Hadoop 平台存储和处理生产数据。再比如，数据分析的结果采用可视化的方式展示，将大大提升用户对数据的理解，能更快、更准确地进行决策。只有提高大数据技术和平台的易用性、实用性、安全性等，大数据技术才能更好地服务于实际应用。

3. 相关政策法律不够完善

在数据开放方面，政府信息公开制度是承认公民对国家拥有的信息有公开请求权，国家对这种信息公开的请求有回答义务的一种制度。在世界范围内，政府信息公开制度都是各国普遍推行的一项制度，全球已有 30 多个国家制定了专门的政府信息公开方面的法律。我国第一部系统规范政府信息公开的全国性法律文件——《中华人民共和国政府信息公开条例》（下称《条例》）于 2008 年 5 月 1 日起正式施行。然而，我国的政府信息公开立法存在着立法技术粗糙、立法层级低、受制于其他更高层级立法掣肘的现实问题。《条例》未明确规定公民享有获取政府信息的权利，反而将不予公开的政府信息在总则中加以强调。《条例》本身立法层级低于《中华人民共和国保守国家秘密法》等法律，因此在确定政府信息公开范围等方面，完全受制于国家秘密的认定。因此我国政府在信息公开制度上还处于起步阶段。在标准化方面，工业和信息化部（下称工信部）促进全国信息技术标准化技术委员会开展了大数据标准化的需求分析、标准体系框架研究及相关标准研制工作，并向相关国际标准化组织提交了大数据研究提案。在省一级大数据立法方面，广东省率先一步，《广东省信息化促进条例》已于 2014 年 9 月 1 日起施行，这意味着包括大数据在内的多个信息化产业在广东省获得了立法扶持。除此之外，关于数

据所有、数据采集、存储、使用权责方面的法律还是空白，这些在宏观层面上直接关系到大数据应用以及大数据的数据安全。国务院发展研究中心信息一处处长李广乾称，中国只有针对国家秘密的国家保密法，缺少保护个人隐私和商业秘密的专门法律，有针对政府信息公开的政府信息公开条例，以及一些部门的法规，比如统计法，但是总体来说约束力不足，尚没有涉及个人隐私保护和商业秘密保护的专门法律。根据工信部 2014 年立法工作安排，目前工信部没有直接涉及大数据产业立法的项目。

4. 数据安全问题

数据的安全一直以来都是个大问题，尤其是在大数据时代，数据的采集、处理、存储、分析等基本上都是分布式的，涉及的终端设备、应用系统和参与方较多，又由于数据量巨大，数据安全问题更加难以解决。比如苹果手机作为一个终端，它存在后门，用户的个人数据经由苹果手机存储处理，很有可能会被窃取。目前有人提出对少量的敏感信息可以做加密处理，存储的是密文，在搜索和计算时，可以借助加密算法的特性直接对密文做相应操作，而这种算法的计算代价高，应用范围较小，不适合于真正的大数据。当前主流的大数据平台自身的安全也存在问题，比如系统缺乏一些基本的安全机制、系统存在漏洞等，与传统的关系型数据库相比，在安全性、可靠性等方面还有待提高。

目前一些大数据技术公司已经开始收购一些安全公司，以弥补大数据安全缺陷。Hadoop 软件和服务公司 Cloudera 公司已经在其软件方案中融入了众多安全措施。例如，Kerberos 能够管理哪些用户有权访问特定 Hadoop 集群，Cloudera 公司将其打造成一系列开源技术成果并作为该公司 Hadoop 发行版的组成部分。Cloudera 公司还亲自推动其他技术方案的普及，包括用于管理哪些用户（或者应用）有权访问保存在 Hive 以及 Impala 环境下的数据与元数据的 Apache Sentry 项目。Cloudera 公司也积极与芯片厂商 Intel 公司合作进行芯片级别的加密机制研发，Rhino 项目就是 Cloudera 公司与英特尔公司进行紧密合作后努力打造出的重要成果。在 2014 年 Cloudera 公司收购了加密厂商 Gazzang，专门为大数据环境提供加密方案，使大数据可以真正走入企业生产环境。Gazzang 的技术方案包括一款用于对 Hadoop 环境内存储数据进行加密的产品，外加一套用于管理哪些用户有权访问密钥、令牌以及其他数据访问类协议的解决方案。Gazzang 技术方案还有助于改进 Rhino 项目。类似的大数据技术公司 Hortonworks 也收购了安全新兴企业 XA Secure，提供基于身份的授权、审计、治理等功能。大数据技术公司在数据安全技术上的投入，有助于促进企业将大数据工作流从研发环境部署到生产环境，甚至部署到公有云中。

4.2.2 大数据产业链和生态环境发展建议

1. 促进数据开放和共享

在数据开放问题上，对于政府和科研机构的核心数据，可以以国家或行业规定的形式，促使其开放非涉密的政府数据及科研数据，成立专门部门进行管理、协调，做到数据交换共享，满足国家战略需求；对于政府和科研机构的非核心数据，以及企业的数据，应以市场为导向，以数据交易的形式进行。对于一些涉及隐私和秘密的信息，可以做变相的共享和交易，比如说数据可以用，但不可以看见（姚期智在 1982 年提出的“millionaires’dilemma”（百万富翁的窘境），就是典型的“可用但不可见”场景）。

上海市政协委员、上海高智科技发展公司董事长刘幸偕认为，出于安全考虑，有些信息是不应该拿出来共享的，哪些能共享、哪些不能，这需要建立一个规范。数据分享出来之后，数据的使用者在数据使用过程中要有审计，确保使用者对数据的操作跟预先设想的一样，不会偷偷泄漏数据。在数据交易方面，需要确立数据定价机制，由于双方数据的价值不一定对等，产生的洞察对各方的用途也不一样，因此基于定价机制，要比大锅饭式的数据共享更有激励性。

针对政府数据开放，李国杰院士的课题组提出六项原则：1）开放原则：政府信息资源以共享为原则，不共享为例外。注意开放与保护隐私的平衡。2）保障安全原则：根据安全等级确定数据共享的范围。3）价值导向原则：开放的政务数据资源应具有经济价值和社会价值，共享的目的是促进数据资源的利用。4）质量保障原则：内容完整可信，数据格式方便使用，内容及时更新。5）责权利统一原则：政务数据拥有部门承担数据开放的责任，依法明确可开放数据的范围。用户对下载后数据的使用行为负责。6）数字连续性原则：被开放的政务数据资源应维护其数字连续性，可检索、可呈现、可理解、可被发现，保证可持续再用。此外，也需要建立评价政府开放数据的量化标准。

李国杰院士提出了评价政府开放数据的五维度框架。1）政策与立法：相关的立法文件和政策；政府数据共享的程序和标准。2）技术：共享数据格式的规范程度；数据质量如何；数据是否通俗易懂；数据更新和维护水平等。3）管理架构与组织形式：监督和合作机制是否恰当，是否明确谁承担保证数据质量的责任；是否提供激励机制让公共部门提供数据（包括对公务员的激励机制）；是否明确谁承担费用，维护政府开放数据的金融机制和费用模式等。4）沟通交流：政府开放数据是否形成良性生态系统；企业界的支持程度；公众对政府开放数据的兴趣；反馈渠道是否畅通；公众对政府开放数据的满意度等。

5）效果和效益：政府数据共享产生的经济效益和社会效益；数据共享对政府本身工作效率的提高等。

2. 推动大数据分析在各个行业中的应用实践

当前阶段，大数据的采集、预处理和存储，对于企业、政府部门等，都不是难点，主要的难点在于对数据的分析也就是价值提取。像腾讯、阿里巴巴、百度、中国移动等企业，对数据的利用和价值提取已经有一定的技术积累，也建立了各自的数据分析模型和利用大数据盈利的模式。对政府部门、小中型企业来说，其价值提取能力相对来说弱一些，一方面是因为技术基础薄弱，一方面是没有建立合适的数据分析模型。首要任务是尽快培育一些专门做价值提取服务的企业，为政府部门、小中型企业做数据分析。

在国外，比如沃尔玛这样的大型企业会和别的公司合作进行数据分析。国外专门做数据分析的公司比如 Palantir 公司，很擅长给各类政府和金融机构提供数据价值提取服务。还有 Kaggle 这样的公开数据建模和数据分析平台，企业和研究者可在其上发布数据，统计学者和数据挖掘专家可在其上进行竞赛以产生最好的模型。Kaggle 的初衷就是试图通过众包的形式使数据科学成为一场运动。根据 Kaggle 官方提供的数据，Kaggle 在全球范围内拥有将近 20 万名数据科学家，专业领域从计算机科学到统计学、经济学和数学。Kaggle 的竞赛在艾滋病研究、棋牌评级和交通预测等方面取得了成果。Kaggle 这样的服务形式的优势就在于能吸引众多数据分析专家，为企业提供了优质的数据分析服务，这是值得国内公司借鉴的。

3. 制定大数据政策和相关法律法规

大数据在我国的发展处于初期阶段，相关法律政策一片空白。政府应为大数据产业发展和生态环境持续良性进化制定积极的政策法规，创建适度宽松的大数据发展环境。国家改革和发展委员会副主任徐宪平对大数据的观点包括：首要任务是推动数据开放，关键要推动数据应用，根本之策是推动数据立法。数据的保护和公开是一体两面，邬贺铨院士呼吁要尽快制定“信息保护法”和“信息公开法”，既要鼓励面向群体而且服务于社会的数据挖掘，又要防止针对个体侵犯隐私的行为，提倡数据共享又要防止数据被滥用。

这方面可借鉴的国家包括：英国制定了《数据保护法》和《信息自由法》来平衡个人隐私数据保护和政府信息公开之间的关系；美国的政府信息公开制度则由《信息自由法》、《隐私权法》和《阳光下的政府法》三部分构成，共同维护着公民的知情权与隐私权，同时也在一定程度上保护了政府的利益。

4. 保障全生命周期的数据安全

对于数据本身的安全，需要分析各行业对大数据的安全需求，分而治之。像电信运营商、金融行业、医疗行业、政府组织对数据的安全需求各不相同，有各自的保护重点，比如数据的完整性、可用性、保密性等，需要不同的安全策略和技术等。对于大数据平台的安全，需要大数据平台厂商、开源组织如 Apache 软件基金会、Cloudera、Hortonworks 等来推动。综合两方面的因素，全面、彻底地解决大数据安全问题，需要大数据企业、大数据技术公司、安全技术公司深入合作，重新设计和构建大数据安全架构和开放数据服务。具体应从网络安全、数据安全、灾难备份、安全风险管理、安全运营管理、安全事件管理、安全治理等多个角度考虑，部署整体的安全解决方案，保障大数据计算过程、数据形态、应用价值的安全。

4.3 大数据人才与教育

4.3.1 教育与科研机构

大数据的应用范围广泛，有很大发展空间，亟须相关方面的技能型专业人才。近年来，世界各国纷纷成立大数据科学研究机构，各大学也成立相关学院培养大数据人才。如波士顿大学和北卡罗莱纳州立大学均开展大数据这一方向的课程体系与专业建设，美国纽约大学、英国邓迪大学也从 2013 年起设立数据科学硕士学位，美国哥伦比亚大学将从 2015 年起设立博士学位。

在国内，工信部教育与考试中心也早在 2003 年就从国外引进了数据分析师培训和认证课程，并于 2008 年开始在国内大规模推广。国内一些经济发达城市的项目数据分析师事务所既承接大量的数据分析项目，也从事数据分析师的培训与认证工作。高校教育方面，香港中文大学自 2008 年起设立了“数据科学商业统计”科学硕士学位，清华大学新成立了数据科学研究院，自 2014 年 9 月起开始招收研究生；北京大学、中国科学院大学、西南大学等高校也先后设立了数据科学研究中心，其他院校也不断推出大数据相关的专业及培训课程。

纵观国内大数据人才培养方向，大致分为三种，即面向商学院、管理学院、财经学院的大数据分析方向，面向计算机学院与软件学院的大数据平台方向，面向理学院等的深度计算分析方向（IBM 大数据价值体系）。

1. 大数据分析方向

（1）清华大学

2014 年 4 月 26 日，清华大学与青岛市人民政府签署共建合作协议，成立“清华—

青岛数据科学研究院”，并召开大数据时代高端论坛。同时，清华大学在 2014 年 9 月起开始培养大数据研究生，每年培养约 150 名，首批硕士将从 2014 级硕士研究生中遴选，2015 年起公开招生。清华大学的大数据战略人才培养工程将包括大数据硕士项目、大数据博士项目、大数据职业培养课程等。

此外，清华大学开设的大数据职业培养课程，旨在推动全校研究生向大数据思维模式转变，共有 5 门选修课，已吸引部分同学参与学习。未来，学校还将以“大数据时代”为主题陆续推出系列论坛，主要包括管理论坛、信息论坛、战略论坛、校友论坛等，推动大数据人才培养和研究的发展。

（2）五所大学联合培养大数据人才

大数据分析硕士培养协同创新平台是在袁卫和纪宏教授的倡导下，由中国人民大学、北京大学、中国科学院大学、中央财经大学和首都经济贸易大学五所大学于 2014 年年初成立的。该平台依托应用统计专业硕士学位，联合了北京地区高校的师资力量与新华社、人民日报、中央电视台、全国手机媒体专业委员会、中国移动、中国联通、中国电信、SAS、JMP、阿里巴巴、华闻传媒产业创新研究院、华通人数据、龙信数据、西部云基地等业界大数据应用翘楚，利用院校协同创新、研究部门与业界部门协同创新的全新模式，以实际社会需求为导向共同开发课程并进行高级大数据分析人才的培养。大数据分析硕士培养协同创新平台的第一期实验班于 2014 年秋天开班。实验班由中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学等五校从 2014 年入学的应用统计专业硕士生中筛选 50 人组成，必修课在中国人民大学联合统一授课。

（3）中国科学院大学

2014 年 8 月 15 日，工信部软件与集成电路促进中心（CSIP) 联合中国科学院大学管理学院推出了“国家信息技术紧缺人才培养工程（NITE)”，该项目将进一步打造大数据信息交流的平台，促进大数据技术与工具在各领域的广泛应用，推动学校大数据研究及企业应用更好融合，提升企业的科学决策及管理水平，培养既具备专业能力又具有创新眼光的全能卓越高科技人才。

此外，2014 年，中国科学院大学开设“大数据技术与应用”专业方向，该专业将面向科研发展及产业实践，培养信息技术与行业需求结合的复合型的大数据人才。该专业首先瞄准的是具有一定行业背景的带着相关行业数据应用问题的在职人员，通过整合中国科学院及其合作伙伴中在大数据技术及应用领域发展前沿的企业资源，打造出独具中

国科学院特色和优势的专业方向。

为实现教育与产业的有效对接，中国科学院大学工信学院、SAP 中国及北京西普阳光公司在整合双方资源优势的基础上，在软件工程领域下开设了大数据专业方向。为此聘请了来自 NEC（中国）大数据研究院、中国移动大数据分析中心、北京云基地大数据实验室、数据堂、腾讯云数据分析中心等知名企业大数据专家共同参与课程设计、课程研发、教学实施、岗位推荐等工作。为学生提供了“实训 + 就业 + 研究生学习”的一体化职业发展通道。

该专业以数据架构、数据分析、数据应用为课程主体，旨在培养具有实战经验的高素质、实用型大数据人才。同时通过知名企业提供的硬件、软件及数据资源，让学生在真实的大数据环境中直接参与项目实践和企业内训，真正把握企业在大数据方向的数据分析、数据管理、系统开发与数据挖掘等方面的核心技能，成为未来大数据领域不可或缺的人才。

（4）大连理工大学

2014 年大连理工大学与 IBM 签约，联合成立“大连理工大学 -IBM 智慧城市与大数据处理联合实验室”。双方将以联合实验室为载体，依托 IBM 全球主机技术研究院及美国“沃森”实验室在企业计算、大数据、云计算、智慧城市等领域的先进技术和解决方案，以及大连理工大学在人才和科研方面的高校优势，主要从事智慧城市、大数据、企业计算相关领域的人才培养、联合科研、产品研发、技术转化等工作。该合作将为大连智慧城市的建设注入新的力量，同时为大连的科技发展培养满足市场需求的高价值人才。

（5）西安交通大学

2012 年，西安交通大学与 MIT 数据科学研究中心合作成立了国内第一个大数据和数据质量研究中心。2012 年，西安交通大学软件学院与 IBM 共建针对研究生的业务分析系。2013 年，学院获得了国家第一个自然科学基金“大数据”重点项目。2013 年 8 月 18 日，西安交通大学管理学院与陕西省信息化工程研究院联合成立了“大数据应用与管理研究中心”、“博士后创新基地”。双方希望将高校的科研成果与市场需求结合起来，理论与实践并重，推动两院的大数据相关领域科学研究向纵深发展。

2. 大数据平台方向

（1）北京大学

2014 年 7 月 26 日，北京大学正式成立“北京大学大数据技术研究院”，主要由北京大学深圳研究生院、北京大学软件与微电子学院、北京大学信息科学技术学院、北京大

学信息工程学院（深圳）发起，推出多学科交叉培养的大数据硕士项目，以应对国家对大数据的战略部署。2014 年 9 月，第一批大数据硕士学位研究生正式开始培养。

（2）武汉大学

2014 年 8 月 17 日，武汉大学计算机学院承办了“2014 大数据与未来计算论坛”，并在会上宣布“武汉大数据发展战略研究院”成立。大数据产业涉及大数据科学、大数据技术、大数据工程和大数据应用等领域，希望在武汉大学的引领作用下，推动大数据基础技术如大数据计算架构、大数据安全与隐私、大数据分析、大数据深度学习、大数据智能、大数据应用技术等共性技术方面的进步，为大数据产业培养和输送有竞争力的大量人才。

（3）北京交通大学

2011 年 9 月 3 日，北京交通大学软件学院与 IBM 公司合作共同发布了信息管理链和人才培养计划，该计划通过北京交通大学软件学院与 IBM 的合作，共同开发信息管理方向的系列培养课程，将业界最新的信息管理技术与软件产品，以及实践教学引入高等教育中，以培养“大数据”时代市场所需的精英型、应用型高端信息管理人才，推动信息管理产业及生态系统的健康、快速发展。

3. 深度计算分析方向

（1）北京航空航天大学

继 2012 年 9 月北京航空航天大学成立大数据科学与工程国际研究中心后，作为布局大数据战略方向的另一重要举措，北京航空航天大学计算机学院、北京航空航天大学软件学院、工信部 CSIP 移动云计算教育培训中心三大权威机构整合优势资源，联合创办了国内第一个“大数据科学与应用”软件工程硕士专业。全国领先的学科体系、国内外知名 IT 企业师资、移动云计算、互联网营销等前沿专业方向的成功建设经验及其创新人才的培养模式，确保了该项目的领先性、权威性、创新性及实战性。

2013 年，北京航空航天大学软件学院在全国范围内首开的大数据技术与应用专业，在技术上，侧重于对 HDFS、Hadoop、MapReduce、HBase、DBA、NoSQL 等专业技能进行培养，着重解决因技术盲点而带来的职场瓶颈。师资方面，则由百度首席云架构师林仕鼎担任大数据专业主任，由北京航空航天大学计算机学院院长吕卫峰，美国卡内基・梅隆大学博士、原 Oracle 公司（美国）高级系统架构师邓侃，软件工程领域国际知名学者蔡维德以及一大批在大数据行业享有盛誉的人担任专业导师，在高手如云的环境

中逐步培养大数据战略思维和工作方法论。软件方面，搭建了开发和处理大规模数据的Hadoop 系统平台，并以其高容错性、高传输率、高吞吐量的特性实现海量数据的计算与处理，进而为大数据的挖掘、存储、分析提供支撑。硬件方面，北京航空航天大学计算机学院、慧科教育、百度公司累计提供上千台服务器共同组建庞大的数据处理中心，为进行大数据领域的研究与实践搭建了共享的多层面 IT 平台。

（2）西南交通大学

西南交通大学公布了学校数字化战略的“七年计划”，将在人才培养、学科建设、学科研究、社会服务、信息化校园等方面推行改革。在 2015 年将形成 5 ~ 8 个大数据相关学科方向，规划招生 20 名“大数据新生”。学校将会投入不少于 1 亿元资金用于相关软硬件的建设。

此外，该校将整合相关资源与机构开展金融大数据领域人才培养和科学研究，计划从 2015 年起招收有计算机、概率统计专业背景的博士生，然后再开始招收硕士生。

西南交通大学大数据教育分为三阶段。

第一阶段：2014 年至 2015 年

- 完成 2 ~ 4 门 MOCC 课程的启动
- 大数据专业（方向）试点本科招生
- 推进省部共建和市校共建
- 形成 5 ~ 8 个稳定的大数据相关学科方向
- 立项建设 10 个左右大数据相关交叉学科培育项目

第二阶段：2016 年至 2017 年

- 初建“学生学习成长档案系统”
- 数字化保留所有学生在线数据
- 推进高铁大数据研究中心建设
- 形成 3 ~ 5 个具有国际影响力的大数据学科
- 建设“数字化”医院

第三阶段：2018 年至 2020 年

- 申报大数据国家级科研基地
- 建设 30 门 MOCC 课程
- 探索高层次综合性大数据人才培养模式

（3）西安交通大学

2012 年，西安交通大学软件学院与 IBM 共建针对研究生的业务分析系。首届业务

分析系学生将在 2015 年 6 月毕业，计划与亟需大数据分析人才的企业合作在 2014 年举办校园招聘会，建立端到端的应用型人才培养渠道。

4. 国际合作方面

（1）国际合作教育

为更好地推动大数据相关学科建设和人才培养，利用 IBM 全球资源与国际该领域产学研结合的领先大学接轨，加强国际化研究应用型师资建设，国家留学基金委员会与 IBM 合作开展大数据方向国际交流项目，助力国内高校培养具有全球视野和竞争力的国际化人才，服务国家教育、科技、经济和社会发展。2014 年 5 月启动的首期项目包括“师资质量提升计划”和“民生科研访问学者计划”，名额 20 人，为期 3 ~ 24 个月，选派学科领域包括信息科学、生命科学、公共健康、交通、电信、能源、环境、农业、材料、制造、经济管理等。计划访问学校包括耶鲁大学、加州大学伯克利分校和诺特丹大学等十多所美国一流大学。

（2）金融大数据教育

2008 年金融危机以来，人们逐渐认识到传统金融业的弊端，而利用大数据分析进行更可靠的金融决策，成为金融领域的研究热点之一。

2014 年 11 英国牛津大学成立首个金融大数据实验室“牛津—聂金融大数据实验室”，该实验室由香港金融数据技术有限公司创始人聂凡淇捐赠，旨在研究分析金融市场各方相关行为和数据，探索市场规律，将大数据技术更好应用于金融领域，推动金融领域的发展。此外香港金融数据技术有限公司还将在牛津大学、剑桥大学、帝国理工大学等英国顶尖学府率先推出“金融创新工场孵化计划”。这一计划包括推出一款模拟证券交易的应用程序，依托移动互联网并结合量化交易和大数据技术，让高校学子能在这一平台上学习锻炼，培养投资兴趣并提高金融“实战能力”。

2014 年 11 月西南交通大学成立了四川省首个金融大数据研究院，计划 2015 年招收首批研究生，研究院将邀请国内证券交易所操盘手、国外金融公司交易员任教，并组织大量实战演练提升研究生的综合能力。在科研方面，研究院还将定期举办金融与大数据研讨会，并将筹办全国首个金融大数据国际学术期刊。此外，为了加强大数据方面的人才培养，西南交通大学还将进行大数据专业试点本科招生，明年预计招收 20 人，学生将从主修金融专业，辅修数学、概率统计等专业的经济管理学院中招生。

（3）互联网社区培训

炼数成金培训的前身是 ITPUB 培训。自 2003 年起至今 10 年内为业界培养了大量人才，堪称数据行业的黄埔军校。该培训机构以数据分析、分布式系统、数据库、企业信息化等作为主要线索，开设了一系列有关课程。

4.3.2 课程体系

大数据人才培养和学科建设应对大数据价值体系进行一个完整梳理，纵观大数据时代对人才的综合需求，可发现，大数据平台教育与大数据分析教育将成为大数据人才培养的两个重要方向。

大数据平台方向，主要着重培养 Hadoop 系统、流计算、数据仓库和信息整合与治理等方向的人才。该方向人才能够打破传统数据库的观念，实现数据管理能力的创新，为企业提供实时大数据分析。总结各高校和机构大数据平台方向的课程体系，该方向的主要课程包括：数据挖掘与分析、信息技术分析与管理、信息安全、云计算架构设计、Hadoop 开发与管理、Oracle 数据库设计与建模、大数据商业管理、大数据营销、信息资源管理、信息化总体架构、数据库管理、企业架构等课程。

大数据分析方向，主要着重培养风险分析、内容分析、决策分析等方向的人才。该方向人才能够为决策者提供全面、统一且准确的信息，帮助他们在海量的数据中洞察商业机遇，发掘商业价值，制定更为有效的决策。总结各高校和机构大数据分析方向的课程体系，该方向的主要课程包括：数据挖掘与分析、数据查询与报告、商务智能、大数据分析与应用实践、精准营销、网络互动营销、客户关系管理、决策管理、商业数据分析、市场预测、供应链数据管理、信息安全与开放等。

4.4 国内外大数据政策与法规

各国已经将大数据上升为国家战略层面，重视数据挖掘和处理技术带来的价值。分析现有政策可知，以下几点已经成为各国及国际层面的共同关注点：1）制定策略促进数据存储、收集及分析能力以推动大数据的发展，挖掘潜在价值；2）仍然将隐私保护放在重要地位，促进经济运行的同时确保隐私保护政策的可行性；3）努力建设透明政府，加快推动政务数据资源开放，保证公众知情权，提高公共决策的预测性和响应性。

4.4.1　国内外数据共享的政策与法规

大数据环境下，信息系统与生俱来的脆弱性以及外部环境的不安全因素促使网络安全风险的不确定性加剧，并导致网络安全事件频发。在此背景下，及时获取和分析网络安全的威胁信息是实现网络安全保障的有效手段。关键基础设施信息、漏洞信息和政府开放数据共享已逐步成为国内外网络安全战略和相关政策的关注点，各国或组织政策法规均作出相应规定。

1. 美国

2007 年 10 月，美国政府发布《信息共享国家战略》，为共享信息提供指南，明确与需要者共享信息的必要性，强调不同部门间和政府与私营部门间信息共享的重要性。关键基础设施保护的信息共享合作项目包括：关键基础设施伙伴关系咨询委员会从事政府内部和公私合作，信息共享，并在关键基础设施保护的整个范围内参与活动；保护关键基础设施信息 / 敏感信息，主要侧重于分析和保护关键基础设施和保障系统，识别安全漏洞，并制定风险评估。

2009 年 11 月，美国国土安全部成立美国国家网络安全和通信综合中心（NCCIC），“常态”期间，NCCIC 的主要工作是汇总各合作部门的网络安全信息同时实现信息共享，并根据这些信息维护国家级的网络安全态势感知系统。NCCIC 可以处理涉密和不涉密等各级别的信息，并通过恰当的渠道与不同级别的合作伙伴进行信息共享。

2011 年 5 月，美国发布《网络空间国际战略》，指出当未来网络空间中的安全事件需要政府干预时，政府官员可以尽早检测威胁，实时共享数据，阻止恶意软件传播，最大限度降低破坏程度，同时也能维持信息不间断传输。当涉及国际犯罪调查时，执法机构能够进行跨国合作，维护和分享证据，并把罪犯绳之以法。此外，事故响应需要加强私营机构和国际社会双方的合作和技术信息共享，这需要各国一起承担责任。

2011 年 7 月，美国国防部通过《网络空间行动战略》，指出发展国际社会共享的网络空间态势感知和预警能力，有助于集体防卫和集体威慑。

2013 年 2 月，白宫新闻秘书办公室发布《改善关键基础设施的网络安全行政命令》，授权美国政府相关部门制定关键基础设施网络安全框架，明确应遵循的安全标准和实施指南，保护隐私权和公民自由权，为关键基础设施所有者和运营商进行安全检查提供依据，促进其与政府共享机密信息并参与政府制定和执行标准的决策中。通过该自愿性信息共享计划，政府将为关键基础设施公司或为其提供安全服务的商业服务提供商提供网

络威胁和技术相关机密信息。

美国总统事务办公厅发布的《大数据：抓住机遇，获取价值》报告在阐述执法与安全问题时，提出利用大数据分析与信息共享加强网络安全的行动建议，指出联邦政府与私人部门在网络安全领域关键基础设施保护方面对大数据利用项目、试点性项目及研究的合作伙伴关系可以帮助加强美国的网络防御能力，尤其是通过共享网络威胁数据的方式。政府部门继续在规定公司共享特定威胁信息及在此基础上捍卫其网络相关责任的同时支持隐私保护立法。政府部门将继续采取行政措施，增加对帮助公共与私人部门阻止和应对网络威胁的信息共享与分析种类的激励，并减少其发展障碍。

除了关键基础设施信息、漏洞信息共享之外，自 2009 年开始，美国政府还不断利用云计算、大数据等技术手段实现政府信息对社会的开放共享，从而推动开放政府的进程。

自 2009 年 1 月美国第 44 届总统奥巴马上任以来，美国国内掀起了一股“数据民主化”浪潮，奥巴马签署的第一份备忘录就是《透明和开放的政府》，指出“政府应该是透明的、具有高参与水平的、合作协调的”，并责成起草《开放政府令》。与此同时，美国联邦政府还推出了“国民开放政府进展报告”，指出“透明”是“通过向国民公开政府信息，尽到行政告知的责任”。为了实现透明，对《信息公开法》进行修改，对 Data.gov 等做出了详细计划。2009 年 5 月 21 日，由美国总务管理局主管的数据门户网站 Data.gov 上线，各政府机构被要求积极向 Data.gov 提供信息。为了便于二次开发，网站将国家基本数据（人口统计、犯罪统计等）、环境（大气污染数据、地质信息等）、保健（各地方肥胖统计等）、经济（就业状况、消费支出等）等 400 多种统计数据同时以 XML、CSV、KML/KMZ、XLS 等多种形式提供。2013 年 12 月，白宫公布第二份《开放政府国家行动计划》文件，承诺通过二度开展开放政府国家行动计划，为美国民众提供更加透明、参与度更高的政府机构，赋予民众参与当地政府财政预算、向政府监管提供建议等活动的权利。

2. 欧盟

2009 年，欧盟通过《保护欧洲免受大规模网络攻击和中断：预备、安全和恢复力的通讯》，提出五个支柱以应对关键基础设施面临的挑战，包括：准备和预防，建立公私信息共享机制；确保所有阶层的防范能力；确立欧洲对抗风险的公私合作；建立成员国之间关于信息共享的欧洲论坛等。欧盟委员会提出在欧盟网络和信息安全局的支持下，建立适当的早期预警机制与欧洲信息共享和预警系统。

2011 年，欧盟通过开放数据战略，明确提出实施路线图，包括：2012 年实施开放数据门户网站；2013 年实施一站式门户网站试点，其网站包含多国语言界面和服务，支持整个欧盟数据的搜索；2013 年所有成员国必须完成数据政策的制定和实施，建立公共数据门户网站，并在 2015 年实现与欧洲数据门户网站对接；2017 年，预计开放数据再利用、开展创新、提高公共部门服务效率，每年利润总计应达 1000 亿欧元[㊀]。

2013 年 7 月，欧盟议会、部长理事会、欧盟经济与社会委员会以及地域委员会联合通过《欧盟网络安全战略：构建一个开放、安全和有保障的网络空间》，强调各机构在网络安全中的责任，提倡多方合作，要求私营企业，尤其是为能源、运输、银行、股市、教育提供信息系统的私营企业，以及重要的互联网服务运营商，向各国的网络安全机构报告重大事故，并运用经济杠杆原理鼓励私营企业积极参与网络安全建设。战略提出：保护基本权利、言论自由、个人数据和隐私是网络安全原则，要求当关系到个人数据的利害时，为了网络安全而共享的任何信息均应遵守欧盟数据保护法律，并充分考虑该领域的个人权利。委员会请求欧盟议会和理事会明确网络信息安全方面的国家能力和准备、欧盟层次的合作、风险管理实践的采纳以及信息共享。委员会请求产业机构从确保强有力的、有效的资产和个人保护角度，主导网络安全方面的高水平投资，开展最佳实践以及与公共机构进行信息共享，特别是通过公私伙伴关系例如欧洲对抗风险的公私合作进行。成员国应当鼓励国内政府部门和私人部门之间的信息共享，以帮助成员国和私人部门保持对不同威胁的全景视图，并能够更好地了解用于更快捷开展网络攻击和反击的新技术。

3. 英国

2010 年 1 月，英国政府数据开放网站（Data.gov.uk）正式上线，旨在通过纳入大量政府数据的方式，使更多的人获得政府提供的数据。Data.gov.uk 已经发展到包含了 5600 多个来自各政府部门的数据集，涉及健康、交通、环保、社区、商务、教育等众多领域。2011 年 5 月，卡梅伦政府又提出了“数据权”的概念，并向全社会开放了英国政府 2005 年以来公共开支的全部原始数据。2011 年 4 月，英国劳工部、商业部宣布了旨在落实、推动全面数据权的新项目——“我的数据”。根据该项目，即使是商业机构收集的数据，只要记录的数据涉及个人信息，个人都有权查看和使用。

2011 年，英国通过《网络安全战略——保护与促进数字世界中的英国》，将打击网

㊀ 《大数据创新》，曹凌，欧盟开放数据战略研究，情报理论与实践，2013（4）。

络犯罪，力争把英国建成全球经商最安全的国家之一作为目标，为此提出的行动建议包括：从 2012 年 1 月开始，充分发挥更广泛的私人部门关于网络安全联合工作的积极性，保证执法在信息共享和降低网络犯罪风险方面充分与商业活动相结合；通过旨在减少影响的建议，确保实行全新的国家网络突发事件响应程序和拓展政府与私人部门之间的合作关系，提供便捷的有关商业威胁的信息共享。

2013 年 3 月，英国成立“网络安全信息共享联盟”，旨在促进公私部门间的网络安全威胁信息共享，确保所有参与者能够及时获取有关网络攻击的实时警报、技术细节、策划攻击的手段及应对措施等信息。

4. 中国

2007 年 4 月，国务院颁布《中华人民共和国政府信息公开条例》，指出行政机关应当主动公开下列政府信息：涉及公民、法人或者其他组织切身利益的；需要社会公众广泛知晓或者参与的；反映本行政机关机构设置、职能、办事程序等情况的；其他依照法律、法规和国家有关规定应当主动公开的。此外，《条例》还详细规定了信息公开的方式、程序、监督和保障措施。

2011 年 4 月，国务院办公厅发布《关于进一步加强政府网站管理工作的通知》，强调要健全机制，充分发挥政府网站的信息公开、互动交流作用。及时准确地在政府网站发布涉及群众切身利益、需要社会公众广泛知晓或者参与的政府信息，尤其要做好财政预决算、公共资源配置、重大建设项目、社会公益事业等领域政府信息的发布工作。凡是可公开的不涉密文件，都要通过政府网站公开发布。涉及群众切身利益的重要决策，要在政府网站公开征求意见；重要政策出台后，要及时通过政府网站做好政策解读工作。

2011 年 8 月，中共中央办公厅及国务院办公厅发布《关于深化政务公开加强政务服务的意见》，进一步重申深化政务公开的重要性，明确公开的总体要求，指出各级政府政务公开的重点内容。要求政府按照公开为原则、不公开为例外的要求，及时、准确、全面公开群众普遍关心、涉及群众切身利益的政府信息。

2012 年 7 月，国务院发布《关于大力推进信息化发展和切实保障信息安全的若干意见》，指出加强地理、人口、法人、统计等基础信息资源的保护和管理，保障信息系统互联互通和部门间信息资源共享安全。

2014 年 5 月，上海率先实行政府数据资源向社会开放，28 个市级政府部门的 190 项数据内容成为重点开放对象。以国内首个政府数据服务网（www.datashanghai.gov.cn）作为

开放统一入口，提供数据查询、浏览、下载等功能。此前，上海已启动政府数据资源向社会开放试点，建成“上海政府数据服务网（一期）”，9 家试点单位开放的数据产品及应用，涵盖地理位置、道路交通、公共服务、经济统计、资格资质、行政管理等 6 大领域。

纵观各国制定的威胁信息共享和数据开放政策法规，虽然可共享数据的对象、方式、途径、时间长短有所差异，但都随着大数据的发展推陈出新。大数据热潮来临之前，国外就意识到信息共享的重要性，在立法中确定互联网服务提供商与政府部门共享关键基础设施、安全漏洞等威胁信息，以加强合作共同解决安全问题。大数据炙手可热的今天，人们对政府透明化的要求促使国外将更多注意转向政府数据共享，在满足了解关键行业和用户自身数据的同时推动商业发展。相比威胁信息共享，目前我国的法律侧重于开放政府数据。政务数据共享有利于保障公民的知情权、建议权和监督权，预防腐败并实现民主制度，也可以促进数据流通，节省政府为满足不同群体重复查找信息的时间。然而，威胁信息共享法律的缺乏不利于突发安全事件的解决，我国还需完善相关立法提高危机处理能力。

4.4.2　国内外数据跨境的政策与法规

大数据环境下，全球政治、经济、文化联系迅速加强，数据的泛在跨境流动或者传输成为常态。作为人权重要组成部分的个人数据在跨境数据流中的比重越来越大，权利也更易受到侵害。在此过程中产生的安全风险成为各国尤其是国际组织的立法焦点，各国个人信息保护立法对信息跨境流动规则作出更多回应。

1. 欧盟

欧盟早在 1995 年就制定了《数据保护指令》（简称《95 指令》），旨在促进成员国之间的信息自由流动，同时确保个人在各成员国之间享受同样的基本权利和自由。根据《95 指令》，欧盟规制数据跨境流动分为欧盟境内成员国之间、欧盟与其境外国家两个层次，数据保护水平的充分性与否是欧盟决定跨境信息流动的基本原则。对内，《95 指令》为成员国确定了数据保护的标准，禁止成员国借数据保护的名义限制个人信息在欧盟境内的自由流动。欧盟成员国都应履行数据保护要求，成员国具备同等数据保护水平是数据在成员国间自由流动的基础，欧盟国家之间数据传输也要建立在各国法律许可的基础上。对外，《95 指令》规定向欧盟境外传输数据将受到限制，除非满足以下三个条件之一：一是第三国经欧盟认定具备所要求的充分保护水平；二是属于数据主体明示同意等例外规定；三是未认定具备充分保护水平的国家的数据控制者采取了所要求的充分保障措施。

为了应对数字时代数据跨境流动的新挑战，确保欧盟规则的前瞻性，欧盟委员会重新审视现有的个人数据保护法律框架，于 2012 年 11 月制定了《一般数据保护条例》草案（简称 GDPR）。GDPR 扩大了欧盟数据保护立法的地域范围，如果非欧盟国家的数据控制者在欧盟向公民提供货物或者服务时处理欧盟公民的个人信息或者进行行为监测，也会成为 GDPR 规制的对象。

2. 经济合作发展组织（OECD）

OECD 在 1974 年成立了一个关于跨国传送个人信息和隐私权保护的专家组，研究个人信息跨国传送中的数据保护问题，此后分别于 1977 年及 1980 年举行了相关议题的研讨会。1980 年 9 月，OECD 理事会提出《保护个人信息跨国传送及隐私权指导纲领》，对个人信息保护做了原则性规定。这个指导纲领分为“国内适用的基本原则”及“国际间适用的基本原则”。国内适用的基本原则包括：搜集限制原则；资料内容正确性原则；目的明确化原则；利用限制的原则；安全保护原则；公开原则；个人参加原则；责任原则。国际间适用的基本原则是自由流通和合法限制原则，主要内容包括：OECD 成员国对于个人资料在国内的处理和传送，应考虑对其他成员国的影响；成员国不能只是阻碍个人资料在国际间的流通，为确保其安全，应寻求合理及适当的方法解决个人资料的安全问题；成员国本国与其他成员国之间，虽应取消限制个人资料在国际间流通的规定，但对其他成员国并未实质地遵守本规则或者前述资料所移送的国家并无隐私权保护规定或者前成员国并无此类保护规则的，则关于某特定的个人资料可以限制其流通；成员国应避免以隐私权及个人自由保护为名，超过保护的必要程度而制定对个人资料国际流通产生危害的法律和政策。

3. 其他国外立法

美国对数据保护采取部门保护方法，以立法、管制和自律相结合，更多地强调行业自律。由于美国未被欧盟认定充分保护的资格，为了实现欧盟向美国的跨境数据传输，经过多次磋商，2000 年 3 月，美国与欧盟达成了安全港协议。欧盟在 2000/520/EC 决定中指出，安全港的法律基础是欧盟委员会根据《95 指令》第 25 条第 6 项作出的充分性认定。该制度仅适用于从欧盟接受个人信息的美国机构，以使其具备安全港资格，并获数据保护充分性的认定。安全港协议的数据处理要求包括：通知、选择、连续传输、安全、数据一致性、获取、执行等。这些原则体现了信息主体的知情权、选择权、获取权、异议权、救济权等，给加入安全港的公司设置了公开透明、目的限定、特殊敏感信

息处理、数据质量、安全措施、异议处理等责任和义务。安全港制度要求对违反义务的公司应进行充分严厉的制裁。美国商业机构每年需向商务部递交一份遵守安全港制度的承诺书，报告他们遵循相关规定的情况，而且这些承诺必须公开披露。

2013 年新加坡正式实施的《个人资料保护法》明确规定机构不得将个人数据转移至境外，除非依据本法的相关要求，确保机构能够提供符合本法要求的个人数据保护水平。但个人资料保护委员会可以通过发布书面通知，豁免相关机构遵守前述义务。

澳大利亚新修订的《隐私保护法》对数据跨境流动问题给予了更多关注，要求数据的海外接收者应受到法律或相关计划的约束等。

2014 年 7 月 4 日，俄罗斯国家杜马批准了一项最新法律规定，要求所有收集俄罗斯公民信息的互联网公司都应当将这些数据存储在俄罗斯国内，其生效时间为 2016 年 9 月 1 日。

4. 中国

我国目前关于数据跨境流动的法律规制较少，主要在国务院发布的《关于大力推进信息化发展和切实保障信息安全的若干意见》中有所涉及。该意见明确指出，严格政府信息技术服务外包的安全管理，为政府机关提供服务的数据中心、云计算服务平台等要设在境内。

2014 年 11 月 3 日，第十二届全国人大常委会第十一次会议初次审议了《中华人民共和国反恐怖主义法（草案）》，第十五条第三款规定：在中华人民共和国境内提供电信业务、互联网服务的，应当将相关设备、境内用户数据留存在中华人民共和国境内。拒不留存的，不得在中华人民共和国境内提供服务。

国际组织在全球经济贸易和技术发展中扮演着重要角色，不断为寻求全球经济发展和各国利益之间的平衡进行努力，完善的数据跨境政策成为首选的平衡点。国际组织制定的政策不仅在本区域内有效，也会对其他国家产生很大影响，尤其是具有前瞻性的欧盟数据跨境措施，经常成为各国借鉴的蓝本。虽然世界上现有的数据跨境政策法律仅限于个人数据，但随着数据收集和萃取技术的进步，企业数据、关键基础设施数据等必将列入跨境数据的立法范畴。我国政策法律目前没有涉及数据跨境传输的相关规定，但政府已经开始站在国家安全的角度进行探索，对为政府机关存储数据的各大互联网服务提供商提出了要求。

4.4.3　国内外隐私保护的政策与法规

中国计算机学会大数据专家委员投票评选出大数据的 10 个问题，其一就是大数据

下的数据安全与隐私保护。为了维护个人数据安全，保护个人身和财产利益，信息化发达国家纷纷制定新法律或者调整现有法律强化隐私保护。

1. 美国

美国 1974 年颁布的《隐私权法案》是迄今为止美国最为全面和重要的隐私保护立法，规范政府如何收集、处理个人数据等行为，在保护个人隐私方面有着无可替代的地位。该法案对个人信息主体的权利、政府机关的义务以及救济途径都做了详细规定。根据该法案，个人信息主体的权利包括三个方面：第一，公开权。信息主体有权利决定是否公开个人信息，对于侵害公开权的行为有起诉并获得赔偿的权利。第二，访问权。个人将其信息提供给相关行政机关是一项法定的公民义务，在符合申请获取自身信息的条件下，信息提供者有权免费获取其个人信息的复印本，并有权知晓其信息在行政机关的状态。第三，修改权。个人信息是否准确、完整，牵扯到个人对外“信息人格”是否准确和完整，关系到个人尊严和荣誉，因此该法赋予信息主体有更正不准确、不新颖以及不完整的信息记录。此外，该法案还明确了政府等公务机关的义务：第一，信息搜集必须与职务需求相关、与国家利益相关并经严格审批手续才允许搜集个人宗教信仰、政治信仰等“敏感”信息。第二，为增加政府搜集信息的透明度，加强社会监督，纠正不当行为，个人信息的搜集必须进行公示公告。第三，政府机关应采取必要措施加强自身管理，确保信息安全、准确、完整以及新颖。任何机关违反上述规定，导致个人利益遭受损失或遭遇不公平待遇的，都有权向当地法院起诉请求民事赔偿。但该法只适用于联邦部门以上机构，约束范围有限，并且没有设立独立的监督机构，使得个人信息保护的条款在实施中问题不断。

美国没有综合性法典对个人信息的隐私权提供保护，主要依靠联邦和州政府制定的各种类型的隐私和安全条例。1986 年颁布的《电子通信隐私法》是目前有关保护网络个人数据最全面的立法，涵盖了声音通信、文本和数字化形象传输等所有形式的数字化通信。该法不仅禁止政府部门未经授权的窃听，也禁止所有个人和企业对通信内容的窃听，同时还禁止对存储于电脑系统中的通信信息未经授权的访问及对传输中的信息未经授权的拦截。然而，随着数字时代的发展，部分企业认为该法案已经不能合理地反映用户对隐私方面的期待。自 2010 年以来，谷歌就一直致力于推动该法案的改革。2014 年 5 月，美国总统事务办公厅发布的《大数据：抓住机遇，获取价值》报告也提出了修正《电子通信隐私法》的行动建议。2014 年 6 月，美国国会表示正在考虑该法的修改草案。修

改草案的重点是防止在无警示的情况下，对数据进行收集的行为，修改该法的决定已在众议员中赢得了多数支持。

1996 年美国通过《医治保险携带和责任法》（简称 HIPAA），通过建立电子传输健康信息的标准和要求鼓励健康信息系统的发展。保障个人健康隐私的完整性和机密性；防止任何来自可预见的威胁、未经授权的使用和泄露；确保官员及其职员遵守这些安全措施。2001 年通过 HIPAA 修正案，目标之一就是保护病人的电子健康记录，并提出保护的具体标准。该修正案详细规定了行政保障措施、物理保障措施、技术保障措施及安全责任的分配问题，对于违反安全标准的实体，规定了最高可达 25 万美元罚款和最长 10 年监禁的严厉惩罚措施。

1998 年美国通过了《儿童网上隐私保护法》，适用于美国管辖之下的自然人或单位对 13 岁以下儿童在线个人信息的收集。该法详细规定了网站经营者必须披露其隐私保护政策，寻求监护人同意的时间及方式，经营者违反儿童隐私保护应承担的责任，禁止营销 13 岁以下儿童的个人信息。但如果经过父母同意，13 岁以下儿童可以合法提供其个人信息。为了顺应大数据的时代潮流，更好地保护儿童的隐私数据，美国联邦贸易委员会修订了该法，自 2013 年 7 月 1 日起生效。新出台的规则把个人信息范畴扩展到地理位置标记、IP 地址、个人照片、音频、网站及 cookie。同时，新规则对一些使用插件或者广告获取信息的公司也同样有效。此外，考虑到 2000 年以来在线技术的发展变化，规则修订了运营商、个人信息、针对儿童的网站或网上服务等术语的定义，并增加了征得父母同意、对相关主体的通知、保密性和安全性以及安全港条款等要求，并引入了数据留存、删除等新术语。

为了配合执法需要，检查罪犯手机已经成为获取证据的优先选择，因此，各国现有数据保护立法的保护主体一般不包括犯罪人员。然而，近几年对人权的呼声愈发高涨，美国开始考虑将罪犯隐私列入立法保护范畴。2014 年 3 月，美国得克萨斯州刑事上诉法院通过一项规定：犯罪分子在入狱之后执法人员无权搜查其手机。新规定以 7 票赞同 1 票反对通过，意味着手机的所有者在入狱之后依然能够确保手机内容不被监听，在美国联邦宪法第四修正案的保护下保有自己的隐私。2014 年 6 月，美国最高法院裁定，在没有得到授权的前提下，警方不得私自查看被捕人员的智能手机。最高法院指出，用户对手机中图片、视频、电子邮件、短信和联系人信息的隐私权高于执法部门的需求，手机中的数字信息本身不可能作为武器来伤害办案警员，也无助于嫌犯逃跑。但是，允许警员查看被捕人员手机的外观获取线索。

“棱镜门”事件爆发后，个人信息的搜集行为引起世界范围内的空前关注，无论是政府出于国家利益的搜集行为，还是类似谷歌、微软、思科等大型跨国企业的搜集行为都令人无法容忍。此外，越来越多的大型科技公司对其向美国政府提交用户数据的方式提出了质疑。为了遏制大规模的数据收集行为，并且保护这些科技公司免受相关起诉，美国白宫方面要求立法部门出台新的改革法案。2014 年 5 月，美国白宫计划递交新法案，据了解，新法案将不再授予国家安全局（NSA）可大规模获取民众数据的权力，取而代之的是特定的必要数据。新法案将改变 NSA 监控与信息收集的方式，禁止 NSA 收集手机通信元数据，其中包括通信号码、通话时长和通话次数，但不包括通话内容。法案还要求美国外国情报监视法院规定，情报机构的每个请求都必须与正在进行的调查相关，并基于“合理准确的怀疑”，这也将提高外国情报监视法院的透明度。2014 年 11 月 18 日，虽然美国参议院以 58 比 42 的投票比率否决了《美国自由法案》，但这类监控项目仍然需要借助新的立法才能继续实施。2015 年 6 月，作为监控项目法律基础的美国《爱国者法案》将会过期，因此 NSA 必须借助新的立法才能继续获取通信数据。

美国的隐私保护立法具有三方面的特征：第一，起步较早。美国在立法上对个人数据的保护具有前瞻性，远远领先于世界其他各国，早在 1974 年的《隐私权法案》中就强调了数据开放性，不仅增加用户对本国互联网服务提供商的信任，还会让用户最大程度了解自己的数据，增加商业模式和就业机会，从而促进经济的发展；第二，高度重视敏感领域和弱势群体的个人数据保护。美国对公民的电子通信、医疗、金融以及儿童的数据等都通过了专门的联邦立法，并且随着数字化的发展不断对其调整和修订，增加新的概念、惩罚措施等规定；第三，重视与时俱进。大数据时代带来的海量信息收集、存储和处理，有利于国际贸易和电子商务的发展，但同时增加了个人数据泄露的安全风险。尤其是“棱镜”事件之后，为了挽回民众对政府的信任，美国站在最有利于商业发展和数据权利保护的高度，通过新法案限制 NSA 的数据收集。此外，美国州立法已经开始关注智能手机等新的风险域。

2. 欧盟

欧盟拥有较为完善的个人数据保护框架，包括一系列的指令、协议、条例等，其中最重要的是《95 指令》、2002 年《隐私和电子通信指令》。《95 指令》是欧盟个人数据保护史上的里程碑，也是目前最重要的数据保护法，旨在推动数据在欧盟范围内的自由流动，在全欧范围内建立数据保护的统一标准以实现完备和充分的数据保护。2002 年《隐

私与电子通信指令》解决了电子通信部门、无线电通信、传真、电子邮件互联网及其他类似服务中的个人数据保护问题，适用于“欧盟范围内在公共无线电通信网络中提供电子通信服务的企业”中处理个人数据的行为。

尽管《95 指令》为成员国个人数据保护立法设立了最低标准，但成员国之间具体实施方式和执法的不同导致个人数据保护水平的差异，严重影响了数据保护的实际效果和欧盟内数据的自由流动。此外，大数据、云计算、移动互联网、智慧城市等技术的快速发展给个人数据保护带来了新挑战。为了应对全新数据世界带来的安全风险，欧盟委员会出台 GDPR，提出个人数据保护立法的一揽子改革计划。2013 年 6 月，“棱镜门”事件曝光后，原本富有争议的欧盟个人数据保护立法改革进程大为加快。2013 年 10 月 21 日，欧洲议会公民自由委员会以绝对多数票通过 GDPR。GDPR 引入了更为严格的数据保护制度，规定：对于违反数据保护法律规定的行为，处罚将提升到公司全球年度收入的 2%；超过 250 人的企业应设立“数据保护负责人”；明确引入“被遗忘和删除权”，即当用户提出需求时，企业必须删除其保存的个人数据。这些更为严格的制度对跨国互联网企业带来重大影响。2014 年 5 月，欧盟法院裁决，搜索引擎必须遵守隐私，移除公民请求删除的相关链接。Google 自 2014 年 6 月 26 日起开始遵循欧洲法院判决结果，正式删除“有权被遗忘”或者过期的搜索结果。

2014 年 4 月，欧盟最高法院作出裁决，明确 2006 年制定的欧洲电信公司保留公民通信数据最多两年的《数据存留指令》无效。根据该指令，无线电通信企业需要保留各种数据，包括呼入和呼出的电话号码、通话时长、IP 地址、网络登入登出的时间、电子邮件活动等。各国必须采取措施，保障留存数据仅用于司法机关以及其他职能机构依法同意的国家机关使用。规定数据存留的最短期限为 6 个月，最长不多于 24 个月。期限届满后，数据必须删除。网络电子通信服务商和网络运营商留存数据的，应该保证数据以及相关信息必要时可以提交给职能部门。数据存留产生的费用由电信企业承担或者给予补偿。然而，奥地利和爱尔兰的法院要求欧洲法院裁决这条法律是否与欧盟的基本权利宪章相一致。最终，欧盟最高法院裁决《数据存留指令》严重干扰了尊重隐私和个人数据保护的基本权利，让公民感觉到他们的私人生活遭到了持续监视。

作为个人信息保护立法的模范和先驱，欧盟最先关注信息通信技术对社会的影响问题，于 1995 年制定了价值倾向明显、覆盖范围广泛且执行机制健全的《数据保护指令》。随着数字时代的变化，个人数据的商业价值一路攀升，欧盟于 2012 年适时提出“被遗忘和删除权”，成为各国数据保护立法的借鉴。当公民数据权利与执法相冲突时，欧盟选择

切实保护公民数据权利并为此废除了方便政府部门执法的《数据存留指令》，这是欧盟隐私保护立法区别于他国的重要特点。

3. 中国

大数据时代的到来，促使我国互联网行业快速发展，对促进经济繁荣起到了积极作用。然而，伴随而来的是用户个人信息的泄露风险和保护难度不断增大。近几年我国已经认识到个人数据泄漏问题的严重性，相关立法节奏明显加快。

2012 年 12 月 28 日，全国人大常委会第三十次会议通过了《关于加强网络信息保护的决定》，明确规定国家保护能够识别公民个人身份和涉及公民个人隐私的电子信息。该决定填补了中国长久以来在个人信息保护方面的立法缺位。该决定对个人信息的保护包括个人信息的收集和使用，规定网络服务提供者和其他企事业单位在业务活动中收集、使用公民个人电子信息，应当经过被收集者的同意，遵循合法、正当、必要原则，明示收集、使用信息的目的、方式和范围。对在业务活动中收集的公民个人电子信息必须严格保密，不得泄露、篡改、毁损，不得出售或者非法向他人提供。同时应当采取技术措施和其他必要措施，确保信息安全，防止在业务活动中收集的公民个人电子信息泄露、毁损、丢失。在发生或者可能发生信息泄露、毁损、丢失的情况时，应当立即采取补救措施。公民发现泄露个人身份、散布个人隐私等侵害其合法权益的网络信息，或者受到商业性电子信息侵扰的，有权要求网络服务提供者删除有关信息或者采取其他必要措施予以制止。

2013 年 7 月 19 日，工信部发布《电信和互联网用户个人信息保护规定》。根据该规定的内容，电信业务经营者、互联网信息服务提供者应当制定用户个人信息收集、使用规则，并在其经营或者服务场所、网站等予以公布。该规定还要求，未经用户同意，电信业务经营者、互联网信息服务提供者不得收集、使用用户的个人信息。此外，电信业务经营者、互联网信息服务提供者在用户终止使用电信服务或者互联网信息服务后，应当停止对用户个人信息的收集和使用，并为用户提供注销号码或账号的服务。违反相关规定的，将由电信管理机构依据职权责令其限期改正，予以警告，可以并处一万元以上三万元以下的罚款，并向社会公告。

2014 年 8 月 8 日，国信办发布了《即时通信工具公众信息服务发展管理暂行规定》，从行业资质、隐私保护、实名注册、备案审核、内容限制等方面对即时通信平台及用户行为进行规范，并明确了对违规行为的处罚措施。

2014 年 10 月 9 日，最高人民法院通过《关于审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定》（下称《规定》），将利用信息网络侵害他人隐私权的行为定性为人身权益民事纠纷案件。根据该《规定》，基因信息、病历资料、健康检查资料、犯罪记录、家庭地址等，都属于比较敏感的个人信息，这些信息一旦向社会公开，不仅会造成个人难以弥补的损害，而且很多情形下会造成整个社会的不安。网络用户或者网络服务提供者利用网络公开这些信息造成他人损害，被侵权人请求其承担侵权责任的，人民法院应予支持。

2014 年 11 月 3 日，第十二届全国人大常委会第十一次会议初次审议了《中华人民共和国刑法修正案（九）（草案）》。将刑法第二百五十三条之一修改为：“违反国家规定，将在履行职责或者提供服务过程中获得的公民个人信息，出售或者提供给他人，情节严重的，处三年以下有期徒刑或者拘役，并处或者单处罚金；窃取或者以其他方法非法获取公民个人信息，情节严重的，依照前款的规定处罚；未经公民本人同意，向他人出售或者非法提供其个人信息，情节严重的，处二年以下有期徒刑或者拘役，并处或者单处罚金；单位犯前三款罪的，对单位判处罚金，并对其直接负责的主管人员和其他直接责任人员，依照各该款的规定处罚。”此外，在《刑法》第二百八十六条后增加一条，作为第二百八十六条之一：“网络服务提供者不履行法律、行政法规规定的信息网络安全管理义务，经监管部门通知采取改正措施而拒绝执行，有下列情形之一的，处三年以下有期徒刑、拘役或者管制，并处或者单处罚金：（二）致使用户信息泄露，造成严重后果的；单位犯前款罪的，对单位判处罚金，并对其直接负责的主管人员和其他直接责任人员，依照前款的规定处罚。”

总体上来说，目前我国没有专门的隐私保护法律，可适用的法律法规主要有《宪法》《关于加强网络信息保护的决定》《刑法》《治安管理处罚法》《未成年人保护法》《居民身份证法》《商业银行法》《邮政法》《档案法》《电信和互联网用户个人信息保护规定》等，这些法律对用户的个人隐私和通信自由作了零散规定。不可否认，我国已经开始高度重视隐私保护，自 2012 年以来连续三年通过了隐私保护的相关法规，尤其是 2014 年 10 月颁布的《关于审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定》涉及对敏感数据的保护，是立法的飞跃性突破。但我国仍然存在效力层级低、法律法规协调性弱、保护内容片面等立法不足，有待于相关部门更深层次地研究和改善。

第 5 章　大数据发展趋势与建议

前面几章从代表性应用领域、技术体系以及产业链与生态环境等角度阐述了大数据应用与产业现状。本章从更宏观的角度，梳理大数据的学科发展现状，总结大数据的学科与技术发展趋势，进而提出我国大数据基础研究与产业化的发展战略和建议。

5.1　大数据学科发展现状与趋势

5.1.1　大数据学科发展现状

1. 大数据研究还处于积累数据、分析现象为主的前科学阶段

不少学者认为，目前的“大数据”主要表现为“研究对象”，是一种需要探索的“现象”。随着采集数据成本大幅度降低，各行各业都涌现出大量非结构化的数据，当前正在探索存储、处理、分析大数据的新方法，尚未形成反映大数据共性规律的科学理论。观察现象，积累科学数据，从现象中发现规律，是形成物理、化学等科学理论走过的路。牛顿力学就是建立在大量天文学观察的基础上。研究人类社会活动规律的社会科学、以复杂网络为研究对象的网络科学等还处在牛顿力学诞生前的积累数据、分析现象阶段。

现有的大数据理论与模型高度依赖于其他学科，如统计分析、机器学习、分布式系统等，还没有建立起独立于其他学科的理论体系与研究方法论。但大数据基础研究可能不是传统科学的复制和延续，大数据有别于传统数据处理的本质是数据之间的相互关联，相互关联的数据跨越了物理空间、信息空间和人类社会，形成了三元空间交织融合的“数据界”(Data Nature)。数据界的存在仅仅是一个现象还是现象之下隐藏着一套全新的“数据科学”理论与“数据哲学”理论，目前尚不清晰。

大数据研究将促使科研第四范式逐渐形成，但第四范式的建立也需要一个过程，要求发展与已有的三种范式不同的科研方法。科研范式的改变和大数据共性规律的发现可能会交织在一起。估计还需要一段时间的努力大数据才能形成独立的学科。

2. 大数据的科学研究与产业应用脱节

当前经济形势下，纯粹依靠物质资源发展经济的老路已难以为继，而数据是贯彻国

家“创新驱动发展”战略的最重要资源。过去几年来，以“ BAT”为代表的大型互联网企业已具有与国际大公司竞争的经济实力和技术基础，他们依托自身拥有的巨量数据和现实的应用需求，已经发展出一些初步满足各自底层次需求的大数据解决方案，但在新技术引领未来的竞争优势方面存在诸多不足。我国在部署大数据科技创新布局时，要抓住当前难得的机会与条件，继续将大数据研究重点放到“网络大数据”方向，真正实现科学研究促进产业的跨越式发展。

另一方面，大数据研究在推进农业生产、工业制造和科学研究等方面尚未出现大规模聚集效应，直接从数据中产生知识的方法论尚未形成体系。所谓“第四范式”还有待研究界从基础问题体系和方法论层面进行提炼和挖掘。由于缺乏真正的大数据，不了解大数据应用的真实需求，科技界对大数据应用发挥的作用还不明显。在大数据基础研究中，科学研究和应用的脱节还表现在信息领域的科技人员与应用领域的科技人员开展深度合作十分困难，而没有这两类科技人员的深度合作，大数据基础研究很难取得突破性进展。

因此，国家在部署大数据基础研究时，一定要特别强调和重视信息领域和其他应用领域科研人员的密切合作，从制度上为跨领域的合作创造条件。同时要加大跨学科人才培养力度，安排充足的经费用于跨学科人才培养。

3. 大数据基础研究的问题体系尚不清晰

从 2012 年以来，科技部、国家自然科学基金委等部门通过“973”计划、重点课题资助计划等陆续支持了若干大数据基础研究类项目，聚集了一批来自于国内高校、科研院所以及企业前瞻研究部门的优秀人才与团队，开展了与大数据处理相关的基础研究。总体上来看，已有科研项目团队对大数据科学问题的定义，大数据研究的角度、粒度、深度等方面存在着较大差异，有些问题在概念层面就非常模糊。研究界尚未形成一个相对清晰的大数据基础研究问题体系。为提高科研效率，促进科学交流，形成真实的创新成果，在一定程度上需要顶层规划和科学引导。

5.1.2　大数据学科发展趋势

尽管针对大数据的科学研究工作还存在上面阐述的各种问题，但以从数据中提取信息和知识进而辅助决策为目标的数据科学逐渐得到认可和关注。在 CCF 大数据专家委员会于 2012 年 12 月发布的《大数据热点问题与 2013 年发展趋势分析》报告和 2013 年 12 月发布的《2014 年大数据发展趋势预测》报告中都预测数据科学将作为一门新的交叉学

科逐步兴起。甚至类似波色子的发现，数学、生物、物理、化学、材料等领域将在一定程度上依赖数据科学才能取得突破性进展。但上述报告同时还指出，数据科学作为一项新的科学，还有很多根本问题没有解决，甚至很多问题还没有被提出。所以，数据科学真正的兴起并成为一个支柱学科，还需要学术界更多的努力。作为对上述预测的一个印证，我们注意到，国家自然科学基金委员会在 2014 年组织的未来五年的“十三五”规划中，特别尝试设立了“数据与计算科学”这一专门面向大数据的学科方向，还具体定义该方向是研究数据的感知、收集、传输、管理、分析与应用的交叉性学科，旨在揭示数据的内在规律，探索数据计算理论，实现从数据到知识的转化，为大数据的科学计算以及在重要应用领域的预测、决策与应用提供基础。该项规划还指出，数据与计算科学主要包括两大内涵：一方面是数据内在规律，主要研究人—机—物三元数据空间的内在规律、大数据关联与演变机理等；另一方面是数据计算理论，研究大数据计算的基础理论、计算模式与新型体系结构等。

与大数据技术与应用走在了大数据研究前面的情形类似，尽管数据科学作为一门学科尚未完全建立，但世界各地的科研院所与培训机构都在积极探索大数据人才培养的课程与学位体系。许多大学（如美国的加州大学伯克利分校、哥伦比亚大学和纽约大学；英国的伦敦大学院、帝国理工大学；荷兰的埃因霍温技术大学；我国的清华大学、中国人民大学、北京航空航天大学、香港中文大学等）都设立了大数据研究中心或研究所。许多大学和研究所已经设立了面向本科生和研究生课程或学位来培养大数据专业人才，包括数据科学家和数据工程师。大数据作为横跨信息科学、数学、社会科学、网络科学、系统科学、心理学、经济学等多个学科的方向，运用到来自许多不同领域的理论、方法与技术，诸如信号处理、概率模型、机器学习、统计学习、计算机编程、数据工程、模式识别、可视化、不确定性推理、数据仓库与高性能计算等。因此，面向大数据的学科体系也将在很大程度上以其他学科的理论与方法为其基础。

5.2 大数据热点问题与技术发展趋势

5.2.1 大数据热点问题

数据科学作为一门新兴的学科，目前尚未建立起完整的基础理论体系，数据学科基础问题体系本身就是大数据领域的研究热点。大数据作为一门以数据及数据处理技术为研究对象的科学，更侧重于具体应用，与常规的信息处理体系框架类似，也存在着功能、性能、易用性、输入输出、系统安全等方面的问题。

1. 大数据科学问题

“科学”的定义是“反映自然、社会、思维等的客观规律的分科的知识体系”，大数据作为一门新兴的科学，其学科基础问题体系尚不明朗，数据科学自身的知识体系还不完备，还有待于学科理论基础的进一步突破。大数据所带来的数据复杂性、计算复杂性和系统复杂性的挑战，将会引领学科基础知识体系的逐步完善和发展。

2. 大数据分析的性能问题

以 Hadoop 为代表的分布式计算框架本身就是为了解决面向大数据的计算性能和可靠性问题而出现的，在这个方向上仍然需要进一步研究。传统的 Hadoop/MapReduce 框架在执行离线批处理任务时会有较好的表现，但在诸如实时流处理、交互式计算等方面却不尽如人意。如何提升分布式计算平台的性能是大数据领域的研究热点，目前也出现了一些积极的研究成果。

3. 大数据分析的功能问题

当前大数据分析的典型应用还体现在如何将传统数据处理平台的功能在更大、更复杂的数据集合上实现，如检索、查询统计、通用数据挖掘算法等。如何充分挖掘蕴藏在数据中的价值，实现基于传统数据处理平台无法突破的新功能，还需要各领域研究人员进一步探索。这方面的一个典型例子是：大数据分析应用方面的领先企业 Google、IBM 等依靠基于大数据的深度学习，在人工智能方面已经取得了突破性进展。

4. 大数据平台的易用性问题

一项新技术的出现能否得到快速、广泛应用，在很大程度上取决于该技术自身是否易于推广使用。现有的大数据平台往往需要编写特定的程序才能实现传统数据处理平台简单的功能，存在较大的易用性和可编程性问题。虽然诸如 Pig、Hive 等开源项目的出现，在一定程度上降低了大数据平台使用的门槛，但对于一般应用领域的数据分析人员以及专业性的大数据程序员来说，还难以像应用传统数据处理平台（如关系型数据库）一样方便地使用大数据平台。因此如何提升大数据平台的易用性，已经成为影响大数据技术应用推广的前提。

5. 大数据技术应用推进和落地问题

大数据技术起源于互联网行业，目前最成功的应用也在互联网行业，在其他行业的应用还处于初级阶段。我国的各级政府机关和各类传统行业，在日常管理和业务运行中

也积累了大量的数据。大数据的真正价值所在是深度价值发现和行业应用，如何推进大数据技术应用，唤醒这些沉睡的大数据资源，实现管理上的科学决策，开创新的业务模式，是这些数据拥有者所关心的问题。

6. 大数据分析的数据来源问题

大数据分析应用普遍存在着技术与资源分离的问题。相对于传统数据分析技术，大数据分析还是一个年轻的领域，缺乏诸如 SQL 之于传统关系型数据之类的通用计算模型，技术应用的门槛较高，最终导致了这样的困境：大数据分析技术集中在少数互联网企业和科研机构中，但这些机构要么只具备自身行业的数据（如互联网行业），要么缺乏验证技术所必需的样本数据，毕竟很难依靠模拟仿真的方式获取有效的分析对象；而真正拥有数据的机构（如政府机关和传统行业），却缺乏必要的大数据分析技术。如何走出这种困境，实现技术和资源的统一，也是大数据行业关注的热点。

5.2.2 大数据技术发展趋势

CCF 大数据专家委员会每年都会面向全体委员进行年度趋势预测调研。与 2014 年不同的是，2015 年 CCF 大数据专家委员会与中关村大数据产业联盟合作，有 50 余位联盟的企业家参与投票。

在此次调研活动中，候选项分为大数据科学、大数据技术、大数据系统和工程、大数据应用、数据资源、产业生态环境等 6 个不同方面，总计 54 个候选项。由 140 位 CCF 大数据专家委员和 50 位中关村大数据产业联盟企业家给出的 2015 年度大数据发展的十大趋势如下。

趋势一　结合智能计算的大数据分析成为热点

大数据与神经计算、深度学习、语义计算以及人工智能等其他相关技术的结合，成为大数据分析领域的热点。大数据分析的核心是从数据中获取价值，体现在从大数据中获取更准确、更深层次的知识，而不是对数据的简单统计分析。要达到这一目标，需要提升对数据的认知计算能力，让计算系统具备对数据的理解、推理、发现和决策能力，其背后的核心技术就是人工智能。近些年，人工智能的研究和应用又掀起新高潮，一方面得益于计算机硬件性能的提升，另一方面依靠以云计算、大数据为代表的计算技术的快速发展，使得信息处理速度和质量大为提高，能够快速、并行地处理海量数据。

趋势二　数据科学带动多学科融合，但其自身尚未形成体系

数据科学将带动多学科融合。但是，作为新兴的学科，数据科学的基础问题体系尚不明朗，其自身的发展尚未形成体系。在大数据时代，许多学科的研究方向从表面上看大不相同，但是从数据的视角来看，其实是相通的。随着社会的数字化程度逐步加深，越来越多的学科在数据层面趋于一致，可以采用相似的思想进行统一研究。数据科学是作为一个与大数据相关的新兴学科出现的，尽管真正支撑大数据发展的学科跨越还没有出现。在大数据处理的理论研究方面，新型的概率和统计模型将是主要的研究工具，学科基础理论的突破在 2015 年还很难出现。

趋势三　与行业数据结合，实现跨领域应用

跨学科领域交叉的数据融合分析与应用将成为今后大数据分析应用发展的重大趋势。大数据技术发展的目标是应用落地，因此大数据研究不能仅仅局限于计算技术本身。由于现有的大数据平台易用性差，而垂直应用行业的数据分析又涉及领域专家知识和领域建模，使得目前在大数据行业分析应用与通用的大数据技术之间存在鸿沟，缺少相互的交叉融合。因此，迫切需要进行跨学科和跨领域的大数据技术和应用研究，促进和推动大数据在典型和重大行业中的应用和落地。

趋势四　与“物云移社”融合，产生综合价值

大数据将与物联网、移动互联网、云计算、社会计算等热点技术领域相互交叉融合，产生很多综合性应用。近年来计算机和信息技术发展的趋势是，前端更前伸，后端更强大。物联网与移动计算促进了物理世界和人的融合，大数据和云计算提升了后端的数据存储管理和计算能力。今后，这几个热点技术领域将相互交叉融合，产生很多综合性应用。

趋势五　大数据多样化处理模式与软硬件基础设施逐步夯实

内存计算将继续成为提高大数据处理性能的主要手段。查询和分析的实时性对于人们能否及时获得决策信息非常重要，大批大数据实时查询分析系统将涌现，基于大内存的计算模式或将成为大数据实时处理的重要手段。以 Spark 为代表的内存计算逐步走向商用，并与 Hadoop 融合共存。

专为大数据处理优化的系统和硬件出现。由于大数据系统在计算、存储、高速缓存方面的需求不平衡，传统服务器与存储已经不适合构建高可靠和高性价比的大数据处理系统。硬件厂商将专门开发适合不同应用的硬件设备，大数据服务商基于现有服务器定

制专用处理设备，最终用户自行定制大数据系统中的硬件设备。在大数据处理过程中，更多的过程将被标准化和模块化，因此引入硬件加速模块（GPU、MIC 或者 FPGA/ASIC 等）成为提高系统性能的必然选择。

大数据处理多样化模式并存融合、一体化融合的大数据处理平台逐渐成为趋势。例如，支撑电信、金融、医疗、安全和电力等关键行业大数据应用的基础软件平台将呈现一体化形态，它以数据为中心，将操作系统、分布式存储、数据库等产品融合起来，对结构化、半结构化和非结构化等全数据进行高效存储与管理，并对应用提供统一的数据服务支撑接口。

趋势六　大数据安全和隐私

大数据安全和隐私问题依然是热点趋势。大数据应用所产生的隐私问题、大数据系统和体系存在的安全防范方面还没有实质性的进展和突破。有分析认为，将大数据安全作为趋势反映的是参加调研专家和用户的一种期盼、理解和关注。

2014 年，大数据安全又被专家提升到一个新的高度，认为大数据以及相关的关键资源涉及国家主权。

趋势七　新的计算模式将取得突破：深度学习、众包计算

尽管这两年深度学习大热，在一些特定的领域发挥了很大作用，但是大数据专家和企业界人士似乎更关注众包技术。

分布式计算是支撑大数据分析的必经之路。分布式计算依然存在两种形态，一种是物理资源集中式，比如具有分布式计算能力的集群、数据中心等；另一种是物理资源分散式，比如以前常用的 P2P 技术等。在很多大数据的应用场合，基于物理资源分散式会有更多的应用场景。典型的应用场景比如大规模的网页爬虫，采用物理资源集中式的方式就极容易被封堵，而采用众包计算这种物理资源分散式的分布式计算平台则可以避免这个问题。

趋势八　可视化与可视分析新方法被广泛引入，大幅度提高大数据分析效能

可视化与可视分析的研究发展迅速，国内专门从事可视化的研究人员将迅速上升。在可视化研究的数据对象从传统的单一数据来源扩展到多来源、多维度、多尺度的同时，用户也从专家扩展到广泛的非特定群体，研究则更强调方法的可扩展性和开发的简捷性。革命性的新方法和原有普遍适用的可视化技术和工具、面向领域和大众的可视化工具库，以及协同与众包的可视分析方式将被更广泛地引入到各种综合分析工具和平台

上，以大幅度提升各类大数据分析的易用性和有效性。

趋势九　大数据技术课程体系建设和人才培养是需要高度关注的问题

大数据技术的快速发展和行业应用需求的快速增长，使得目前技术市场上掌握大数据技术的人才严重短缺。因此，政府、高等院校和科研院所将加快建立大数据技术人才教育和培养体系，发展数据科学和工程专业，梳理和构建跨学科和领域交叉的大数据课程体系，融合计算机、数学分析统计、应用相关的学科，推动交叉学科数据分析技术的发展以及人才的培养。

趋势十　开源系统将成为大数据领域的主流技术和系统选择

以 Hadoop 为代表的开源技术拉开了大数据技术的序幕，大数据应用的发展又促进了开源技术的进一步发展。开源技术的发展降低了数据处理的成本，引领了大数据生态系统的蓬勃发展，同时也给传统数据库厂商带来了挑战。据统计，目前有超过 150 种开源大数据平台。这个数字还在增长中。

5.3　中国大数据发展战略与建议

5.3.1　大数据基础研究的发展战略与建议

基于前面对大数据产学研用现状与趋势的认识，为了科学合理地规划布局国家在“十三五”甚至更长时间内大数据基础研究的方向，建议我国在未来一段时间内对大数据科学重大基础研究的布局围绕着数据科学的学科体系、大数据计算系统与技术的基础理论、大数据驱动的颠覆性应用的基础问题三个层面展开。三者相互呼应、相互促进，形成国际领先的中国大数据发展战略和大数据科研生态环境。

1.“数据界”内涵与数据科学的学科体系研究

由于科学技术的进步，人类社会、物理世界和信息空间趋向于深度融合、交互影响、相互转换，三元世界的深度关联映射形成了一个庞大的数据空间，这个融合空间称为“数据界”(Data Nature)。人们认识数据界的普适规律所形成的基础学科可以称为“数据科学”。

当前，数据科学尚处于各个学科分割观察和各个行业独自发展的阶段，数据界的共性科学问题和内在基本机理尚不清晰。为此，建议融合数理科学、计算机科学、社会科学以及各类应用学科，采取归纳和演绎相结合，实验、唯象和理论框架

相结合，以研究相关性和复杂网络为主，探讨数据科学的学科体系。一些可能的研究方向包括：

数据谱系与分类体系研究。类似基因组方法，通过对三元世界以及各行业应用的大数据现象的系统性观察；研究数据的分类体系、内部数据基元（基本量纲）的相互作用力与关联关系规律；提出数据分类学理论。

数据计算思维与数据计算范式研究。探讨“数据→知识→智能决策”的认知机理；探讨从以算法的“计算复杂性”理论为核心的计算机科学研究发展到以大数据关联与融合分析的“数据复杂性”理论为核心的数据科学研究；从数据利用层面探讨数据的质量度量模型，研究类似基于“数据熵”的数据计算理论。

数据的社会效应理论研究。人是各种社会关系的总和，基于大数据研究个体人的各种关系，进一步研究不同社会群体和社会发展的数据化存在规律；基于社会的自组织、自洽、进化规律，研究数据界的仿社会计算理论，预测数据的社会效应；基于社会效应，研究数据价值模型和数据经济学理论。

数据的网络效应理论研究。类似物体的运动状态存在，任何数据在数据界的存在是网络化关联的。复杂网络理论是 20 世纪末、21 世纪初人们发现的普适性理论。在交通、经济、商业、健康等各行各业中的数据均以复杂网络形态存在。研究数据界的复杂网络规律和网络效应，发现数据存在的基本网络特征和数据关联的演化规律，进而支撑重大实际需求中的关键问题求解。

2. 大数据计算的系统与技术瓶颈问题

随着计算机和网络软硬件技术的发展，传统的计算机系统、网络传输、数据库、数据挖掘等 IT 技术虽然可以通过并行化、集群化在某些具体应用需求上缓解大数据处理带来的挑战，但现有系统与技术上的改良无法适应大数据复杂关联的网络效应、数据规模的指数增长、数据全生命周期内的迁移、异构多源数据的融合分析以及普适泛在的“人—机—物”交互所带来的压力。从系统与技术的基础方法层面来说，亟需摆脱传统 IT 技术的束缚，开放思路，研究面向大数据计算的新型系统体系和大数据分析理论，突破大数据认知与处理的技术瓶颈。具体而言，建议开展如下几个方面的研究：

面向大数据创新应用的新型系统结构研究。包括基于运行时系统的数据流计算模型，具体研究多核和分布式众核环境下的运行时系统及其数据流并行计算模型；面向智能计算的新型计算机体系结构，具体研究面向智能计算的神经元处理器及相应的计算机

体系结构；面向全生命周期的数据存储与实时计算的弹性架构与系统，具体研究异构网络数据全生命周期的存储、处理和实时计算的分布式弹性系统架构，大数据处理系统的效能评估模型与测试试验床；以及面向弱一致性约束和低成本冗余的异质数据管理与查询系统等。

大数据融合分析与深度挖掘的创新模型与方法。包括复杂网络分析与异质大图计算模型；跨媒体大数据的内容关联分析与协同智能计算模型；群体智慧与社会计算；人在回路的数据感知、预测与调控。

大数据计算的数理基础。包括数据采样与非完整数据环境下的统计学习与近似计算基础理论；高维特征空间和多模型融合计算的数学模型等。

大数据应用基础共性技术。包括面向物联网环境的绿色低能耗的大数据采集和通信；网络空间大数据的感知、测量、清洗与数据质量判定；大数据的交互式可视化与可视分析；大数据挖掘分析与机器学习工具和开发环境等。

3. 大数据驱动的颠覆性战略性应用基础研究

尽管各个行业都高度关注大数据，但就当下的情况而言，大数据无论是科学研究还是技术研发都还有很长的路要走，不宜不分行业地开展大数据的应用研究，以免造成大范围的人力、财力和物力的浪费。因此，建议通过研讨，挑选能对行业的传统模式带来颠覆性变革的几个行业进行试点。通过在试点行业的探索和应用，使大数据的技术方案逐步成熟，实现大数据技术与行业的无缝耦合之后再进行更大范围的拓展。为此，特别推荐如下几个领域：

网络大数据。互联网是大数据技术发展最快速也相对最成熟的领域，同时也是迄今为止产业最多颠覆性变革的领域。根据 2013 年发布的《中国互联网络发展状况统计报告》，目前最典型、最主要的互联网服务和应用包括网络新闻、搜索引擎、网络购物 / 网上支付、网络广告、旅行预订、即时通信 / 社交网络、博客微博、网络视频 / 网络音乐、网络游戏等，对当中的许多服务和应用，大数据的新理论、新技术大有用武之地，大数据将助推互联网服务和应用得到更好发展，反过来也将使大数据的新理论、新技术在互联网行业找到新的应用点，从而实现互联网与大数据两大新兴领域的有机结合。

安全大数据。这里主要是指网络空间安全。网络的开放性、复杂性和跨国性决定了网络安全是全球性的严峻挑战，它不仅是一个科技竞争，而且是一个与政治、社会、军事等问题紧密相关的，多方位、多层次和多领域的错综复杂的综合问题。当

前，网络安全已经成为国家安全的核心。国家在战略层面上对网络空间安全的重视，引发巨大的投资驱动。而大数据在处理网络空间复杂性等问题上具有先天优势，两者的结合使得网络空间安全问题成为大数据现实应用中最为活跃的领域之一。大数据技术将对面向网络安全的信息萃取技术、人机结合的网络安全分析、网络武器的防御与对抗技术等带来前所未有的挑战和机遇。同时，组建国家网络安全力量是网络安全大数据应用的重要步骤。

金融大数据。金融领域有着良好的大数据基础。借助新兴大数据技术的支持，金融业的两大根基——征信与风控，即将发生革命性的变化，同时将产生很多全新的金融与商业服务模式（如 O2O、互联网金融和众筹等）。具体而言，以大数据为代表的新型技术正在或将在两个层面对传统金融业引起颠覆：一是金融交易形式的电子化和数字化，具体表现为支付电子化、渠道网络化、信用数字化，是运营效率的提升；二是金融交易结构的变化，其中一个重要表现便是交易中介脱媒化，服务中介功能弱化，是结构效率的提升。伴随着大数据应用、技术革新及商业模式创新，金融业中的银行和券商也迎来巨大的转变。

生物组学大数据。基因组学可以说是大数据最经典的应用之一。基因测序的成本在不断降低，同时产生着海量数据。公司和研究机构可以通过研发基于基因组学大数据和云计算的高级算法来加速基因序列分析，让发现疾病的过程变得更快、更容易、更便宜，这为个性化医疗带来很多机遇。对制药公司来说，同样也可以预测哪些药物对特定的变异病人有效，作出更为科学和准确的诊断和用药决策，更大程度地提高药物疗效通过的概率。

健康医疗大数据。伴随着我国医疗卫生服务的信息化进程推进，已经产生了大量的医疗健康数据，主要包括来自医院的大量电子病历、区域卫生信息平台采集的居民健康档案等。而随着个人健康管理的推进，将产生越来越多的个人日常健康监测信息，这个数据的规模和增长速度将远超想象。但尽管目前各地的医疗信息系统已经电子化，但绝大多数的医疗数据处于归档状态。预计大数据解决方案将在居民健康档案数据管理和服务、医院大数据管理和服务，以及其他医疗领域（如医疗保险、生物制药、生命科学等）被广泛应用并带来颠覆性变化变革，譬如，使当今“One-Size-Fits-All”的医疗模式向个性化医疗模式转变。

4. 大数据的科研模式改革

目前，大数据已经渗透到了各行各业，每个行业都可能通过对大数据的研究与应用

大受裨益。但大数据这种跨领域、跨行业的特性，使得面对大数据的研究员需要对传统的模式进行改革。我们提出如下几项建议：1）产业界拥有真正的大数据，以及对大数据的迫切实际应用需求，因此大数据的研究与应用需要产业界和学术界密切配合；2）大数据的产学研用要快速发展，必须依靠良好的生态环境支持。而良性生态环境（包括政策、法律、法规等）的构建需要依托公共的而不是私有的数据平台体系，这就需要政府主导企业和科研院所参与的公共数据中心；3）在现有科研项目资助模式之外，探讨资助类似美国普林斯顿高等技术研究院以及持续性的交叉学科科学沙龙方式促进大数据学科创新研究。

5.3.2　大数据产业的发展战略与建议

1. 加快大数据开放共享的进度

在大数据时代，国家竞争力将部分体现为一国拥有数据的规模、活性以及该国解释、运用数据的能力；而国家数据主权体现了对数据的占有和控制。2012 年 7 月 10 日，联合国发布大数据政务白皮书《大数据促发展：挑战与机遇》指出，各国政府应当使用丰富的数据资源，更好地响应社会和经济指标。在大数据时代，数据主权将是继边防、海防、空防之后，另一个大国博弈的空间。

中国在大数据发展方面颇具数量上的优势。但相较于网络大数据和企业大数据的应用，中国对大数据，特别是政府大数据的利用几乎尚未起步，民众和企业几乎无法利用政府的大数据进行更加全面的决策来提高社会生产力。

根据调研，甚至包括乌拉圭、摩洛哥、摩尔达维亚等小国都已建立了国家级数据开放网站，将政府各部门的数据有机地整合到开放平台上，并进行详细分类和说明，网站及时更新，保证数据的时效性和有效性。所涉领域包括农业、气候与天气、基础设施建设、能源、金融与经济、环境、健康、犯罪、政府政策、法律、就业、公众安全、科学与技术、教育、社会与文化、旅游以及交通等。其中，金融与经济、健康、就业、教育等几大类是所调研的 26 个国家几乎都予以公布的数据大类。

以美国为例，美国的国家级数据开放网站 Data.gov 平台是一个由美国政府官方建立的政府信息的数据集散地和索引库。目前 Data.gov 提供了 97 274 个数据集，349 个数据应用工具，137 个手机应用程序，涵盖 173 个政府专业部门和子部门的数据。其目标是最大程度地将政府数据向公众开放，通过鼓励民众创新，使政府的数据得到更多的创新性应用。

在大数据时代各国争夺数据主权的关键时期，建议我国政府尽快高度重视制定大数据的国家战略，建立 data.gov.cn 这一国家级的数据开放网站，逐步开放政府拥有的大数据，让民众利用大数据提高社会生产力。中国作为世界上最大的发展中国家，在数据资源的管理和有效利用方面同样要探索一条符合中国国情的道路，尽快实现跨越式追赶和超越。

具体的建议有以下几条：

第一，国家层面建立 data.gov.cn 这一国家级的数据开放网站。从数据可获性、可分类性、异源融合、安全性等多角度全面地设计和实施建设大数据开放平台。

第二，成立数据管理的专门部门，建立数据管理的行政体系，通过国家数据平台发布管理制度。通过成立数据管理方面的职能部门，负责管理国家的数据资源。并通过建立数据管理的行政体系，从制度上为数据管理提供良好的规范流程。

第三，建立国家层面的数据标准，进而上升为国际标准。通过制定国家层面的数据资源的采集、存储、处理标准化体系。为国家层面的数据管理提供具体的操作指标。

2. 加快大数据隐私保护的立法

随着大数据时代的到来，收集、处理、传输、利用个人信息越来越简单，在促进社会经济发展、给人们带来便利的同时给犯罪分子可乘之机，也可能侵犯当事人的人格利益，甚至造成社会秩序紊乱，危害社会和谐稳定，个人数据及隐私安全问题日益凸显。而且，当今社会已出现一些专门收购和买卖个人数据信息的“生意人”或中介机构，信息交易已经呈现出专业化、行业化特点。个人信息的泄露渠道多种多样：商家通过问卷调查、会员登记等方式收集用户信息；消费者在就医、求职、买房、买保险或办理各种银行卡时填写的个人信息被出售；网络登录申请邮箱、注册进入聊天室或论坛时填写的个人信息被非法搜索或链接；物业泄露业主信息等。可以说，目前个人信息的泄露已经形成了产业链，此外网络流行的“人肉搜索”，让公民个人信息无处藏身。而大数据时代，存在于网络空间的多源数据更加复杂多样，信息泄露的渠道更加虚拟化、抽象化。

目前，我国还没有关于个人数据信息保护的专门法律法规，且信息产业的行业力量、行业组织不够强大，企业自律难以实现，政府的调控和保护能力不够强。而按现有的法律规定，泄露个人信息造成个人名誉权侵害后果、情节严重的才予以追究责任。对绝大多数受“骚扰”的人来说，受侵害的后果并不是很明显，也很难追究举证，因此有必要专门确立一部个人信息保护法。另外个人信息保护刑法条文草案的出台虽然能在一定程度上遏制严重侵犯个人信息的不法行为，但《刑法》是所有法律的最后屏障，单个

条文的《刑法》显然无法承担起保护个人信息的重任，所以一部“个人信息保护法”的出台已是众望所归。

为此，提出如下三条具体建议：

第一，通过立法在我国尽快建立个人信息和隐私保护制度，为公众创造一个良好的信息和隐私安全环境。

第二，个人信息和隐私保护法是一个涉及宪法、刑法、民法、行政法等多个部门法的综合体系，在刑法条文的细化与完善之外，应同步跟进行政处罚和相关的民事责任，以规制政府部门和金融、电信、交通、医疗等单位泄露个人信息和隐私的行为。

第三，应尽快推行网络实名制登录等，强制要求现实中的网民在虚拟世界不要忘了自己的责任，从而更规范个人言行。

3. 加快大数据基础设施的建设

基础设施是大数据产业高速发展的前提和保障。因此，我国需要加快推进“宽带中国”战略，加快下一代互联网、4G 通信网络、公共无线网络、电子政务网和物联网等网络基础设施的建设。具体而言需要采取以下几个方面的措施。

第一，国家财政加大对大数据基础设施的投入。中国的铁路、公路等基础设施建设正是有了国家财政的大力支持和投入才有了实质改善，大数据基础设施作为具有很强正外部性的设施，对于社会发展具有基础性带动作用，要从国家战略角度认识大数据基础设施的重要性，中央和地方财政需要充分认识大数据基础设施作为基础性设施的重要性，安排专项预算支出支持相关基础设施建设。

第二，采取基础设施建设优惠政策鼓励企业加大大数据基础设施建设投入。企业是大数据基础设施的主力军，也是大数据环境下，必须采取有效措施鼓励企业特别是民营资本进入大数据基础设施建设，将大数据基础设施建设作为国家重点公共设施项目，对其所得税采取实行“三免三减半”的税收优惠，即从取得经营收入的第一年至第三年可免交企业所得税，第四年至第六年减半征收。

第三，采取有效金融措施支持大数据基础设施建设。采取有效措施鼓励企业通过发行股票、债券等从资本市场进行大数据基础设施建设融资。鼓励商业银行发放贷款支持大数据基础设施建设，对中小企业参与大数据基础设施进行风险补偿机制，设立专门担保公司，解决相关企业从银行贷款的担保难问题，政府对相关基础设施建设贷款设立贴息制度，降低基础设施建设的财务成本。

第四，建立产业园区产生集群效应。中央和地方政府依据客观产业布局，规划大数

据产业园区，对入园企业给予房租、水电等费用的优惠，鼓励大数据企业入园创业。

4. 加大大数据人才培养的力度

政府应充分认识大数据的重要性和战略地位，从整个国家的角度积极布局，引导大数据全面发展。现如今，虽然大数据受到各行各行的高度关注，但在世界范围内都存在着大数据人才匮乏的现象，因此需要加大跨学科大数据人才项目的支持力度。在国家高等院校、科研机构建立大数据人才培养机制，国家资助或成立专项基金支持大数据关键技术研究。具体包括以下几项措施。

完善相关专业设置。在计算机或管理学相关学院设置大数据本科专业，根据企业对大数据从业人员专业技能需求，完善专业课程设置，建立产学研用合作机制，加大学校师资技能培训，增强学生课程实用性。根据国家专业硕士发展，在高校研究生院设置大数据专业硕士，鼓励多元知识背景学生报考，培养专业知识扎实、面向应用的大数据人才。加大高层次大数据科研人才培养力度。鼓励大数据企业到高校和科研院所中设立大数据教育奖学金，支持大数据人才的培养。

在相关学科中增加大数据相关课程。一方面大数据是一个以技术为本的方向，鼓励在数学、计算机科学与技术、软件工程等理工类专业中适当增加大数据相关的课程，鼓励相关学科学生在理工技术基础上从事大数据工作。另一方面大数据是一个应用驱动的领域，正是大规模数据的应用需求让整个社会认识到了数据的价值，要在经济学、金融、管理科学与工程等专业中适当增加大数据应用课程，使应用领域专业学生培养中渗透大数据的理念，推动大数据在应用领域的发展。

建立大数据从业人员认证机制。依托行业协会或者政府部门建立大数据从业人员的认证机制，允许大学在校学生和社会相关从业人员报考，根据需求设置不同门类，并根据难度设置不同的级别。认证要面向应用。

提高大数据从业人员待遇。依据大数据的价值创造能力，给予大数据从业人员与创造价值相符的待遇。对于基础扎实、做出重大创新、具有突出贡献的专家人才要通过设置专门的津贴机制，鼓励从业人员长期从事大数据工作。设立大数据青年人才基金，对在大数据从业的青年人才进行扶持。

附　录

附录 A　开源组织

A.1　ASF（Apache 软件基金会）

Apache 软件基金会（也就是 Apache Software Foundation，简称为 ASF），是专门为支持开源软件项目而办的一个非盈利性组织。在它所支持的 Apache 项目与子项目中，所发行的软件产品都遵循 Apache 许可证（Apache License）。

Apache 软件基金会正式创建于 1999 年 7 月，它的创建者是一个自称为“Apache 组织”的群体。这个“Apache 组织”在 1999 年以前就已经存在很长时间了，这个组织的开发爱好者们聚集在一起，在美国伊利诺伊斯大学超级计算机应用程序国家中心（National Center for Supercomputing Applications，简称为 NCSA）开发的 NCSA HTTPd 服务器的基础上开发与维护了一个叫 Apache 的 HTTP 服务器。

下面介绍它的发展历史。

最初 NCSA HTTPd 服务器是由 Rob McCool 开发出来的，但是它的最初开发者们逐渐对这个软件失去了兴趣，并转移到了其他地方，造成没有人来对这个服务器软件提供更多的技术支持。因为这个服务器的功能如此强大，代码可以自由下载、修改与发布，所以当时这个服务器软件的一些爱好者与用户开始自发起来，互相交流，分发自己修正后的软件版本，并不断改善其功能。为了更好进行沟通，Brian Behlendorf 自己建立了一个邮件列表，把它作为这个群体（或者社区）交流技术、维护软件的一个媒介，把代码重写与维护的工作有效组织起来。这些开发者们逐渐地把他们这个群体称为“Apache 组织”，把这个经过不断修正并改善的服务器软件命名为 Apache 服务器（Apache Server）。

这个名称是根据北美当地的一支印第安部落而来，这支部落以高超的军事素养和超人的忍耐力著称，并在 19 世纪后半期对侵占他们领土的入侵者进行了反抗。为了对这支印第安部落表示敬仰之意，取该部落名称（Apache）作为服务器名。但一提到这个名称，这里还流传着一段有意思的故事。因为这个服务器是在 NCSA HTTPd 服务器的基

础之上，通过众人努力，不断地修正、打补丁（Patchy）的产物，被戏称为“A Patchy Server”（一个补丁服务器）。而“A Patchy”与“Apache”是谐音，故最后正式命名为“Apache Server”。

后来由于商业需求地不断扩大，以 Apache HTTP 服务器为中心，启动了更多与 Apache 项目并行的项目，比如 mod perl、PHP、Java Apache 等。随着时间地推移、形势地变化，Apache 软件基金会的项目列表也不断更新变化中——不断有新项目启动，项目中止以及项目拆分与合并。比如一开始，Jakarta 就是为了发展 Java 容器而启动的 Java Apache 项目，后来由于 SUN 的建议，项目名称变为 Jakarta。但当时该项目的管理者也没有想到 Jakarta 项目因为 Java 的火爆而发展到如今一个囊括了众多基于 Java 语言开源软件子项目的项目，以至后来不得不把个别项目从 Jakarta 中独立出来，成为 Apache 软件基金会的顶级项目，Struts 项目就是其中之一。

最近，为了避免 SCO 与 UNIX 开源社区之间的纠纷降临在 Apache 软件基金会（ASF）身上。Apache 软件基金会开始采取一些措施，让众多的项目进行更多协调的、结构化管理，并保护自己的合法利益，避免一些潜在的合乎法律的侵犯（potential legal attack）。

Apache 软件基金会包括以下项目。

- HTTP Server：可以在 UNIX、MS-Windows、Macintosh 和 Netware 操作系统下运行的 HTTP 服务器的项目；
- Ant：基于 Java 语言的构建工具，类似于 C 语言的 Make 工具；
- AXIS2：Web 服务（SOAP、WSDL）的处理器，基于 AXIS1.X 重新构建；
- APR：（即 Apache Portable Runtime）C 语言实现的便携运行库的管理工具；
- Beehive：为了简单构建 J2EE 应用的对象模型；
- Cocoon：一个基于 XML 和组件技术的 Web 应用开发框架；
- DB：关于数据库管理系统的几个开源项目集合；
- Derby：一个纯 Java 的数据库管理系统；
- Directory：基于 Java 语言的目录服务器，支持 LDAP 等目录访问协议；
- Excalibur：Apache Avalon 项目的前身；
- Forrest：一个发布系统框架的项目；
- Geronimo：J2EE 服务器；
- Gump：整合管理器；
- Hadoop：并行运算编程工具和分布式文件系统；

- Harmony：一个兼容 Java 标准的 Java 语言的开源实现；
- HiveMind：一个服务（service）与配置（configuration）的微内核；
- iBATIS：一个基于 Java 语言的数据持久化框架；
- Incubator：为了帮助那些希望获取 Apache 软件基金会支持的计划进入 Apache 软件基金会的审核项目；
- Jackrabbit：内容仓库 API 标准（Content Repository for Java Technology API，即 JSR-170）的一个开源实现项目；
- Jakarta：在 ASF 中，基于 Java 语言的一组开源子项目的集合，现在包含的子项目有 BCEL、BSF、Cactus、Commons、ECS、HttpComponents、JCS、JMeter、ORO、Regexp、Slide、Taglibs、Turbine、Velocity；
- James：Java 语言实现的邮件新闻服务器；
- Labs：为基金会成员提供最新变更的思维的计划；
- Lenya：内容管理系统；
- Logging：一个开发可以在 C++、Java、Perl、PHP、.NET 计算机语言下运行的通用日志工具项目集合；
- Lucene：高性能的，基于 Java 语言的全文检索项目；
- Maven：项目集成构建工具；
- MyFaces：一个 JavaServer Faces（JSF）的实现框架；
- mod_perl：为 Apache 服务器提供 Perl 语言整合的项目；
- POI：提供 API 以供 Java 程式对 Microsoft Office 格式档案的读 / 写；
- Portals：与门户（Portal）技术相关的几个项目集合；
- Santuario：发展 XML 安全性方面的项目；
- Shale：在 Struts 之后发展起来，基于 Java 语言的 Web 应用框架；
- SpamAssassin：垃圾邮件过滤器；
- Struts：一个基于 J2EE 平台的 MVC 设计模式的 Web 应用框架；
- Subversion：一个软件版本管理系统；
- Tapestry：另一个 J2EE 平台的能产生动态、高性能 Web 应用的框架；
- TCL：为 Apache 服务器提供 Tcl 语言整合的项目；
- Tomcat：一个运行 Java Servlet 与 JavaServer Pages（JSP）的容器；
- Web Services：与 Web Services 技术相关的项目集合；

- Xalan：XML 转换处理器；
- Xerces：一组可以在 Java、C++、Perl 计算机语言下使用的 XML 解析器项目；
- Apache XML：XML 解决方案；
- XMLBeans：基于 Java 语言 XML 对象绑定工具；
- XML Graphics：发展 XML 与图形进行转换的计划项目；
- ……

与大数据有关的项目包括 Hadoop、HBase、Spark 等。下面介绍一下这些项目。

1. Hadoop

Hadoop 实现了一个分布式文件系统（Hadoop Distributed File System），简称 HDFS。HDFS 有高容错性的特点，用来部署在低廉的（low-cost）硬件上；而且它提供高吞吐量（high throughput）来访问应用程序的数据，适合那些超大数据集（large data set）的应用程序。HDFS 放宽了（relax）POSIX 的要求，可以以流的形式访问（streaming access）文件系统中的数据。

Hadoop 框架最核心的设计就是：HDFS 和 MapReduce。HDFS 为海量的数据提供了存储功能，MapReduce 为海量的数据提供了计算功能。

Hadoop 是一个能够对大量数据进行分布式处理的软件框架。Hadoop 以一种可靠、高效、可伸缩的方式进行数据处理。

Hadoop 是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理。

Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度。

Hadoop 还是可伸缩的，能够处理 PB 级数据。

此外，Hadoop 依赖于社区服务，因此它的成本比较低，任何人都可以使用。

Hadoop 是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在 Hadoop 上开发和运行处理海量数据的应用程序。

Hadoop 得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载（ETL）方面上的天然优势。Hadoop 的分布式架构将大数据处理引擎尽可能靠近存储，这对例如像 ETL 这样的批处理操作相对合适，因为类似这样操作的批处理结果可以直接进行存储。Hadoop 的 MapReduce 功能实现了将单个任务打碎，并将碎片任务（Map）发送到多个节点上，之后再以单个数据集的形式加载（Reduce）到数据仓库里。

2. HBase

HBase 是一个分布式的、面向列的开源数据库，该技术来源于 Fay Chang 所撰写的 Google 论文《 Bigtable ： 一个结构化数据的分布式存储系统》。就像 Bigtable 利用了 Google 文件系统（file system）所提供的分布式数据存储一样，HBase 在 Hadoop 之上提供了类似于 Bigtable 的能力。HBase 是 Apache 的 Hadoop 项目的子项目。HBase 不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是 HBase 是基于列而不是基于行的模式。

HBase 利用 Hadoop HDFS 作为其文件存储系统；Google 运行 MapReduce 来处理 Bigtable 中的海量数据，HBase 同样利用 Hadoop MapReduce 来处理 HBase 中的海量数据；Google Bigtable 利用 Chubby 作为协同服务，HBase 利用 Zookeeper 作为对应。

3. Spark

Spark 是 UC Berkeley AMP lab 所开源的类 Hadoop MapReduce 的通用的并行计算框架，Spark 基于 MapReduce 算法实现的分布式计算，拥有 Hadoop MapReduce 所具有的优点；但不同于 MapReduce 的是 Job 中间输出结果可以保存在内存中，从而不再需要读写 HDFS，因此 Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法。

Spark 比 Hadoop 更通用。

Spark 提供的数据集操作类型有很多种，不像 Hadoop 只提供了 Map 和 Reduce 两种操作。比如 map、filter、flatMap、sample、groupByKey、reduceByKey、union、join、cogroup、mapValues、sort、partionBy 等多种操作类型，Spark 把这些操作称为 transformation。同时还提供 count、collect、reduce、lookup、save 等多种 action 操作。

这些多种多样的数据集操作类型，给开发上层应用的用户提供了方便。各个处理节点之间的通信模型不再像 Hadoop 那样就是唯一的 Data Shuffle 一种模式。用户可以命名，物化，控制中间结果的存储、分区等。可以说编程模型比 Hadoop 更灵活。

不过由于 RDD 的特性，Spark 不适用那种异步细粒度更新状态的应用，例如 Web 服务的存储或者增量的 Web 爬虫和索引。换句话说，Spark 不适合那种增量修改的应用模型。

在分布式数据集计算时通过 checkpoint 来实现容错，而 checkpoint 有两种方式，一个是 checkpoint data，一个是 logging the updates。用户可以控制采用哪种方式来实现容错。

Spark 通过提供丰富的 Scala、Java、Python API 及交互式 Shell 来提高可用性。

A.2 Linux Foundation（Linux 基金会）

Linux 基金会是一个非盈利性的联盟，其目的在于协调和推动 Linux 系统的发展，以及宣传、保护和规范 Linux，该组织是 2007 年由开源码发展实验室（Open Source Development Lab，OSDL) 与自由标准组织（Free Standards Group，FSG) 联合成立的，其中 MeeGo 是 Linux 基金会管理下的 Linux 操作系统。

下面介绍它的发展历史。

2000 年，Linux 基金会成立，旨在赞助 Linux 创始人 Linus Torvalds 的开发工作，并通过领先的技术和来自世界各地的开发人员合作。Linux 基金会保护其成员和开源开发社区资源，以确保 Linux 仍然是免费的、技术上先进的 Linux。

2010 年 5 月 10 日，Linux 基金会对外透露，基金会任命了 Linux 业界资深人士 Cliff Miller 出任基金会中国区运营总监。这也是 Linux 基金会首次在中国设立分支机构，主要任务是在中国推广 Linux 的应用，包括所有基于 Linux 的产品，如 MeeGo、Android 和其他 Linux 操作系统。

2010 年 6 月 29 日，Linux 基金会执行董事吉姆·策姆林（Jim Zemlin）表示：在嵌入式领域，Linux 排名第一，在超级计算机领域也是 Linux 领先，服务器方面 Linux 与微软的 Windows 各占半壁江山，而桌面则微软领先。从中可以看到 Linux 在绝大部分领域都有出色表现。

2011 年 4 月 8 日，Linux 基金会表示 Linux 已经战胜微软：Linux 基金会执行理事吉姆·策姆林在接受《 Network World》采访时称，在莱纳斯·托瓦尔兹（Linus Torvalds）开发出他著名的操作系统内核 20 年之后，Linux 与微软之间的斗争已经结束，并且 Linux 取得了胜利。

截至 2011 年 12 月，Linux 已经发展成为计算领域中的强有力工具，从纽约证券交易所到手机再到超级计算机和消费电子设备，Linux 都在默默无闻地奉献着。

A.3 Free Software Foundation（开源软件基金会）

自由软件基金会（Free Software Foundation，FSF）是一个致力于推广自由软件的美国民间非盈利性组织。它于 1985 年 10 月由理查德·斯托曼建立。其主要工作是执行 GNU 计划，开发更多的自由软件。

从其建立到 20 世纪 90 年代中期自由软件基金会的基金主要用来雇用编程人员来发展自由软件。但是从 20 世纪 90 年代中期开始写自由软件的公司和个人太多了，因此自由软件基金会的雇员和志愿者主要在自由软件运动的法律和结构问题上工作。

自由软件基金会最早的目的在于促进自由软件的开发，但自由软件基金会也有开发 GNU 操作系统的任务。

自由软件基金会具有施行 GNU 通用公共许可证和其他 GNU 许可证的能力和资源，但自由软件基金会只对它拥有版权的软件负责。其他软件必须由它们自己的拥有人来负责，原因是从法律规定上自由软件基金会无法为这些其他软件负责。自由软件基金会每年约接触到 50 个违反 GNU 通用公共许可证的事件，但它会试图不通过法院使对方遵守 GNU 通用公共许可证。

GNU 通用公共许可证是自由软件工程中最普及的许可证。目前的版本（版本 2）是 1991 年发表的，但自由软件基金会正在进行版本 3 的工作。自由软件基金会还发布了 GNU 宽通用公共许可证和 GNU 自由文档许可证。

自由软件基金会列出了一个高优先计划（high priority project）列表，FSF 认为这些计划需要自由软件社群们的注意，这些计划所开发的项目目前并没有自由软件可以用来取代非自由软件。

目前计划：

- GNU PDF
- Gnash——自由的 Flash Player 软件
- Coreboot——一个自由 BIOS 的运动
- 一个代替 Skype 的自由软件
- 一个自由的视频编辑软件
- 一个代替 Google Earth 的自由软件
- gNewSense——一个只使用自由软件的 GNU/Linux 操作系统
- GNU Octave——一个代替 Matlab 的自由软件
- 代替 OpenDWG 函式库
- 一个支援 RAR 第 3 版的自由软件
- GDB 的可逆性侦错
- 自由的网络路由器驱动程式
- 一个代替 Oracle Forms 的自由软件

A.4 开源软件中心（中国开源软件推进联盟）

中国开源软件推进联盟（英文全称：China OSS Promotion Union，英文缩写：COPU，以下简称“联盟”）是在政府主管部门指导下，由致力于开源软件文化、技术、产业、教学、应用、支撑的企业、社区、客户、大专院校、科研院所、行业协会、支撑机构等组织自愿组成的、民主议事的民间行业联合体，非独立社团法人组织。该联盟在2004 年 7 月 22 日于北京成立。

联盟的宗旨是为推动中国开源软件（Linux/OSS）的发展和应用而努力；为促进中日韩以及中国与全球关于开源运动（Linux/OSS）的沟通、交流与合作而努力；为促进全球开源运动（Linux/OSS）做出贡献而努力。

联盟的作用是为推动 Linux/OSS 的发展，充分发挥联盟在政府与企业之间有关立法、政策、规划和环境建设方面的桥梁、纽带与促进作用；充分发挥联盟在企业与用户、企业与企业、企业与社区、中外企业 / 社区间、企业与科研、教育、支撑机构之间关于研发、生产、教育、培训、测试、认证、标准化、应用等方面沟通、交流、合作、推进的桥梁、纽带与促进作用。

附录 B　产业园与政策措施

B.1　我国各地大数据产业园区介绍与相关政策

《国民经济和社会发展十二五规划纲要》指出，“推动中小企业调整结构，提升专业化分工协作水平。引导中小企业集群发展，提高创新能力和管理水平。”这为产业园区指明了发展方向，奠定了理论基础。从目前的地方经济发展趋势看，各种产业园区确实逐渐成为区域经济发展的引擎，带动着区域整体实力提升。

1. 重庆市大数据产业发展计划及措施

全产业链目标：打造 2 ~ 3 个大数据产业示范园区，培育 10 家核心龙头企业。

根据《重庆市大数据行动计划》内容的要求，到 2017 年，重庆要在虚拟技术、云计算平台技术、海量数据存储、数据预处理、新型数据挖掘分析、信息安全技术、大数据关键设备等 7 大领域，突破一批关键技术，推动大数据技术在电子政务、民生服务、城市管理及相关重点行业广泛应用，将大数据产业培育成全市经济发展的重要增长极。打造 2 ~ 3 个大数据产业示范园区，培育 10 家核心龙头企业、500 家大数据应用和服务企业，引进和培养 1000 名大数据产业高端人才，形成 500 亿元大数据产业规模，建成国内重要的大数据产业基地。

引自：重庆市人民政府网

链接：http://www.cq.gov.cn/publicinfo/web/views/Show!detail.action?sid=1111664

2. 陕西西咸新区沣西大数据产业园发展规划

发展思路及目标：以实现数据的“规模化集中吞吐、深层次整合分析、多领域社会应用、高效益持续增值”为方向，大力发展数据存储、呼叫中心、IDC 中心、灾备中心、数据交换共享平台等业态，力争用十年左右时间，将园区建设成为国家政务信息资源聚集地、社会商务资源集散地和西部超算中心。

战略规划和推进路径：第一步，以“存”为基础，引进龙头企业，形成行业核心数据的存储优势；第二步，以“用”为核心，推动跨行业、跨部门数据的分析整合，创新大数据产业发展模式；第三步，以“强”为目标，以拉长产业链、提升价值链为重点，辐射带动相关领域，实现信息产业的集群化发展。

引自：陕西省西咸新区沣西新城管理委员会

链接：http://www.fcfx.gov.cn/zwb/xwzx/jqdsj/fxxcdsjggfb/2358.htm

3. 福建省大数据产业园十条发展措施

福建省政府出台支持大数据产业重点园区加快发展十条措施，推动数字福建（长乐）产业园、中国国际信息技术（福建）产业园建设成为全省大数据产业重点园区和“数字福建”建设的重要承载基地。

福建省将从完善园区发展规划、引进培育产业龙头、推动资源汇聚开发、建设大数据创新平台、加强人才引进培养、做好园区用地保障、确保园区用电需求、强化园区网络支撑、实施财税优惠政策、提高安全保障能力等方面，支持大数据产业重点园区加快发展。

引自：福建省人民政府网

链接：http://www.fujian.gov.cn/ztzl/jkjshxxajjq/zcwj/201410/t20141020_885355.htm

4. 武汉市大数据产业发展行动计划

为抢抓大数据产业发展机遇，加速武汉市信息化建设和信息产业发展，提高自主创新能力，形成大数据产业和应用特色优势，争取在我国相关产业发展的先导地位，参照《国家新型城镇化规划（2014-2020 年）》、《2006-2020 年国家信息化发展战略》、《武汉市国民经济和社会发展“十二五”规划》、《武汉市信息产业和信息化发展“十二五”规划》和《武汉市智慧城市建设总体规划和设计》等，特制定《武汉市大数据产业发展行动计划（2014-2018）》

武汉市通过明确指导思想，制定发展目标，加强大数据产业基地建设，创立大数据中心，打造大数据龙头企业和知名品牌，搭建大数据交易平台，加强大数据安全管理，完善组织保障、政策保障、人才保障、资金保障等方面的保障措施，支撑武汉市大数据产业发展行动计划全面实施。

引自：湖北省武汉市人民政府办公厅

链接：http://www.whbgt.gov.cn/documents.php?c=5&id=1121

5. 天津市滨海新区大数据产业园行动方案

按照《滨海新区大数据行动方案（2013-2015）》，新区将规划建设大数据产业发展园区，促进大数据领域企业、项目、人才等资源向新区聚集，构筑高端产业体系。新区将按照“聚集创新资源、突破关键技术、推动创新应用、促进辐射带动”的模式，启动应用示范工程，部署建设大数据产业园区。

到 2015 年将实现“2111”发展目标，即聚集 200 家大数据企业，引进 10 个信息中

心和数据中心项目，实施 10 项典型应用示范项目，形成 10 项杀手锏技术产品。新区将规划建设 1 个占地 2.5 万平方米的大数据产业示范基地和 3 个产业园区，并通过出台政策和专项资金，引进国内外知名大数据企业研发中心落户，重点推进大数据在港口、海洋以及交通物流等领域的应用。到 2017 年，建成具有国际竞争力的大数据产业基地和数据资源聚集服务区。

引自：天津市工业和信息化委员会

链接：http://www.tjec.gov.cn/jwnew/50618.htm

6. 贵阳大数据产业行动计划

借助大数据产业到来的东风，贵阳市积极行动，迅速出台《贵阳大数据产业行动计划》。计划明确：“到 2016 年底，贵阳大数据相关产业规模达到 540 亿元，约占全市信息产业总产值的 30%，成为经济发展的重要增长极。”

以此为目标，贵阳市将以中关村贵阳科技园为统筹，以贵安新区大数据基地为重要载体，实施“四大工程”，“软硬结合”聚焦发展大数据云服务与智能终端两大产业集群，“云、管、端垂直整合”建设大数据特色产业基地，形成“4 园 4 基地”的大数据产业园区布局，将大数据产业打造成贵阳重要的战略性新兴产业。

引自：贵州省贵阳市人民政府官网

链接：http://www.gygov.gov.cn/art/2014/5/14/art_18325_606276.html

7. 上海推进大数据研究与发展三年行动计划（2013-2015 年）

《上海推进大数据研究与发展三年行动计划（2013-2015 年）》，通过强化顶层设计形成主体架构，建立协同共享机制，加强统筹规划，充分沟通、协调、调动各方资源，延伸大数据技术链、服务链、价值链。并以市场需求为导向，加强基础研究，突破大数据关键技术瓶颈，不断探索创新商业模式，培育和挖掘满足国内市场特性的新业态、新模式，支撑和促进经济社会发展。同时营造和完善大数据技术和产业发展所需的政策环境、融资环境、创业环境以及公共服务体系，推动大数据技术与城市经济社会各领域相关应用的深度融合。

发展目标如下：

1）研究数据科学基础理论，突破大数据共性关键技术，研制具有自主知识产权的若干大数据硬件装备，达到国际领先水平；

2）遵循市场需求牵引、应用导向的业务发展模式，开发一批具有产业核心竞争力

的大数据软件产品；

3）突出企业创新主体地位，建设 6 个以上行业大数据公共服务平台，支持 6 类以上大数据商业应用系统的研制，培育一批带动本地数据产业发展的行业龙头企业；

4）汇聚产业和行业创新活力，制定有利于大数据产业发展的标准、规范和政策，培养和引进千名高端数据人才。

引自：上海市科学技术委员会

链接：http://www.stcsm.gov.cn/gk/ghjh/333008.htm

8. 中关村管委会关于《加快培育大数据产业集群推动产业转型升级的意见》

为加快培育中关村大数据产业集群，充分发挥大数据在工业化与信息化深度融合中的关键作用，推动中关村国家自主创新示范区产业转型升级，根据《中关村国家自主创新示范区战略性新兴产业集群创新引领工程（2013-2015 年）》，特制定《加快培育大数据产业集群推动产业转型升级的意见》。

到 2016 年，中关村要形成大数据完整产业链和产业集群，培育 500 家大数据企业和一批领军企业，具备大数据应用能力的企业数量超过 5000 家，建成 10 个以上行业大数据应用平台，实现对国内外大数据创新资源的聚集整合，带动产业规模超过 1 万亿元。

引自：中关村科技园区管理委员会

链接：http://www.zgc.gov.cn/zcfg10/93637.htm

参 考 文 献

[1] Harter T, Borthakur D, Dong S, 等 .Analysis of HDFS under HBase: a facebook messages case study[C]//FAST. 2014: 199-212.

[2] Chen W, Chen Y, Mao Y, 等 .Density-based logistic regression[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 140-148.

[3] Chen W, Chen Y, Weinberger K Q. Maximum variance correction with application to A* search[C]//Proceedings of the 30th International Conference on Machine Learning (ICML-13). 2013: 302-310.

[4] Hölzle U. Brawny cores still beat wimpy cores, most of the time[J]. IEEE Micro, 2010, 30(4).

[5] Albrecht C, Merchant A, Stokely M, 等 .Janus: Optimal Flash Provisioning for Cloud Storage Workloads[C]//USENIX Annual Technical Conference. 2013: 91-102.

[6] Ousterhout J, Agrawal P, Erickson D, 等 .The case for RAMClouds: scalable high-performance storage entirely in DRAM[J]. ACM SIGOPS Operating Systems Review, 2010, 43(4): 92-105.

[7] Yang D, Zhong X, Yan D, 等 .NativeTask: A Hadoop compatible framework for high performance[C]//Big Data, 2013 IEEE International Conference on. IEEE, 2013: 94-101.

[8] Crotty A, Galakatos A, Dursun K, 等 .Tupleware: Redefining Modern Analytics [J]. arXiv preprint arXiv:1406.6667, 2014.

[9] Lim H, Han D, Andersen D G, 等 .MICA: a holistic approach to fast in-memory key-value storage[J]. management, 2014, 15(32): 36.

[10] Murray D G, McSherry F, Isaacs R, 等 .Naiad: a timely dataflow system[C] // Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles. ACM, 2013: 439-455.

[11] 维克托·迈尔 – 舍恩伯格，肯尼思·库克耶 .（译）。《大数据时代：生活，工作与

思维的大变革》[J]. 盛杨燕，周涛，译。2013.

[12] Morstatter F, Pfeffer J, Liu H, 等 .Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose[C]//ICWSM. 2013.

[13] Agarwal S, Milner H, Kleiner A, 等 .Knowing When You’re Wrong: Building Fast and Reliable Approximate Query Processing Systems[J].

[14] Zhu, Jonathan J. H., Qian Mo, Fang Wang, Heng Lu. A random digit search (RDS) Method for sampling of blogs and other user-generated content. Social Science Computer Review (2010): 0894439310382512.

[15] Leskovec J, Faloutsos C. Sampling from large graphs[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 631-636.

[16] Wong P C, Shen H W, Johnson C R, 等 .The top 10 challenges in extreme-scale visual analytics[J]. IEEE computer graphics and applications, 2012, 32(4): 63.

[17] Tu S, Kaashoek M F, Madden S, 等 .Processing analytical queries over encrypted data[C]//Proceedings of the VLDB Endowment. VLDB Endowment, 2013, 6(5): 289-300.

[18] Yao A C. Protocols for secure computations[C]//2013 IEEE 54th Annual Symposium on Foundations of Computer Science. IEEE, 1982: 160-164.

关键词索引