中国计算机学会
China Computer Federation

中国大数据技术与产业发展白皮书

（2013）

中国计算机学会大数据专家委员会

2013年12月1日

序　言

近两年来，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 6 月成立了“大数据专家委员会”，其宗旨是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，并对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精心组织下，花了大半年时间撰写了这本《中国大数据技术与产业发展白皮书（2013 年）》。

中国计算机学会大数据专家委员会的 110 位专家（不包括最近正在遴选的第三批专家委员）来自大学、科研单位、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，白皮书的编写集中了各个领域众多专家的知识和智慧，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《中国大数据技术与产业发展白皮书》的目的在于为业界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。白皮书包括六部分内容，第一章介绍大数据的发展背景，第二章阐述大数据典型应用领域的现状，第三章阐述大数据技术体系的发展现状，第四章讨论大数据 IT 产业链与生态环境，第五章分析了大数据人才资源情况，第六章探讨大数据的发展趋势并提出相关建议。

大数据成为热点以后，众说纷纭。推动者认为是“上帝给中国崛起准备的礼物”；泼冷水者认为是又一场“泡沫”。实际上所谓大数据主要是干三件事：一件是提高“数据意识”，用已经掌握的技术大力推动数据产业，这方面主要是企业界要做的事。在企业看来，不管是大数据还是小数据，只要能给企业带来价值，就是好数据。对于数据意识薄弱的发展中国家，经过大数据浪潮的洗礼，提高对数据资源的掌控能力，无疑是一件好事。第二件事是解决现有计算机系统和软件不能对付急剧增长、种类繁多的数据（尤其是网络数据）这一挑战问题，研究各种采集、整理、存储、处理和呈现大数据的变革性技术。各国专家对大数据的定义大都是着眼于这一挑战，这主要是科技界（包括大企业的研发机构）要做的事。介于这两者之间的第三件事是，推广近几年开始应用的不同于传统事务处理、传统数据库和小样本建模分析技术的大数据处理新方法，如深度学习、MapReduce 、

Hadoop 软件和数据中心的分布式服务器集群等技术。这是从传统的数据处理转向大数据处理的过渡阶段。

本白皮书洋洋洒洒 8 万字，其中分量最重的是第二章和第三章。第二章介绍大数据的典型应用，对应上述第一件事和第三件事。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。发现典型的大数据应用案例，宣传推广应用大数据技术的经验是本白皮书的主要动机，今后我们会更加关注应用案例的分析介绍。

第三章分析大数据技术体系的现状，对应上述第二件事。专家委员中多数是科研工作者，最熟悉的是本领域科学技术研究的进展，最擅长的是探讨技术发展趋势，分析科学研究和技术开发中面临的问题与挑战。本白皮书的主要价值可能体现在对大数据技术的分析方面。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。这部分内容反映在第六章 6.2.2 节“大数据的技术发展趋势”中，希望能对读者有所启迪。在其他几章，企业界和政府部门的专家也表达了一些真知灼见，如第四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四个趋势、大数据产业发展的主要瓶颈等都有独到的观点。第五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。

由于时间和篇幅有限，白皮书只选择的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白皮书。在后续工作中，大数据专家委会将继续不断完善和丰富白皮书的内容，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单行本。本白皮书是专家委第一次组织撰写，虽反复修改了十余次，但书中肯定还存在一些内容和文字的错误，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和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

2013 年 12 月 1 日

致　谢

众多大数据专家委委员参与了白皮书的撰写工作，其中，第一章大数据的发展背景部分主要由赵国栋完成，第二章大数据典型应用现状由潘柱廷、苗凯翔和张自力负责整理，其中互联网与大数据由沈烁、查礼、雷涛等撰写；网络通信与大数据由童晓渝、孙少陵、罗圣美、张宝峰等撰写，网络空间安全与大数据由潘柱廷、金波、杜跃进、何利文、胡晓峰等撰写；城镇化、智慧城市与大数据由苗凯翔、李剑等撰写；金融与大数据由赵国栋、石勇、白硕等撰写；健康医疗与大数据由苗凯翔等撰写；生物信息、制药与大数据由胡斌等撰写。第三章大数据技术体系现状由杜小勇、舒继武、黄宜华、王文俊、李翠平、于戈、刘伟、袁晓如等撰写，第四章大数据 IT 产业链与生态环境由朱扬勇、施水才、齐红威等撰写；第五章大数据人才资源由朱扬勇、王元卓、靳小龙等撰写，第六章主要由李国杰、程学旗、潘柱廷、王元卓、靳小龙等撰写。程学旗、王元卓、靳小龙负责材料组织和统稿等工作。方锦清、张学工、季统凯、邓波、张师超、陈继东、王意洁、王国胤、周霞、顾宁等大数据专家委委员积极参与了白皮书的撰写，不仅提供了素材，还参与了白皮书的修改工作。由于白皮书经过了多次反复的修改，对参与专家的统计可能还有遗漏，在此表示歉意。对所有参与白皮书编写的专家表示感谢。

目 录

第一章 大数据的发展背景

1.1 大数据的起源

信息科技经过 60 余年的发展，已经渗透到国家治理、经济运行的方方面面。政治、经济活动中很大一部分的活动都与数据的创造、采集、传输和使用相关，随着网络应用日益深化，大数据应用的影响日益扩大。根据国外一些机构测算，全世界数据总量以每两年翻一番的速度递增。换句话说，最近两年产生的数据总量相当于人类有史以来所有数据量的总和。在这个大背景下，从公司战略到产业生态，从学术研究到生产实践，从城镇管理乃至国家治理，都将发生本质的变化。

近些年来，我国一些代表性的企业，如华为，开拓美国市场屡屡受阻，已经传达了明确清晰的信号，即美国政府对自家数据安全的重视程度，已经到了不能让任何外国信息基础设施产品供应商染指的地步。备受世人瞩目的“棱镜门”，更是深刻暴露一些大国在利用信息技术领域的优势，有计划、有步骤的采集各国的“数字 DNA”。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、活性以及解释、运用数据的能力；国家网络空间主权[1]体现对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。没有数据安全，也就没有国家安全。

大数据技术虽然发源于信息科技，但其影响已经远远超出信息行业。数据已经存在于全球经济中的每一个部门，就如固定资产和人力资本等生产要素一样，如果没有它，许多现代经济活动就不会发生。我们观察到一些新兴的互联网公司，利用新技术，大规模地收集数据，预判客户行为，然后在不同的行业纵横捭阖。他们剑锋所指，现代服务业无不受其锋芒所迫，或随波逐流，或奋起反击。但缺少数据资产、缺少强大的数据分析能力，这类第三产业公司无疑处在被颠覆的边缘。另一方面，也看到传统行业的公司，数十年如一日的坚持积累当时被视作“废料”的数据，现在回头审视这些数字化的资产，居然一跃成为人类的宝库。凭借独一无二的“数据资产”[2]，公司进入相关行业，易如反掌。我们回头审视产业

[1]网络空间英文译为 cyberspace

[2]数据成为资产，参见国金证券大数据系列研究报告《大数据时代的三大发展趋势及投资方向》

的起起伏伏，就会发现决定产业兴衰的根本性因素，已经不是一城一地的争夺，也不仅仅是依靠土地、人力、技术、资本这些传统的生产要素，需要对“数据资产”重新进行优化配置。

大数据时代，有两点非常有利于中国信息产业跨越式发展。第一，大数据技术以开源为主，迄今为止，尚未形成绝对技术垄断。即便是 IBM、甲骨文等行业巨擘，也同样是集成了开源技术和该公司已有产品而已。开源技术对任何一个国家都是开放的，中国公司同样可以分享开源的蛋糕，但是需要以更加开放的心态、更加开明的思想正确地对待开源社区。第二，中国的人口和经济规模决定了中国的数据资产规模冠于全球。这在客观上为大数据技术的发展，提供了演练场。第二点亟待政府、学术界、产业界、资本市场四方通力合作，在确保国家数据安全的前提下，最大程度地开放数据资产，促进数据关联应用，释放大数据的巨大价值。

大数据的诞生是信息技术发展的必然结果。如交通业，在初期需要修建、疏通道路，当道路发展到一定的里程，就为汽车产业的发展提供了基础。当汽车普及时，人们关注的焦点就会迁移到汽车运输的“货物”。信息产业的发展亦可以以此类比。宽带网络建设是信息高速公路，物联网、云计算等技术相当于汽车和仓库，而大数据则是大家普遍关注的“货物”。

➢ 信息科技进步

如果把信息技术的不断进步，看成世界万物持续数字化的过程，则会理出一条清晰的主线。信息科技具有三个最核心和基础的能力：信息处理、信息存储和信息传递。几十年来，信息科技的这三个能力的飞速进步，是人类科技史上最为激动人心的故事之一。在这段波澜壮阔的历史中，信息的处理和储存能力获得了成千上万倍的提升。

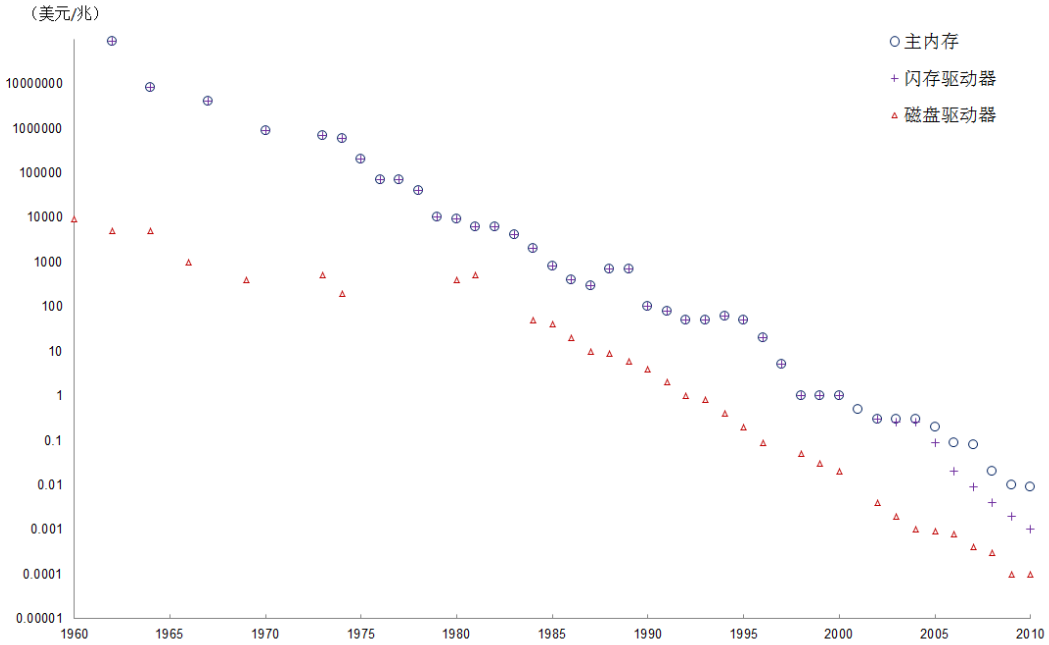

图1-1： 存储价格的下降 [3]

1977 年，世界上第一条光纤通信系统在美国芝加哥市投入商用，数据传输速率为 45Mb/s，自此拉开信息传输能力大幅跃升的序幕。有人甚至将光纤传输带宽的增长规律称为超摩尔 [4]定律，认为带宽的增长速度比芯片性能提升的速度还要快。事实上，存储的价格从上个世纪 60 年代 1 万美元 1M，降到现在的 1 美分 1G 的水平，其价差高达亿倍。在几年前在线实时观看高清电影还是难以想象的，而现在却变得习以为常了。网络的接入也从有线连接方式向高速无线连接的方式转变。毫无疑问，网络带宽和大规模存储技术的高速持续发展，为大数据时代提供了廉价的存储和传输服务。因而本书假定存储和带宽不再是制约数据应用的因素。

[3]来源： Plattner and Zeier, “In-Memory Data Management”, 2011, p. 15-16; * Driscoll, “Big Data Now”。

[4]摩尔于 1929 年出生在美国加州的旧金山，曾获得加州大学伯克利分校的化学学士学位，并且再加州理工大学获得物理化学博士学位。20 世纪 50 年代中期，他和集成电路的发明者罗伯特·诺伊斯一起在威廉·肖克利半导体公司工作。1968 年，摩尔和诺伊斯创办了大名鼎鼎的英特尔公司。自 1982 年起的 10 年间，微电子技术共有 22 项重大突破，其中由英特尔公司开发的就有 16 项之多。摩尔在 1974 年至 1987 年间担任英特尔公司的总裁和首席执行官，英特尔公司在微机时代和微软公司一道主宰了整个产业的发展。

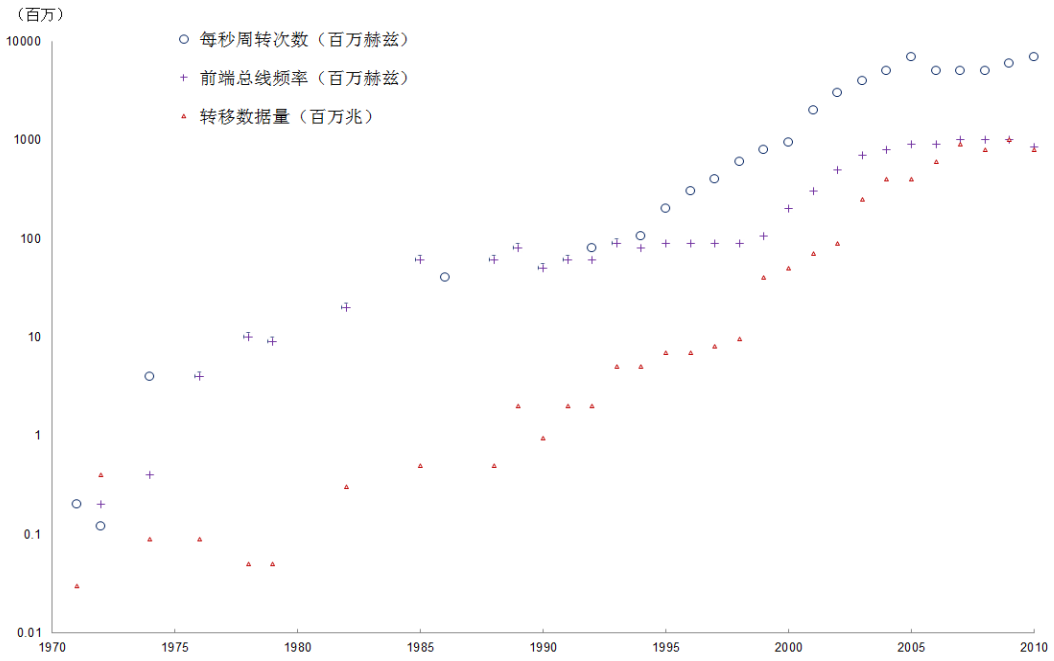

图1-2： 网络带宽的增加 [5]

➢ 互联网

互联网的出现是科技史上可以比肩“火”与“电”的发明。互联网把每个人桌面上的计算机连接起来，改变了人们的生活，成为人们获取各类数据的首要渠道。

互联网的内在机理，使其成为更接近消费者、最理解消费者的工具和平台。互联网没有删除键，人们在互联网上的一言一行都被忠实地记录。古代皇帝身边总有一位兢兢业业的史官，随身携带纸笔，记下皇帝的起居作息、金口玉言。互联网就像每个人的“史官”，它从不知疲倦，事不分大小，悉心而精准地记录着一切。事实上，这位“史官”记录的就是大家的数字化生活。

[5]来源： Plattner and Zeier, “In-Memory Data Management”, 2011, p. 15-16; * Driscoll, “Big Data Now”

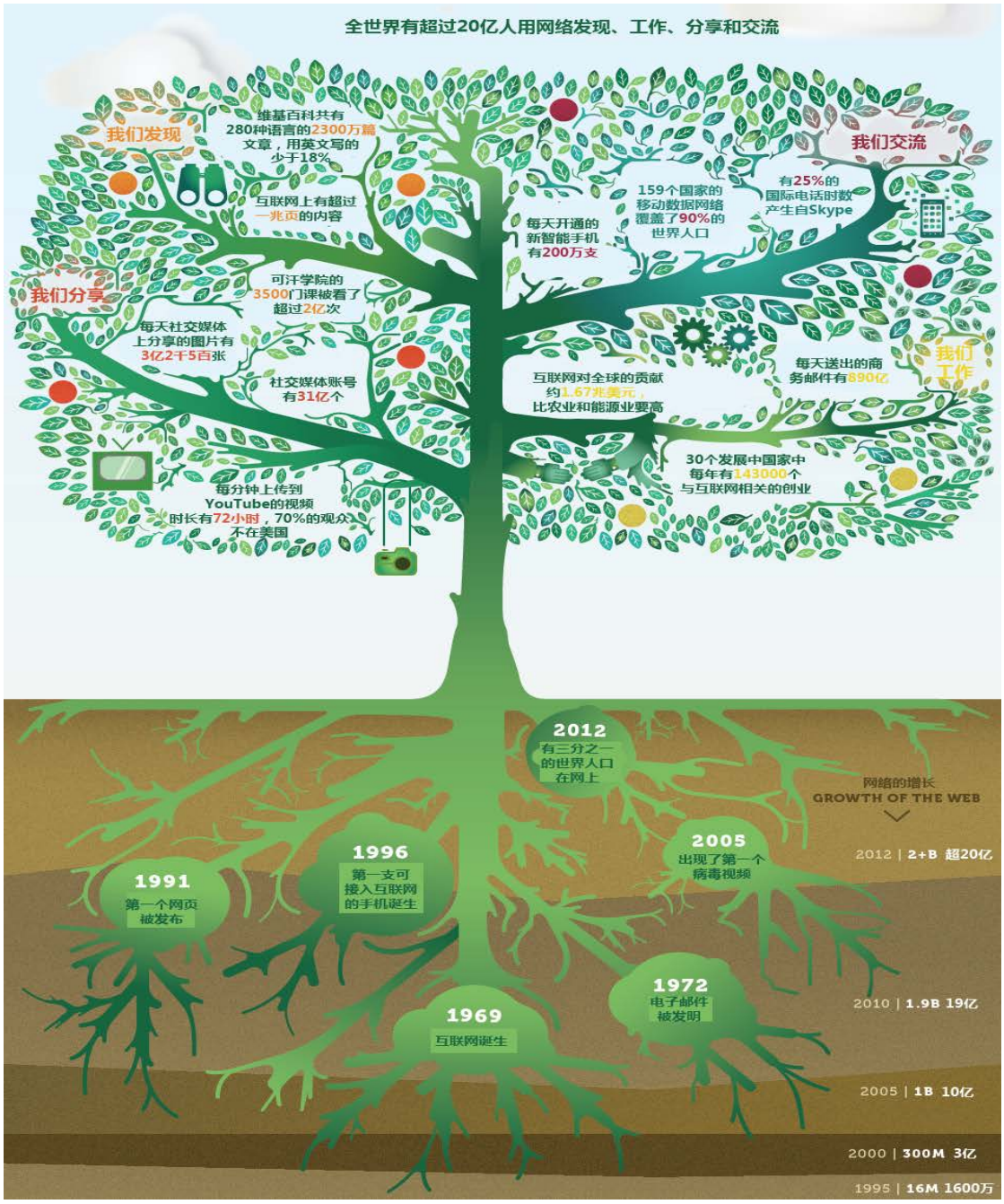

图1-3： 网络生活 [6]

➢ 云计算

云计算，再一次改变了数据的存储和访问方式。在云计算出现之前，数据大多分散存储在每个人的个人电脑、每家企业的服务器中。云计算，尤其是公用云计算，把所有的数据集中存储到“数据中心”，也即所谓的“云端”，用户通过浏

[6]来源：Google，https://www.google.com/takeaction/

览器或者专用应用程序来访问。

一些大型的网站，通过提供基于“云”的服务，积累了大量的数据，成为事实上的“数据中心”。“数据”是这些大型网站最为核心的资产，他们不惜花费高昂的费用，付出巨大的努力，来存储这些数据。谷歌公司甚至购买了单独的水力发电站，为其庞大的数据中心提供充足的电力。根据一些公开资料显示，谷歌在全球分布着大约 36 个数据中心。

近几年，国内各地兴起了建设云计算基地的风潮，客观上为“大数据”的诞生准备了必备的储存空间和访问渠道。各大银行、电信运营商、大型互联网公司、政府各个部委等都拥有各自的“数据中心”。绝大多数的银行、电信、互联网公司都已经实现了全国级的数据集中的工作。

云计算是大数据诞生的前提和必要条件。没有云计算，就会缺少数据集中采集和存储的商业基础，而云计算为大数据提供了存储空间和访问渠道；大数据则是云计算的灵魂和必然的升级方向。

2012 年业内所有的云计算大会，无论官方背景还是民间主办，都是把“大数据”作为一个核心的主题，甚至有时候都分不清楚，这是云计算的会，还是大数据的会。

- **物联网**

物联网是信息技术领域的另一个热词，遍布大街小巷的摄像头是大家可以直观感受到的一种物联网形态。物联网，究其本质是传感器技术进步的产物。在人们的生活中，传感器几乎无处不在，从监测大气的温度、压强、风力，到监测桥梁、矿井的安全，再到监测飞机、汽车的行驶状态等。大型器件，如一架军用战斗机上的传感器多达数千个；小型器件，如日常使用的智能手机，就包括重力感应器、加速度感应器、距离感应器、光线感应器、陀螺仪、电子罗盘、摄像头等诸多种类的传感器。这些不同类型的传感器，无时无刻不在产生大量的数据，其中的某些数据被持续的收集起来，成为大数据的重要来源之一。

- **社交网络**

社交网络是互联网发展史上的又一个重要的里程碑。它把人类社会真实的人际关系完美的映射到互联网空间，并借助互联网的特性而大大升华。广义地看，社交网络使得互联网甚至具备某些人类的特质，譬如“情绪”：人们分享各自的

喜怒哀乐，并相互传染和传播。社交网络为大数据带来一类最具活力的数据类型——人们的喜好和偏爱。更重要的是，在社交网络中，如何利用网民的关系链来传播喜好和偏爱，为研究消费者行为打开了一扇方便之门。如果深入地分析社交网络，就会发现，大型的社交网络平台，事实上构成了以“个人”为枢纽的不同的数据的集合。借助“分享”按钮，人们在不同网站上的购物信息、浏览的网页都可以“分享”在社交网络上。就像雪地上的脚印，社交网络把网民在不同网站上留下的“脚印”链接起来，形成完整的行为轨迹和“偏好”链。

图1-4： 反映社交网络 Facebook 上人们活跃程度的世界地图 [7]

图 1-4 是 Facebook 的一个实习生把网站中人们相互联系的数据通过建模、渲染得到的一幅图片，越是明亮的地方，人们相互交流越是活跃。现在 Facebook 是世界上最大的社交网站，每月的活跃用户数已突破 10 亿人。

➢ 智能终端普及

古人只能用“大漠孤烟直，长河落日圆”等诗词歌赋，来主观地描述他们的所见所闻，而现代人则可以掏出手机、照相机、摄像机等终端设备，再现美丽的风景，与亲朋好友分享。在迷路时，性情的古人索性信马由缰不问归路 [8]，而现代人则可以拿出智能手机，使用导航软件找寻目的地。

[7]来源：Facebook，http://www.facebook.com

[8]《晋书·阮籍传》中记载，“时率意独驾，不由径路，车迹所穷，辄恸哭而反”。籍非迷路，刻意为之。正文是夸张的说法。

智能终端不仅仅局限于个人应用，许多行业都已经开始大规模地部署终端产品。举一个“美丽”的例子，如婚纱摄影行业，以前影楼需要租用大面积的场馆、位置优良租金高昂的门店，携带大型的、笨重的写真集，展示给准新娘们用以挑选照片。而现如今利用 iPad，可以做出令人心醉神迷的实景效果，如 360 度旋转等特效。准新娘只需要一部 iPad，就可以全面地看到最终的拍摄效果，并利用其交互特性提高样片选择的精准度。

KPCB[9]（凯鹏华盈）是美国最大的风险投资基金之一，其合伙人 Mary Meeker 在 2012 年发布的一份趋势报告中指出，在 2010 年第四季度，智能手机加平板电脑的出货量已经超越台式机和传统笔记本电脑，（参见图 1-5），并且预计将在 2013 年第二季度，智能移动终端全球保有量也将实现超越。（参见图 1-6[10]）。

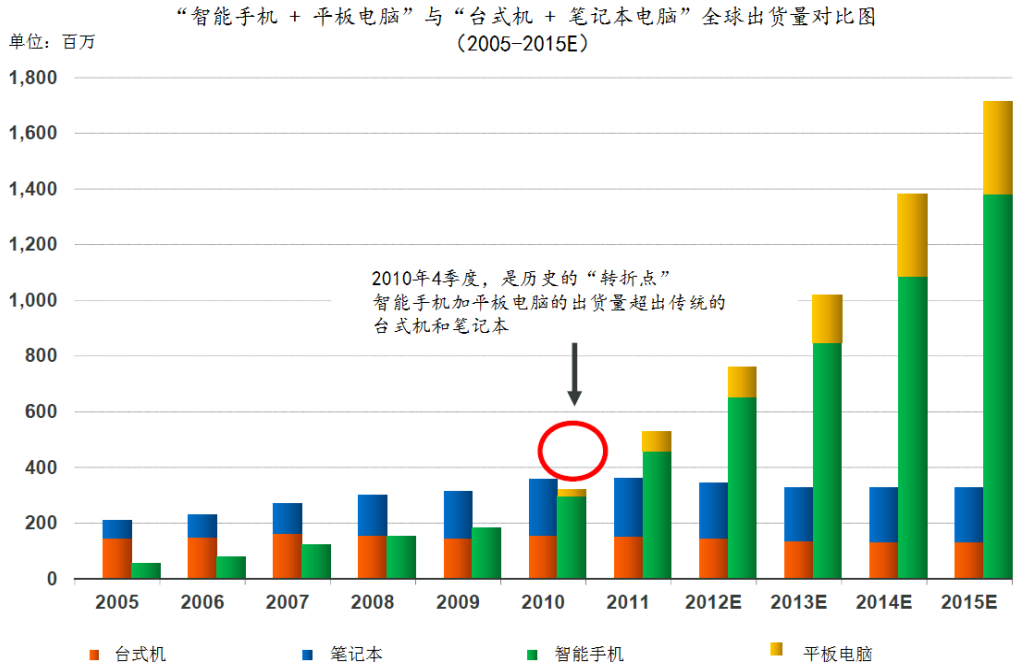

图1-5： 移动设备与传统台式机、笔记本电脑的全球出货量对比图 [11]

[9]KPCB 公司（Kleiner Perkins Caufield & Byers）成立于 1972 年，是美国最大的风险基金，主要是承担各大名校的校产投资业务。KPCB 公司人才济济，在风险投资业崭露头角，在其所投资的风险企业中，有康柏公司、太阳微系统公司、莲花公司等这些电脑及软件行业的佼佼者，随着互联网的飞速发展，公司抓住这一百年难觅的商业机遇，将风险投资的重点放在互联网产业，先后投资美国在线公司、奋扬公司（EXICITE）、亚马逊书店、网景公司、谷歌、Intuit 等公司。

[10]计算保有量，预计保有量假定台式机的换机周期是 5 年，笔记本电脑的换机周期是 4 年，智能手机 2 年，平板电脑 2.5 年。

[11]来源：Katy Huberty, Ehud Gelblum, Morgan Stanley Research. Data and Estimates as of 9/12

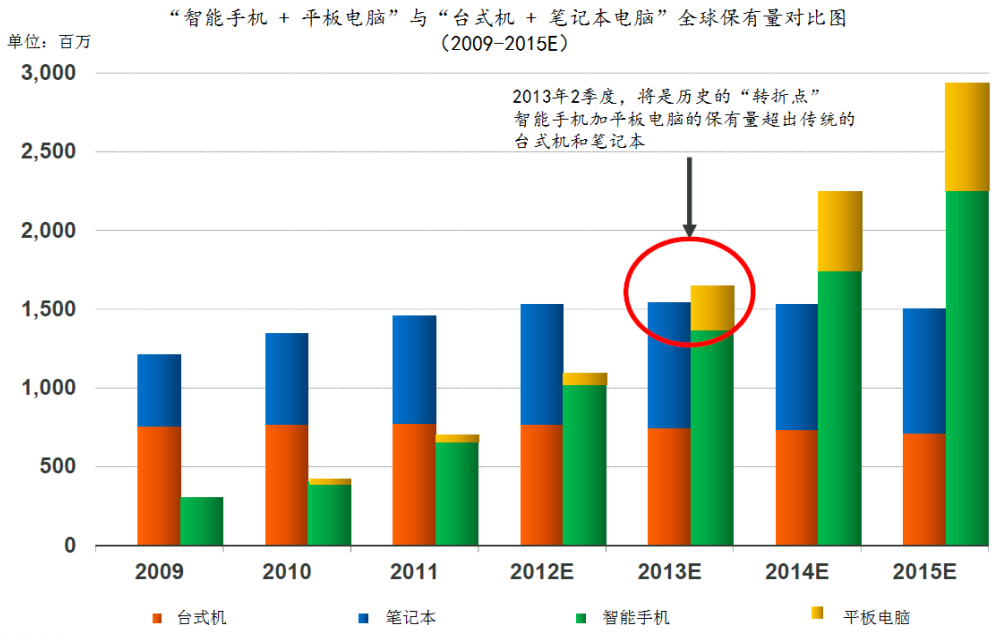

图1-6： 移动设备与传统台式机、笔记本电脑的全球保有量对比图 [12]

智能终端的普及给大数据带来了丰富、鲜活的数据。苹果公司 2012 年公布的一组运营数据，反映了智能终端上人们的活跃程度。其中，iMessage 功能目前每秒为用户传递 28000 条信息，iCloud 已经为用户提供了总计 1 亿多份的文档，GameCenter 的账号创建数达到了 1.6 亿，当前 iOS 应用总数突破 70 万，支持 iPad 的应用则达到了 27.5 万，AppStore 的应用下载量突破 350 亿次，通过分成付给应用开发商的分成总额已达 65 亿美元，iBooks 中的图书总数已达 150 万册，下载量也超过了 4 亿次。

1.2 大数据的概念和内涵

麦肯锡（美国首屈一指的咨询公司）是研究大数据的先驱。该公司在报告《大数据：创新、竞争和生产力的下一个前沿领域》中给出的定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。其同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据 [13]。

国际数据公司（IDC）从四个特征定义大数据，即海量的数据规模（Volume）、快速的数据流转和动态的数据体系（Velocity）、多样的数据类型（Variety）和巨

[12]来源：Katy Huberty, Ehud Gelblum, Morgan Stanley Research. Data and Estimates as of 9/12
[13]参见麦肯锡，《Big data: The next frontier for innovation, competition, and productivity》，2011 年。

大的数据价值（Value）。

亚马逊（全球最大的电子商务公司）大数据科学家 John Rauser 给出了大数据的简单的定义：大数据是任何超过了一台计算机处理能力的数据量。（Big data is ‘any amount of data that’s too big to be handled by one computer’.）

维基百科中则只有短短的一句话：“巨量资料(big data)，或称大数据，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯”。

大数据是一个宽泛的概念，见仁见智，但是上面几个定义都无一例外地突出了“大”字。诚然“大”是大数据的一个重要特征，但远远不是全部。《大数据时代的历史机遇》一书的作者认为：大数据是“在多样的或者大量数据中，迅速获取信息的能力 [14]”。前面几个定义都是从大数据本身出发，这个定义更关心大数据的功用，即大数据能帮助人们干什么？在这个定义中，重心是“能力”。一般而言，“大数据”是指难以在可接受的时间内，用传统数据库系统或常规应用软件处理的、巨量而复杂的数据集[1]

“大数据”与钱学森老先生提倡的“大成智慧学”的要义非常接近。钱老将“大成智慧” 翻译成“Wisdom in Cyberspace”，强调“必集大成，才能得智慧”。有了数据，有了信息，不等于就有智慧，出智慧的关键在“集”。大数据中包括的全部事实、经验、信息都是“集”的对象和内容。采集到的原始数据往往是些“零金碎玉”，没有什么逻辑，不一定能直接用现在掌握的科学技术解释，需要集成融合各个侧面的数据，才能挖掘出前人未知的大价值。每一种数据来源都有一定的局限性和片面性，事物的本质和规律隐藏在各种原始数据的相互关联之中。只有融合、集成各方面的原始数据（带毛的数据），才能反映事物的全貌。开展大数据研究和应用，切忌“瞎子摸象”、“坐井观天”，一定要大协作，大集成。

大数据不仅仅是一种工具，而是一种战略、世界观和文化，要大力推广和树立“数据文化”。智慧来源于数据而不是主观臆断，要提倡用数据说话，少犯官僚主义、形式主义、主观主义和经验主义的错误。从这种意义上讲，推动“大数据”技术的应用也是贯彻中央精神，破除“四风”的有力抓手。

大数据产业的生产活动涵盖数据的获取、整理（curation）、存储、处理、可

[14]参见《大数据时代的历史机遇》清华出版社 2013.07.

视化、应用服务和信息共享等，其业务模式包括网络数据与信息服务、企业和政府智能化管理决策、企业流程改造与变革等，应用领域涉及信息服务、智慧城市、金融、制造业、国家安全和科学研究等，几乎渗透到国民经济的所有部门。

目前能产生经济利益的数据主要是网上数据，即“在线数据”，写在纸上的数据很难被快速挖掘出大的价值。互联网领域是大数据的标杆应用领域。众多互联网服务厂商让上亿用户免费为其打工，其海量的原始数据和用户行为数据来自于用户的信息消费过程，通过网络爬虫和用户点击日志获取等技术手段，信息消费和信息获取、分析已融为一个整体。

虽然有些学者认为关系数据库和事务处理不能算作大数据，但Forrester公司的调查统计表明（见图1-7）：目前大数据实际应用最多的（占被调查企业的72%）是公司的事务处理数据，而视频图像数据（13%）和科学数据（12%）还不是大数据应用的主流。大数据分析可以更全面地了解客户偏好和需求，通过这种深入的了解，各类企业均可以从中受益。因此在发展大数据产业时需要高度重视企业的事务处理的智能化，引导企业从传统的小型机和关系数据库走向新的大数据处理平台。

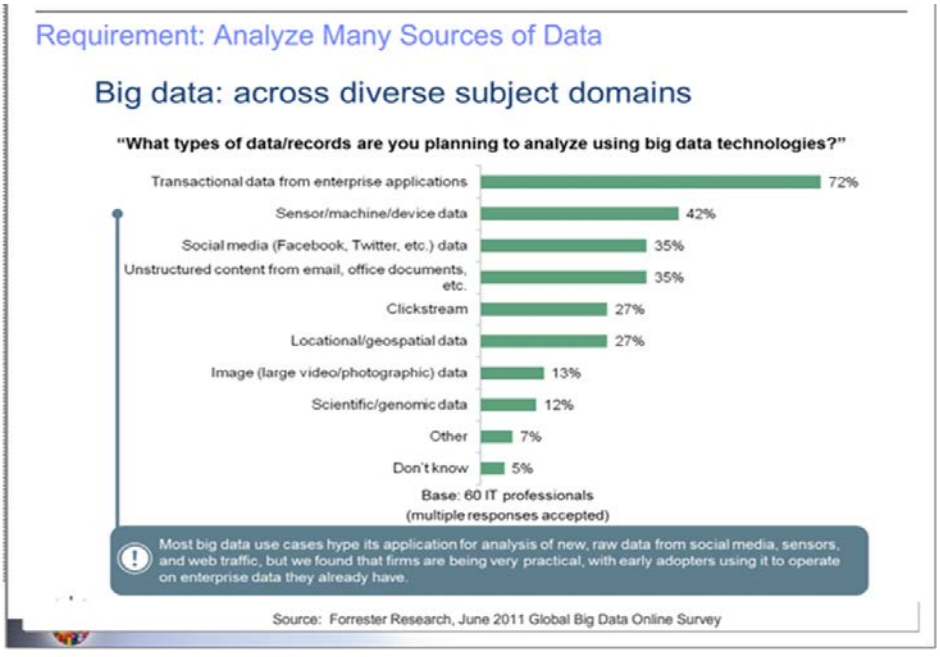

图1-7： 企业对大数据技术的需求统计

在发展大数据产业过程中，还需要关注大数据对制造业、材料、化工、制药等传统产业的变革性影响。未来对经济影响较大的可能是“数据材料”、“数据

化学”、“数据药物”等新产业，需要重视“材料基因组学”、“化学基因组学”、“药物基因组学”等的研究。

国内外软硬件厂商都将大数据处理作为重要的新兴应用负载，研究开发新产品，大力提升大数据能力。应用软件厂商（如SAS和SAP公司）已推出支持大数据的新产品，甲骨文、IBM、曙光等系统厂商正在推出支持高效大数据处理的一体化服务器。设备和软件厂商必将是发展大数据产业的主要驱动力之一。

1.3 大数据的发展历程

麦肯锡于 2011 年 5 月发布的《大数据：创新、竞争和生产力的下一个前沿领域》报告将大数据概念从技术圈引入企业界。国金证券[15]率先将大数据概念引入中国资本市场，连续推出三篇报告，令资本市场沸腾。巧合的是，美国政府在国金证券大数据研究报告发布不久就推出了《大数据研究发展计划》[16]，将大数据上升至国家战略层面，形成国家意志。之后，Splunk 成为在美国成功上市的首家大数据公司，让“数据人”一时扬眉吐气，深感数据工作的春天到了。

正如哈佛大学量化社会科学学院院长 Gary King 所说：“这是一种革命，我们确实正在进行这场革命，庞大的新数据来源所带来的量化转变将在学术界、企业界和政界中迅速蔓延开来，没有哪个领域不会受到影响。”毫无疑问，上述的种种事件无不向世界传递一个讯息：大数据时代已经到来！

麦肯锡的研究报告指出全球数据正在呈爆炸式增长，数据已经渗透到每一个行业和业务职能领域，并成为重要的生产因素。大数据的使用将成为企业成长和竞争的关键，人们对大数据的运用将支撑新一波的生产力增长和消费者收益浪潮。

麦肯锡的报告深入研究了美国医疗卫生、欧洲公共管理部门、美国零售业、全球制造业和个人地理信息等五大领域，用具体量化的方式分析研究大数据所蕴含的巨大价值。大数据的合理有效利用，为美国医疗卫生行业每年创造价值逾 3000 亿美元，为欧洲公共管理部门每年创造 2500 亿欧元（约 3500 亿美元），为全球个人位置服务的服务商和最终用户分别创造至少 1000 亿美元的收入和 7000 亿美元的价值，帮助美国零售业获得 60%的净利润增长，帮助制造业在产品开发、组装方面将成本降低 50%。

[15]http://www.gjzq.com.cn
[16]http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

通过对上述五大领域的重点分析，麦肯锡提出了五种可以广泛适用的利用“大数据”的方法：

（1）创造透明度，使利益相关者更容易及时获取大数据将产生的巨大价值。

（2）启用实验来发现需求，呈现可变性，提高性能。数据驱动的组织在已有经验成果的基础上做出决定，这种方法的好处已经被证实。

（3）细分人群，采取灵活行动。随着技术的进步，可以接近实时地进行细分，并通过更精确的服务满足客户需求。

（4）使用自动化算法代替或辅助人类决策，基于大数据的深入分析可以大幅降低决策风险，提高决策水平。

（5）创新商业、产品和服务，大数据使各类企业拥有了改善和创新现有的产品和服务的机会，甚至建立全新的商业模式。

此外，麦肯锡在报告中也指出了在挖据大数据潜能时所面临的各种挑战，包括隐私、安全、人才、技术等。

麦肯锡的报告充分肯定了大数据蕴藏的巨大价值，并试图帮助不同地域、不同部门的领导者及政策制定者了解如何利用大数据的潜在价值。整篇报告为大数据时代的蓬勃发展拉开了序幕。

表1-1： 大数据发展大事记

时间	大数据事件	里程碑
2011 年 5 月	麦肯锡全球研究报告《Big data: The next frontier for innovation, competition,and productivity》。	首开先河
2011 年 5 月	EMC World 2011 在拉斯维加斯开幕，会议主题为“云计算适逢大数据”，参会者超过 10000 人,现场有超过 500 场讲座，以及来自上百家领先 IT 厂商的上百个动手实验室和展示。EMC 公司董事长兼首席执行官乔图斯先生发表主题演讲为四天的大会开幕，他着重介绍了云计算和大数据给 IT 带来的变革。同期举办 Momentum 大会（企业内容管理大会）、数据科学家峰会（Data Scientist	

	Summit）、大数据存储峰会和 CIO 峰会。	
2011 年 5 月	IBM 推出的大数据分析软件平台 InfoSphere BigInsights 和 Streams 是目前业内最先推出的针对大数据分析的产品。两款产品将包括 Hadoop MapReduce 在内的开源技术紧密地与 IBM 的系统集成起来。研究 Hadoop 开源技术的人很多，但是 IBM 是真正将其变成了企业级的应用。	
2011 年 7 月	Yahoo 宣布成立新公司 Hortonworks 接手 Hadoop 服务，Hadoop 也迎来了新的发展机会。针对大数据领域，有很多技术提供商参与了 Yahoo 的项目。Apache Hadoop 是一个开源项目，Yahoo 是其中最大的贡献者；MapReduce 是 Hadoop 架构的一个主要组件，开发出的软件可以用来分析大数据集，它在目前的火爆程度已经无需赘言；Cloudera 是 Hadoop 最早的技术支持、服务和软件提供商，它今后将直接与 Yahoo 的 Hortonworks 展开竞争。此外，EMC 还推出了付费的基于 MapR Technologies 公司技术的 Hadoop 产品。	
2011 年 8 月	微软宣布推出了两个基于 Hadoop 的大数据处理的社区技术预览版连接器组件，一个用于 SQL Server，另一个用于 SQL Server 并行数据仓库（PDW）。该连接器是一个部署在 Linux 环境中的命令行工具。SQL Server Hadoop 连接器是在微软大数据之路上最重要的一步。另外，微软还宣布将推出 LINQ Pack、LINQ to HPC、Project “Daytona”以及 Excel DataScope，这些产品都将专为研究人员和业务分析师打造，用以在 Windows Azure 上做大数据分析。	

2011 年 10 月	甲骨文宣布收购为企业用户提供非结构化数据管理、网络商务和商务智能技术的企业搜索和数据管理公司 Endeca Technologies。	
2011 年 12 月	中国资本市场发布第一篇大数据主题研究报告《大数据时代即将到来》。	掀起资本市场热潮
2012 年 1 月	中国资本市场发布第二篇大数据主题研究报告《大数据时代三大发展趋势和投资方向》。	提出系统的大数据认知框架
2012 年 3 月	IDC 发布大数据（Big Data）市场预测报告，预估该领域的市场规模将从 2010 年的 32 亿美元成长到 2015 年的 169 亿美元，每年的平均成长率接近 40%。	
2012 年 3 月	亚马逊 CTO Werner Vogels 在 Cebit 上发表的主题演讲“无限的数据”称，企业在思考大数据的时候，需要注意的不仅是需要分析大量的数据，还包括信息的存储方式。此外，他鼓励企业思考大容量图片的问题，他还介绍了用于实施大数据系统的亚马逊云蓝图。	
2012 年 3 月	美国奥巴马政府宣布推出“大数据研究和发展计划”。该计划涉及美国国家科学基金、美国国家卫生研究院、美国能源部、美国国防部、美国国防部高级研究计划局、美国地质勘探局等 6 个联邦政府部门，承诺将投资两亿多美元，大力推动和改善与大数据相关的收集、组织和分析工具及技术，以推进从大量的、复杂的数据集合中获取知识和洞见的能力。美国奥巴马政府宣布投资大数据领域，是大数据从商业行为上升到国家战略的分水岭，表明大数据正式提升到战略层面，大数据在经济社会各个层面、各个领域都开始受到	标志大数据上升为国家战略，体现国家意志

	重视。	
2012 年 4 月	中国资本市场发布第三篇大数据主题研究报告《以数据资产为核心的商业模式》。	系统阐述大数据商业图景
2012 年 4 月	SAP 计划斥资近 5 亿美元来吸引用户使用其 Hana 数据处理产品，从而加大与甲骨文之间的竞争。Hana 平台的设计目的是迅速分析海量的销售和运营信息，以及对电子邮件和社交媒体等非结构化数据进行分析，依靠计算机存储器而非磁盘驱动器来加速这一程序。	
2012 年 4 月	企业数据软件公司 Splunk 以每股 17 美元的价格在纳斯达克进行 IPO，融资 2.3 亿美元，首个交易日市值突破惊人的 30 亿美元。	首家上市的大数据公司
2012 年 4 月	谷歌正式推出在线存储服务 Google Drive。	
2012 年 5 月	Google 推出的一项企业级大数据分析的云服务 BigQuery，用来在云端处理大数据。BigQuery 将有助于企业在没有硬件基础设施的情况下分析他们的数据。同时可以建立应用程序和数据共享的所有服务。	
2012 年 5 月	IDC 发布研究报告指出，“大数据”概念正在引领中国互联网行业新一轮的技术浪潮，截至 2011 年底，中国互联网行业持有的数据总量已达到 1.9EB(1EB 艾字节相当于 10 亿 GB)。IDC 预计，这一规模到 2015 年将增长到 8.2EB 以上。	
2012 年 6 月	IDC 发布研究报告《中国互联网市场洞见：互联网大数据技术创新研究，2012》，对中国互联网行业围绕大数据的技术创新进行了专题研究。报告指出，大数据正在引领中国互联网行业新一轮的技术浪潮。	

2012 年 6 月	为推动大数据（Big Data）这个交叉学科的发展，推动学术、应用和产业的发展，中国计算机学会决定成立“CCF 大数据专家委员会”（暂定），并责成 CCF 名誉理事长、中国工程院院士李国杰作为牵头人，开展有关工作。	
2012 年 7 月	联合国在纽约发布了一份关于大数据政务的白皮书《大数据促发展：挑战与机遇》，总结了各国政府如何利用大数据更好地服务和保护人民。	
2012 年 9 月	北京拓尔思信息技术股份有限公司联手华为技术有限公司倾力推出拓尔思-华为大数据一体机系列。拓尔思-华为大数据一体机系列包括拓尔思-华为信息采集一体机、拓尔思-华为检索一体机，后续还会有相应的大数据一体机问世。	
2012 年 10 月	中国通信学会大数据专家委员会成立大会暨首届大数据论坛在北京召开。会上，大数据的飞速发展以及给我国相关产业带来的机遇和挑战成为与会专家讨论的焦点。	
2012 年 10 月	市场研究公司 Gartner 发布研究报告称，大数据产业今年将在全球范围内带来近千亿美元的 IT 开支。Gartner 在报告中预测，今年，大数据对全球 IT 开支的直接或间接推动将达 960 亿美元；到 2016 年，这一数字预计将达到 2320 亿美元。	
2012 年 10 月	IBM 和牛津大学联合发布了一份大数据研究报告，研究包括：大数据的实际使用情况；创新型企业如何从不确定数据中提取有价值数据。	
2012 年 11 月	淘宝和天猫今年的交易总额在 11 月 30 日突破 1 万亿人民币，为支撑这巨大规模业务量的直接间接就业人员，已经超过 1000 万人。	

2012 年 11 月	首届数据科学与信息产业大会在北京国际数学研究中心召开。标志学术界、产业界、资本市场形成共识。	数据科学登上产业舞台

1.4 大数据的热点问题

数据是与自然资源、人力资源一样重要的战略资源，掌控数据资源的能力是国家数字主权的体现。大数据研究和应用是现有产业升级与新产业崛起的重要推动力量，如果落后就意味着失守战略性新兴产业的制高点。大数据也正在引发科学思维与研究方法的一场革命。

大数据将颠覆过去的商业思维，未来企业核心竞争力，主要不是资金，也不是现有市场规模，而是对大数据的掌控分析能力。大数据对产业生态环境的颠覆基于以下三大趋势：软件的价值同它所管理的数据的规模和活性成正比；越靠近最终用户的企业，将在产业链中拥有越大的发言权；数据将成为核心资产。

大数据不仅来源于谷歌、百度等厂商，工业领域也在源源不断地产生数据，其规模可能比网络服务厂商还大。飞机汽轮机压缩器叶片的监控数据为588GB/天，是世界最大的微博公司（Twitter）每天产生数据（80GB）的7倍。制造业是数据分析的广阔天地，应充分挖掘工业领域大数据的价值。

“数据量大”是存储、分析大数据的一个难关，但不是最大的挑战。比数据量大更难应对的是数据的多样性、实时性和不确定性。而判断一个数据集是否有价值也是很困难的事，也许今天认为没有价值的数据将来会找到很大的价值。因此，我们应关注的并不是PB级或EB级的数据，而是从巨量模态多样、真伪难辨的数据中及时获得价值的“能力”。

大数据与智慧产业有本质的联系，智慧城市的关键技术就是大数据的获取与分析。只有“让数据说话”，才能做出明智的决策。企业的大数据应用正是从过去的“商业智能(Business Intelligence)”发展起来的。大数据技术本质上是机器学习等人工智能技术的深化与推广。

新技术发展之初有一个相当长的酝酿期。美国国家研究理事会对信息领域的统计结果表明，一个产业从基础研究到形成100亿美元的市场，一般需要20-40年。

新兴产业的成长曲线是一个开口向上发展的指数曲线，开始的2-5年发展并不是很快，产业的规模常常低于人们的预期。政府官员和企业界往往对新兴产业的短期成效估计过高，而对其长期发展的潜力又估计不足。我们应改变短视的习惯思维，着眼长远发展。

采集和利用大数据是一个持续发展、不断升级的过程，“大数据”从“小数据”开始。目前大多数单位还是处于“小数据”处理阶段，但只要在纵向上有一定的时间积累，横向上有丰富的记录细节，做仔细的数据分析，就可能产生大的价值。在实际工作中，我们不要太在意“大数据”和“小数据”的区别，不必花精力对大数据的定义做无谓的争论。不管“大数据”、“小数据”，能挖掘出价值就是好数据。

数据获取和分析是人类几千年来从未间断的活动，今后也会永远持续下去。目前大家热议的“大数据”只是漫长旅行中一个里程碑，标志着人类对数据的重视进入一个新阶段。预计到 2015 年，大数据的“炒作”将进入了低谷幻灭期，开始进入良性平稳发展的阶段。

1.5 各国大数据发展战略

历史上曾经上演过类似的一幕。1993 年美国“信息高速公路”计划一出台就引起了各国的强烈反应。日本政府在 1993 年 6 月发布拟建大规模超高速“研究信息流通新干线”计划，决心通过高速通信线路将全国研究机构与大学连接起来，并于 1994 年 5 月前后提出了日本版“信息高速公路”计划——《通信基础结构计划》和《通向 21 世纪智能化创新社会的改革》两个报告，决定分成三个阶段逐步实施网络建设。欧洲也不甘落后，在 1993 年 6 月哥本哈根欧盟首脑会议上，欧盟主席德洛尔首次提出“构建欧洲信息社会”的倡议，随后在同年 12 月，欧盟发布了旨在“振兴经济，提高竞争能力和创造就业机会”的白皮书。白皮书中明确提出构建欧洲版“信息高速公路”的设想，并成立了专门的工作小组负责计划的推进。此外，加拿大、韩国、新加坡等发达国家为争夺高新技术的发展优势，迎接 21 世纪的发展挑战，也都纷纷选择立即跟进，投入巨额资金，推出各自国家的“信息高速公路”计划，在全球范围内掀起一场热潮。

➢ 美国的动向

美国奥巴马政府在 2012 年 3 月正式启动“大数据研究和发展计划”，该计划涉及美国国防部、美国国防部高级研究计划局、美国能源部、美国国家卫生研究院、美国国家科学基金、美国地质勘探局等 6 个联邦政府部门，宣布将投资 2 亿多美元，用以大力推进大数据的收集、访问、组织和开发利用等相关技术的发展，进而大幅提高从海量复杂的数据中提炼信息和获取知识的能力与水平。该计划并不是单单依靠政府，而是与产业界、学术界以及非营利组织一起，共同充分利用大数据所创造的机会。这也是继 1993 年 9 月美国政府启动“信息高速公路”计划后，国家层面发力在信息领域的又一次“狂飙猛进”。

图1-8：美国从“信息高速公路计划”过渡到“大数据发展计划”

随着网络技术、计算机技术以及通信技术的快速发展，人类社会的数据总量呈现指数级增长，根据 Google 前 CEO、现董事会主席埃里克•施密特的说法，截止至 2003 年，人类社会总共创造了 5EB 的数据，而现在仅需要两天就能创造相同的数据量，这对数据的收集、运输、存储、分析利用和安全等技术应用和数据的管理工作提出了更高的要求和更大的挑战。信息作为三大社会资源之一（另外两个是物质和能量），如何充分利用信息资源、如何更经济更快速地从海量、不

同结构类型的复杂数据中快速提取价值成为关键。

美国前总统克林顿开展的“信息高速公路”计划是通过高速率的通讯网络搭建人们的信息交流网络，进而带动经济的快速发展，该计划促使海量数据的产生，却未能实现对数据资源进行充分利用，尤其是在大数据时代的今天，海量数据具有的巨大价值被白白浪费。根据 2011 年美国总统科学技术顾问委员会提出的一份建议显示，大数据相关技术具有重要战略价值，但美国联邦政府对其研发投资却明显不足。而通过“大数据研究和发展”计划可以深度挖掘大数据的潜在巨大价值，带动产业的升级换代。从这个角度说，“大数据研究和发展”与“信息高速公路”计划可谓一脉相承，层层推进。

- **欧盟开放数据平台——Open Data Portal**

欧盟委员会全新的开放数据平台（以下简称为 ODP）Beta 版已经向公众开放（http://open-data.europa.eu/open-data），和美国政府的数据开放平台类似，致力于推动开放、透明的政府，促进创新。

2010 年 4 月欧盟委员会发起欧洲数字化议程，致力于利用数字技术刺激欧洲经济增长，帮助公众和企业最大化利用数字技术。ODP 是欧洲数字化议程的一部分，欧盟委员会副主席 Neelie Kroes 说：“这将打开一个金矿，通过这个系统，公众获得这些数据会更便捷，成本更低，获得的数据内容更广泛”。

截至 2013 年 1 月 12 日，ODP 已经开放 5815 个数据集，其中的 5638 个数据集来自欧盟统计局 Eurostat，数据包括地理、大气、国际贸易、农业等各类信息。

ODP 提供的不仅是数据，还建立了数据的统一语法规则，保证数据发布机构、公众、应用开发者都能够利用这些数据，任何人都可以在这里下载数据，利用这些数据开发新的应用。和美国政府的数据开放平台一样，ODP 开放最原始的、粒度最小的、未经过加工的数据，保证数据的真实性，让公众各取所需，各尽其用。目前数据提供 dft、sdmx 和 tsv 三种标准格式供下载使用。

表 1-2 是 2003-2011 年关于欧盟成员国内陆货运水路的数据。根据 2006 年 9 月 6 日欧洲议会通过的 1365/2006 号欧共体条例收集。这些数据是了解欧盟成员国货运情况的基础。

表1-2：欧盟国家内陆货运水路（单位：公里）[17]

国家	2003	2004	2005	2006	2007	2008	2009	2010	2011
保加利亚	4316	4259	4154	4146	4143	4144	4150	4098	4072
芬兰	5851	5741	5732	5905	5899	5919	5919	5919	5944
意大利	15965	15916	16225	16295	16335	16529	16686	16704	16726
立陶宛	1774	1782	1771	1771	1766	1765	1767.6	1768	1768
拉脱维亚	2270	2270	2270	2269	2265	2263	1884	1897	1865
荷兰	2811	2811	2810	2797	2801	2888	2896	3013	3013
波兰	19900	20250	20253	20176	20107	20196	20360	20228	20228
罗马尼亚	11077	11053	10948	10789	10777	10785	10784	10785	10777
斯洛伐克	3657	3660	3658	3658	3629	3623	3623	3622	3624
……	……	……	……	……	……	……	……	……	……

➢ 日本重启 ICT 战略计划

美国推出“大数据研究和发展”计划之后，日本政府重新启动曾在日本大地震后一度搁置的 ICT 战略研究，将重点关注大数据应用。

2012 年 5 月，日本总务省信息通信政策审议会下设的 ICT 基本战略委员会召开会议，战略委员会大数据研究主任、京东大学教授森川博之强调，美国在大数据技术上处于领先，其 Google、Amazon 等网络企业在大数据应用领域方面有很强的优势，日本非常有必要在大数据方面制定综合性的战略。随后，日本文部科学省在 7 月发布了以学术云为主题的讨论会报告。文部科学省指出为迎接大数据时代学术界面临的挑战，将重点推进大数据收集、存储、分析、可视化、建模、信息综合的各阶段研究，构建大数据利用的模型。

[17]来源： http://open-data.europa.eu/open-data

➢ 联合国发布大数据白皮书

联合国也随后发布了《大数据促发展：挑战与机遇》的白皮书，全球范围内对大数据的关注达到了前所未有的热度。

2012 年 7 月，联合国在发布的白皮书中指出大数据时代已经到来，大数据对于联合国和各国政府都是一次历史性的机遇。报告讨论了如何利用大量丰富的数据资源帮助政府更好地响应社会需求，指导经济运行，并建议联合国成员国建设“脉搏实验室”（Pulse Labs），挖掘大数据的潜在价值。印度尼西亚在首都雅加达建立的“脉搏实验室”由澳大利亚提供资助，于 2012 年 9 月投入运行；此外，乌干达也率先在首都坎贝拉建立了“脉搏实验室”。

第二章 大数据典型应用现状

大数据将给各行各业带来变革性机会，但真正的大数据运用仍处于发展初级阶段。据《证券日报》市场研究中心统计分析发现，目前我国在公共领域对大数据的运用主要集中在电力行业、智能交通、电子政务、司法系统等四个方面。**电力行业：**大数据对该行业的应用主要体现在智能电网上，通过获取人们的用电行为信息，智能电网能够实现优化电的生产、分配以及消耗，有利于电网安全检测与控制（包括大灾难预警与处理、供电与电力调度决策支持和更准确的用电量预测）、客户用电行为分析与客户细分，电力企业精细化运营管理等多方面，实现更科学的电力需求管理。**智能交通：**交通运输部今年 7 月份下发通知，将对公共交通信息化应用系统建设、相关支撑系统建设、数据资源与交换系统建设提供资金支持。**电子政务：**通过政府信息化，大数据能够提高政府决策的科学性和精准性，提高政府预测预警能力以及应急响应能力，节约决策的成本。以财政部门为例，基于云计算、大数据技术，财政部门可以按需掌握各个部门的数据，并对数据进行分析，做出的决策可以更准确、更高效。另外，也可以依据数据推动财政创新，使财政工作更有效率、更加开放、更加透明。**司法系统：**公安市场大规模的信息化和装备投资产生了海量的非结构化数据，公安的实战应用是大数据的重要应用领域。

2.1 互联网与大数据

互联网作为一个数据平台、一个数据集散地，聚集了海量的数据，完全可以借助新的大数据理论和技术，分析其中蕴含的丰富内容、发现其中存在的统计规律，以便为互联网提供更好的服务和应用、为互联网行业今后实现更好更快的持续发展提供定量化的依据。

根据 2013 年 7 月发布的最新一期《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、新技术大有用之地，将助推互联网服务和应用得到更好发展，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。

大数据在互联网领域的应用现状以及未来发展：

（1）电子商务：近年来，淘宝、京东等网络零售第三方交易平台和电子商务网站的蓬勃发展，使其上聚集了大量的经营者、消费者和商品、服务，并因此而衍生出了大量的数据，利用大数据理论和技术，对网络购物、网络消费、网络团购、网上支付等数据进行深度挖掘、深入分析，将可发现大量有价值的信息与统计规律，对布局和推动今后中国互联网经济的健康有序发展、对进一步规范经营者和消费者的电子商务活动、加强国家对该领域的宏观调控和监管等，均将产生积极的影响。

当前的电子商务平台主要面对两类用户，一是最终消费者，二是大量的商家。对于最终消费者而言，电子商务平台目前主要通过积累和挖掘用户消费过程的行为数据，来为消费者提供商品推荐服务。某些电子商务平台还将时间、地理位置、社交网络等因素融入到用户行为数据中，进一步进行精准推荐。在实际的推荐系统中，主要利用的是机器学习、自然语言理解、大数据分布式存储和并行处理等技术；然而，目前针对第二类用户——商家的大数据分析挖掘服务还较少。一方面，相对于最终消费者，商家更注重数据的隐私性，对于某些数据他们是不愿意被第三方获知的；另一方面，商家的许多商业行为并不都是在线上完成的，有很多是在线下完成的，平台难以获得较为全面的数据。促使商家开放数据或者部分数据，需要在数据安全、数据使用的商业模式和技术等多个层面的创新以及观念的改变才能实现，还需要一定的时间。然而，针对商家数据的分析的确具有很高的价值。例如，通过对商家进货、库存、销售、客户关系等多方位数据的获取和分析，可以有效地为商家推荐优质的上下游业务，帮助商家建立起上下游的产业链关系；可以通过平台数据的分析为商家推送有关税收、融资、法律等与企业经营活动相关的专业服务，帮助商家更好的发展，帮助政府更好对企业进行监管和扶持。目前，国内专注于企业领域的一些公司正在大力开展这方面的工作。

（2）网络广告：利用大数据理论和技术，可深入分析网络广告的效果及其对商品销售等的影响、广告“读者”对之的反应等。

（3）网络新闻、搜索引擎：利用大数据理论和技术，通过对网民阅读/搜索内容、习惯、爱好、行为、关键词等的深入分析，可为新闻门户网站的建设、搜

索引擎技术的改进、互联网舆情的监控与引导等提供依据。

（4）旅行预订：网上预订旅行产品、旅行行程、车票机票等，已成为一项非常重要的互联网服务和应用，并因此聚集了大量的有关游客/乘客、景区/景点、宾馆/饭店等的数据，利用大数据理论和技术对此做深入、精细分析，可为更好地布局和推动我国旅游经济和假日经济的发展、更好地为游客提供旅游产品和旅游服务、更好地建设景区和景点等提供参考和依据。

（5）即时通信、社交网络、博客微博：即时通信、社交网络、博客微博成为互联网时代民众新的通信、社交和发表见解手段，利用大数据理论和技术对此进行深入分析，可更好地发现民众新的交往习惯与方式、民众关注的社会问题与社会热点、民情民意，为改善互联网时代的通信和社交服务、更好地体察民情和改进社会管理等提供参考。

（6）网络视频、网络音乐、网络游戏：网络视频、网络音乐、网络游戏等为互联网时代的民众带来了新的娱乐形式，带动了新的经济增长，当然也带来了网瘾、网络安全等问题，利用大数据理论和技术对此进行深入分析，可更好地发现民众新的娱乐形式和爱好、掌握青少年网游习性和规律，为更好地推出网络娱乐和网络游戏产品与服务、推动网游经济发展、保障青少年上网安全等提供依据。

大数据分析在互联网上的一个重要应用就是基于用户的各种海量在线行为来分析用户的兴趣和需求。对于网络游戏来说，越来越多的游戏厂商也意识到了大数据分析的重要性，特别是对于游戏的研发和运营中的三个重要环节的作用，即降低用户获取成本，提高用户留存和提高用户付费率和付费额。他们开始建立实时大数据平台收集用户在游戏中的行为数据，通过分析理解每个用户如何玩游戏、他们的动机和潜在的价值，来调整游戏的设计，并对这些用户进行实时自动的营销，以更好的满足这些用户的需求。例如基于游戏内用户行为利用数据挖掘和机器学习算法对每个用户进行评估和分类，然后可以使用这些细分的用户类别，推送及时、相关和个性化的消息（如促销信息）来留住用户。同时基于行为数据对用户细分后，还可以进行跨游戏的用户营销，对不同的用户类型推送不同类型的游戏。

面向游戏的大数据分析仍有三个挑战，第一关于数据质量，不同的游戏之间或者不同玩家的数据的预处理面临的问题是，接口不规范、杂乱无章导致数据比

较差，如何能够选出高质量的数据。第二个问题在用户的隐私和个性化之间找到一个平衡点，这对整个互联网上的用户行为分析来说都是有挑战的问题，这个挑战不光是技术，还有政策法规。第三个问题是未来跨设备、跨平台、跨应用的手机游戏，网页游戏和电视游戏将为用户提供更加无缝的娱乐体验，如何收集用户的完整的行为数据以了解他们的需求将是挑战性的任务，同时如何将大数据中的预测性分析技术应用于游戏分析，提供更加个性化的游戏成为未来的另一个方向。

总之，通过对新兴的大数据理论和技术对互联网应用的分析，能够掌握行业现状、发现潜在问题、谋划未来发展，推动互联网和大数据这两大新兴领域的结合、互动，推动二者的共同繁荣。

2.2 网络通信与大数据

对于“大数据”时代的到来，运营商普遍认为：随着信息成为企业战略资产，市场竞争要求越来越多的数据被长期保存，每天都会从管道、业务平台、支撑系统中产生大量有价值的数据，基于这些数据的商业智能应用将为运营商带来巨大的机遇。

根据 GSMA 预测，2012 年~2018 年，全球移动数据流量将以每年 50%的复合增长率增长。到 2018 年，全球移动数据流量将比 2012 年增加 12 倍。中国的发展更为迅速，2011 年全国移动数据流量为 5.77 亿 GB，预计到 2013 年底将达到 14.13 亿 GB。在 2012 年一年时间，手机的数据流量同比增长 119%，但流量的爆炸式增长也给运营商带来了前所未有的机遇与挑战。流量收入成为运营商最主要的新增长点，而语音则出现逐步下滑的局面。根据计世资讯预测，未来三年，中国电信业大数据应用市场将保持快速增长势头，到 2015 年，电信业大数据应用市场规模预计将达到 18.3 亿元。

中国联通、中国移动、中国电信 3 大运营商加速推进大数据应用的具体举措见下表。

表2-1： 推动大数据应用的举措

中国联通	2012 年底，中国联通已经成功将大数据和 Hadoop 技术引入到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 亿投资重庆大数据计划，显现了其发展大数据，转型自身业务的决心。
中国移动	中国移动在大云 1.5 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，并在英特尔至强＋Hadoop 平台上运行，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做了充分准备。
中国电信	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。在“流量经营”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，并发掘移动互联与之结合的商机，重塑转型之路。

目前，中国移动企业信息化系列产品已经得到 270 万家企业客户的认可，广泛应用于金融、交通、物流、IT、制造等领域，随着企业信息化业务增长快速，中国移动充分利用自身在数据的获取、存储、分析等众多技术与应用的集合等方面的优势，为企业客户提供更为丰富和有针对性的信息化产品和解决方案。除了电话会议、视频会议、专网专线服务、无线宽带接入、集团 V 网、IDC 数据中心等基础通信服务外，移动办公、会议助理、企业一卡通、移动财务等办公管理服务，移动 400、商户管家、移动 CRM 等营销服务，M2M 应用、视频监控、车务通等生产控制服务。仅以融合通信业务为例，这一新商用业务，截至当年 10 月收入达 3.21 亿元，客户覆盖了政府、教育、金融、电力、制造、公安、酒店等重点行业。

2012 年中国联通成功将大数据和 Hadoop 技术引入到“用户上网记录集中查询与分析支撑系统”，并已经部署了 4.5PB 的存储空间。系统已经具备了每天处理 700 亿条上网记录的能力，每天新增数据量达 20 多个 TB，每年正以 70%的速度在递增。通过该大数据项目，联通在全球运营商中率先提供了用户上网记录的清单查询服务，为移动互联网时代移动上网流量的明明白白消费提供了技术上的保证。同时，也为中国联通的移动互联网业务精细化运营、流量提升、移动网网络规划和优化提供了有效支撑。

从 2009 年开始中国电信成立的八大基地，在运营过程中都用到了大数据的概念。目前为止中国电信在全国拥有 300 座以上的机楼，计算能力已经超过了 100 万处理器核心，存储能力已经达到 EB 的级别，在北京、上海、广东、四川部署了集团级的资源池，而且这些资源池的能力还在不断的扩张。

中国电信提出了大数据发展思路，并以综合平台、智能管道为依托，以丰富大数据为基础，聚焦重点大数据应用，特别是聚合更有价值的四大大数据商业应用模式，依托自身核心业务，以实现利润最大化。中国电信最有价值的大数据应用表现在四方面：语音数据分析、视频数据分析、流量分析、位置数据分析。

1. 利用大数据处理平台分析海量语音数据，建立呼叫中心测评体系和产品关联分析，为如保险公司等提供基于自动语音识别的大数据分析系统；
2. 基于智能图像分析能力的视频索引、搜索、摘要服务，从海量视频挖掘有价值的视频信息，提供公用视频图像分析，中国电信全球眼智能系统在智慧城市、平安社区、交通监管等领域大规模的使用；
3. 通过分析流量及协议信息，对一般性网络使用者的行为习惯分群组提供有针对性的网络便利性服务，比如精准广告；
4. 通过系统平台，对使用者的位置和运动轨迹进行分析，实现热点地区的人群频率的概率性有效统，比如根据景区人流进行优化。

2.3 网络空间安全与大数据

当前，网络空间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。总的来看，大数据在网络安全领域的应用呈现如下几个特征：

（一）实体-行为模型是大数据在网络安全领域应用的理论支撑

大数据技术的网络安全应用是个系统工程，迫切需求理论指导下构建体系框架，探索多种技术创新应用的方法学。从当前几个代表性项目来看，将着眼点放在各类实体的行为之上、通过重构网络空间行为模型构建大数据应用理论框架是通用做法。

这些实体包括物理实体，也包括虚拟实体。卡巴斯基、曼迪昂特等网络安全

产业公司的核心技术都是将主机恶意软件作为实体对象，通过对恶意软件行为分析来识别威胁。而 MITRE 公司则是在此基础上面向整个网络安全产业，通过恶意软件威胁信息的结构化，实现国家、私营网络安全产业公司甚至个人提交的威胁信息的自动交换，并作归因枚举、攻击模式分析、可观察化、自动化管理等研究，最终实现在国家应急反应框架下的网络威胁自动化处理。斯诺登所揭露的 X-keyscore 项目则将“人”作为实体，通过分析人在网络空间中的多跳信息交换行为 [18]，聚类分析得到可疑分子。而美军当前正在推进的 X 计划项目，则是将网络空间所有物理与虚拟实体---包括人在内，也包括路由器、服务器、终端、业务系统、软件工具---作为研究对象，以网络地图的方式实现网络空间态势感知 [19]，服务于网络攻防作战。

（二）信息萃取技术成为网络安全大数据应用的关键

实体行为体现于实体生产、存储、处理、转发各类信息的过程。这些信息在网络空间又可以分为元数据和内容数据两大类型，可以用来描述实体的行为。从中萃取实体的行为信息，依据实体行为模型进行综合分析，是网络安全大数据应用的基础和前提。高级信息萃取技术直接关系到实体行为分析的广度与深度。传统的元数据或者内容数据萃取技术包括：从通信语音识别、语种判断，到浏览器请求数据的语言识别，截获邮件的时间、地址分析、收发件人等信息分析，再到地理影像信息的判读、视频信息分析等。新兴的信息萃取技术则涉及面更广，有些甚至让人无法想象：包括从 office、PDF 等电子文档中分析该文档的传播路径、编辑修改作者信息、编辑修改计算机的 MAC 地址等信息；从浏览器 session 信息中获得上网人员的手机、住址、用户名、密码等信息；从移动设备中获得用户的地理位置信息；从用户使用 Google 地图的数据中获取用户敏感目标信息；从软件 MD5 信息中获得软件编辑环境、编辑作者、编码习惯等信息；从软件的操控对象或者网络访问目录中得到指挥控制中心信息；从软件二进制代码中分析得到软件演化行为的特征信息；从 VPN 网络中截获传输中的加密信息；从网络上直接嗅探发现所有可被攻破的主机信息 [20]；通过对海量加密文档样本的分析，获

[18] Hops Analysis，逐条分析，根据六度分离理论，地球上任何人都可以通过六跳形成关系。目前，美国宣称已经具有全球数据的三跳分析能力。

[19]网络地图，Cyber map,是将网络情报（cyber ISR）与全球地理信息相叠加，试图重构虚拟的网络地球，支持网络空间行为可视化分析。

[20]以上所述技术有些是成熟技术的大规模应用，有些是创新技术，更有的是通过与类似微软、谷歌、推特等信息服务巨头的隐秘合作实现的。

得当前主流加密算法的蛮力破解方法[21]。这些信息萃取技术五花八门，通过这种全方位、大纵身信息的萃取，形成物理实体或者虚拟实体的行为信息，再通过地址聚合分析、联系链逐跳分析等方法[22]，就可以逐步聚合得到各类的实体。

（三）人机结合成为网络安全领域大数据分析工具的基本形态

单以全球每天产生的海量信息来看，以当前软硬件能力存储、传递、分析所有萃取出的元数据与内容数据仍然相当困难。为此有很多解决方案，包括“将元信息自动分析预警与重点内容数据采集”相结合、“采集数据滚动式缓存与验证信息结构化提取”等等，都可以减少海量信息采集与存储带来的压力。但这样仍然不能解决海量信息的分析处理需求。当前的主要解决方案是将人机结合作为大数据分析工具的基本形态：通过“形式化”专家学者的智能经验，提升分析工具的智能化程度；将关键领域任务模型与规范化工作流相结合，打造任务敏感性分析方式；将情报分析专家与技术专家相结合，实时修订信息处理策略提升信息处理能力；以发展态势可视化技术为基本方向，做好人机结合式的网络安全态势理解。美国人正在做的系列项目，都是试图汇聚所有美国及其盟国情报系统的力量和资源，尽最大可能实现更广泛意义上的“人机”结合[23]。

（四）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

网络安全领域的大数据应用，要求数据的完整性、综合性，需要多种信息处理手段相结合，更需要高效组织运作机制的保障。当前，这些情报来源已经由传统的军方以电子监听为主[24]，转为以电子监听、网络监视相结合，甚至网络监视的业务量可能更大。这种组织构架全面、综合处理各类信息，能够有效的保证及时、准确的威胁告警与安全反应。2012 年 5 月美国击毙本拉.登的组织领导者就是中央情报局，美海军陆战队只是执行人。2013 年 8 月 4 日美国政府发布关闭全球大使馆安全告警，随后在也门打死六名基地组织武装人员，背后也都有大数据技术的有效支持。目前美军正在依照国家安全局的模式组建 100 个服务于联合作战的“营”级规模的网络作战分队，其核心装备仍需要大数据技术的支撑。国家力量对网络安全领域大数据的投入，彻底改变了国际网络安全力量格局，促成

[21]甚至解密本身，也成为大数据在网络安全领域的重要应用。

[22]斯诺登所暴恒星风项目，其基础就是人的地址数据库。而针对某地址所宿人员之间信息联系链条的逐跳分析是基本方法。

[23] X-keyscore 项目就是以此为目的，该项目汇聚了美、英、加拿大、澳大利亚、新西兰五国的情报力量，目前正在发展德国加入。

[24]电子监听主要来自于美军全球基地，各国大使馆，地面、水下、空中监视平台。

网络空间事实上的军事化。

（五）在大数据驱动下必须加强积极防御“网络武器”的研发工作

美国“网络武器”的研究已多达 2 千多种，其中最值得注意的是，“震网”病毒是世界上首个专门针对工业控制系统编写的破坏性病毒，被称为“网络空间的精确制导武器”。它能够利用 Windows 系统和西门子 SIMATIC WinCC 系统的漏洞进行攻击。攻击西门子公司控制系统的数据采集与监视控制系统(SCADA)，该系统广泛应用于能源、交通、水利、石油化工等领域，实现生产过程控制与调度的自动化。“震网”病毒侵入系统后，对可编程逻辑控制器进行重新编程，以达到破坏目的。在发起攻击破坏离心机成功后，同时用假数据欺骗操作和监控人员，病毒会向监控设备发送虚假的状态信息，使监控人员无法及时察觉设备的异常，从而最大限度地达到破坏效果。目前“网武”的危害性日益严重，已对国家重要工业与国防基础设施造成严重破坏，迫切要求解决相关威胁的大数据及其技术难题。因此，我国必须吸取伊朗核设施等工业系统被攻击的经验教训、启发和警示，大力加强大数据下国防重要基地和装置的网络安全技术，大力提高我国积极防御网络武器的技术水平[25]。

（六）数据采集政策成为影响网络安全大数据应用的根本

当前，影响大数据在网络安全领域应用的障碍除了资金与技术这两个因素以外，还有个更为根本性的问题就是数据采集政策与规范。曼迪昂特与当前网络安全产业公司不同之处在于，它与客户之间构建了一种信任关系。服务对象同意将本公司所属信息资源自动在线交给曼迪昂特，而曼迪昂特公司也承诺不会透露相关信息、损害相关公司的商业利益。而传统网络安全产业公司一直受困于不能过多的采集到用户主机上的数据。这就需要网络安全产业政策的支持。如果没有网络安全产业政策，曼迪昂特这种企业化服务模式只能针对大型企业，而且规模不可能无限拓展，因为规模扩大以后就无法保障隐私承诺。数据采集政策与规范已经成为影响网络安全产业发展的一个瓶颈。

没有数据采集政策，就没有大数据在网络安全领域应用的法律依据。但从长远看，网络空间公私合作治理是一个趋势，网络安全国际合作也是必然，数据采集政策前景应该是光明的。但要避免出现网络空间成为某国谋求自身利益的战略

[25]方锦清，我国网络空间安全与和领域网络面临的挑战与思考，核科技观察，2013 第 4 期，1-8。

工具，其中必然需要长久的国际斗争，这又涉及到更根本的政治问题。

因此，网络安全面临严峻挑战，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[26]。网络的开放性、复杂性和跨国性决定了网络安全是全球性挑战，单靠一国之力难以有效应对，必须通过国际合作共同破解这一难题。为此，我国政府一直遵循八字方针：“和平、安全、开放、合作”。中国国际战略学会已经提出的四点建议：共同制定网络空间规则；深化打击网络犯罪合作；加快网络防护技术发展；完善网络安全对话机制。今后各国应继续推动政府、企业、学术界建立对话交流机制，进一步加强网络安全事件应急处理方面的对话合作，增强互信，分享经验，化解分歧，共克难题。

2.4 城镇化、智慧城市与大数据

如果我们回顾科学技术革命的宏观规律，对比科技革命周期和经济波动周期，我们会发现，从全球的范围看，大科技革命的下一波高潮呼之欲出。在本世纪二十年代，前苏联学者康德拉季耶夫提出，在经济生活中存在着 45~60 年的长期波动。这种长期波动被人们称为康德拉季耶夫周期。而科学技术的发展是一个辩证地自我否定的过程。这种自我否定的力量是在它的发展过程中蓄积起来的，是一个由弱到强、再由强到弱的过程。科学技术的周期和经济波动的周期有紧密的相关性。21 世纪初期正处于下一次科技革命的前夜。如果我们进一步深入观察信息技术革命的小周期，我们会有一个结论：信息技术领域有些大事情正在发生。

今天在“智慧化”这个前沿，移动互联网、云计算、物联网和大数据是当前最重要的 4 架马车。作为 4 架马车中的一员，大数据正在带来全新的应用模式。中国智慧城市建设已拉开序幕。北京围绕城市智能运转、企业智能运营、生活智能便捷、政府智能服务等方面，全面启动“智慧城市”建设工程。南京重点铺开“智慧城市”，由政府指导、电信运营商建设围绕物联网技术，打造江苏智慧城市 13 个市分站。上海将建设国际型智慧城市。在 2013 年底达到“基础设施能级跃升、示范带动效应突出、重点应用效能明显、关键技术取得突破、相关产业国

[26]方锦清，我国网络安全面临的严峻挑战与若干应对建议，”复杂系统与复杂性科学”网络科学专刊，2014 年第 1 期。

际可比、信息安全总体可控”的目标。广州着力建设“智慧树”，建设“树”型智慧城市框架，囊括交通、信息服务、电子政务、城市综合管理、医疗、社区、市民卡等各方面。至 2012 年底，我国已有几百个城市提出建设智慧城市，41 个地级以上城市在“十二五”规划或政府工作报告中正式提出建设智慧城市，80%以上的二级城市明确提出建设智慧城市的发展目标。可以说，智慧城市已在中国遍地开花[27]。

大数据给我们带来了前所未有的新思维： 大数据的数据量够“大”，数据不再是稀缺资源；大数据的数据够“杂”，数据来源广泛，格式五花八门，用户需要从大量的数据里提炼出有价值的数据，个体数据的精确性不再重要，重要的是大多数数据群共同指出的结论；大数据的数据非常“快”， 数据产生得快，数据增加得快，数据随时间的折旧也快，数据的时效性成为关键。

因此，我们需要新的方法学。在数据极大丰富的前提下，我们需要新的分析思维和技术。

表2-2： 大数据方法与传统方法的对比

对比	**传统方法**	**大数据方法**
数据采集手段	采样数据	全局数据
数据源	单数据源	多数据源整合
判断方法	基于主观因果假设	机械穷举相关关系
演绎方法	孤立的推算方法	大数据+小算法+上下文+知识积累
分析方法	描述性分析	预测性和处方性分析
对产出的预期	绝对的精确性更重要	实时性更重要

智能交通的数据源是多数据源的集成。我们看到在今天的北京市的实际应用中，浮动车 GPS 每天产生 20MB 数据，交通监控中的视频/图像数据和元数据是每天几百 TB，GIS 数据（如手机位置信息）是每天 18MB，出租车运营数据是每天 1MB，交通卡数据是每天 19MB，高速公路收费数据是每天 0.5MB。另外城市供水系统、 智能电网、居民睡眠质量分析、社交网络情感分析都是智能交

[27]数据来源于 CCID：http://www.ccidnet.com

通的数据源。（部分数据源于北京 TOCC）.

在如今各大城市相继建设智能交通的进程中，各种路测和车载智能传感器以及信息化的交通业务系统，产生了大量的车辆信息、道路信息、出行者信息和管理服务信息，包含了城市道路、公路、地面公交、轨道交通、出租汽车、省际客运、公安交通管理、民航、铁路，甚至气象等零零总总的数据内容。这些交通数据容量大、增长快、结构多样化，不少数据价值密度低，有待深入的处理挖掘。随着中国智能交通建设进程的逐步推进，交通数据已经从稀缺走向了极大丰富，并带来了交通大数据的严峻挑战：

●□ 极大丰富的交通数据未能有效整合，数据依类别、行业、部门、地方被隔离，数据之间的关联性被遗忘，道路视频作为最大的交通信息源没有被充分利用，公众无法获取准确连贯的出行服务信息；

●□ 数据来源众多，存储方式多样，数据类型复杂，包含大量视频、图像等半结构、非结构化数据，并且数据无统一标准，在组织、融合、清洗和转换这些数据时的难度较大；

● 为深入挖掘交通数据的潜在价值，需要一个数据管理平台来处理各种类型和规模的数据，该管理平台还需要同时能处理结构化数据、半结构化数据和非结构化的数据。

● 针对高增长、规模日益庞大的交通数据，我们需要一种高效的大数据处理技术，对交通数据进行快速有效的挖掘分析，从中提炼出高价值的信息，并灵活支撑日益增多的各类交通业务应用需求。

面向大数据的交通数据活化应用主要包含三大机制和五大转变：

三大机制：

- 第一时间发现问题机制：实时的交通大数据分析能力，能够帮助人们在海量的交通数据中快速发现异常，定位症结，锁定线索。
- 第一时间处置问题机制：分布式的交通大数据处理能力，可支撑高并发多用户访问，协助人们在紧急事件中多方协作、快速处置；
- 第一时间解决问题机制：高效的交通大数据挖掘能力，能够快速碰撞发现海量交通数据中的内在关联规律，帮助人们高效分析解决疑难问题。

五大转变：

- 化被动为主动：传统模式下人们往往是被动处理各类紧急事件，而大数据模式下将提供预测、预警机制，可主动部署人力、调动资源。
- 化僵化为灵活：传统统计报表多为一天一生成或一月一生成，程序僵化、变动不易，而大数据模式下用户可自由生成各种统计报表，而无需系统事先预制报表。
- 化低效为高效：传统模式下的海量数据模糊查询和统计分析无法达到用户的实时使用需求，而大数据模式则提供秒级响应的用户体验。
- 化单一为互动：传统的数据应用多为单表挖掘分析，一旦涉及到跨表关联碰撞就会因效率问题而无能为力，而大数据模式则擅长复杂的跨表关联分析，推动数据串并关联，产生更大的价值。
- 化粗放为精细：从原来粗放式的设备购买和升级更换，变为精细化的设备集群化平行扩展、按需精准投资；从原来粗放式的数据访问体验（每次可查询内容较少，为达目的需叠加多次操作），变为精细化的数据访问体验（系统支持自动关联和推送式信息服务，用户一次查询可获得更丰富的信息内容）。

2.5 金融与大数据

大数据是改变国家命运、产业格局的关键技术。从数据角度看，金融无非是各种数据的排列组合。纵观历史，从 19 世纪 30 年代电报的兴起，到后来电话、计算机，乃至今天互联网、移动互联网，每一次通讯信息技术的变革都对金融业产生了巨大的影响。大数据时代，凡是拥有独特数据资产的公司，都可以涉及金融。利用新兴的大数据技术，金融业的两大根基——征信与风控，即将发生革命性的变化。

互联网金融是当下的一个热词，言下之意是指利用互联网技术、大数据思维进行的金融业务再造。总体而言体现在两个方面，一是金融机构依靠互联网技术和思维自我变革，如招商银行和平安银行，二是互联网企业跨界开展金融服务业

务撼动传统金融格局，如阿里金融、腾讯微信支付和京东金融等。尤其是这些新兴的互联网金融机构源源不断涌现出来并推动着金融业在更大空间、更广地域进行着深刻而卓有成效的金融创新，促使金融业由量变到质变，推动着金融业由不可能走向可能、由不完备走向完备、由不受关注走向备受关注，如在小额贷款和中小企业融资领域的 P2P 和众筹融资模式。因此，互联网金融不仅是互联网、大数据等技术在金融领域的应用，更是基于大数据思维而创造出的新的金融形态。前者是金融企业将越来越多甚至全部的金融业务搭载在互联网平台上，后者是互联网企业以互联网技术平台为优势而不断搭载金融业务，两者不断趋同，但各有各的优势。虽然由于监管和历史优势的原因，难以在短期内处于完全竞争状态，但是互联网企业的金融业务创新越来越对传统金融企业构成巨大的竞争威胁，甚至会蚕食金融企业的优势空间，尤其是在目前大多数金融企业想做而做不好的中小企业融资、小额和第三方支付领域等巨大蓝海区域，将是互联网企业开拓金融业务的巨大空间。

全球领先的软件和技术服务企业 SunGard 发布了 2012 年金融服务业各部门“大数据”发展的十大趋势。它们分别是：

第一、市场数据集变得越来越庞大，业务对数据的细分粒度要求越来越高，以满足预测模型、业务预测和交易影响评估的需求。

第二、新的监管和合规要求更强调治理和风险汇报，推动了全球性金融机构对更深入和透明的数据分析需求。

第三、金融机构不断完善自身的企业风险管理框架，基于主数据管理策略开发的框架可协助企业提高风险透明度，加强风险的可审性和管理力度。

第四、金融服务公司都希望能充分利用各种服务交付渠道（如分公司、网络、移动通信等）的海量客户数据，开发新的预测分析模型，实现对客户消费行为模式进行分析，提高客户转化率。

第五、在巴西、中国和印度等新兴市场，经济和业务增长机会正在超越欧洲和美国，大量投资投向了本地和云数据处理基础设施中。

第六、“大数据”在存储和处理框架两方面的优势将帮助金融服务企业充分掌握业务数据的价值，降低业务成本并发掘新的套利机会。

第七、面对“大数据”所带来的不断增加的数据量要求，需要对传统的数据

传输工具 ETL（提取、转换和加载）流程进行重新设计。

第八、大量历史客户支付行为数据的信用风险预测模型正在零售与公司贷款催收中得到大量应用，通过该技术，银行可以通过对不同客户违约和还款资料进行分析，对催收次序进行优化。

第九、随着以平板电脑和和智能手机为代表的移动应用和互联网工具的迅速普及，技术基础设施和网络在对不同来源、不同标准数据进行处理、编索和整合方面的压力不断增大。

第十、“大数据”推动了对数据处理算法的需求，提出对数据安全和访问控制的重视，并可有效降低对现有系统的影响。

目前，中国金融行业已形成共识——数据是重要资产。数据的真正价值在于能够洞察企业内部规律，数据的洞察力成为金融企业的核心竞争力。在中国金融行业信息化建设中，与信息加工密切相关的大数据管理正逐渐成为与核心业务系统建设、渠道建设和前置建设同等重要的领域。经过多年的发展与积累，目前中国的大型商业银行和保险公司的数据量已经达到 100TB 以上级别，并且非结构化数据量在迅速增长。中国金融行业已步入大数据时代的初级阶段，并且呈现快速发展势头。

金融业面临众多前所未有的跨界竞争对手，市场格局、业务流程将发生巨大改变，未来的金融业将开展新一轮围绕大数据的 IT 建设投资。据悉，目前中国的金融行业数据量已经超过 100TB，非结构化数据迅速增长。分析人士认为，中国金融行业正在步入大数据时代的初级阶段。优秀的数据分析能力是当今金融市场创新的关键，资本管理、交易执行、安全和反欺诈等相关的数据洞察力，成为金融企业运作和发展的核心竞争力。

目前，以大数据为代表的新型技术将在两个层面改造金融业。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。此外，腾讯、阿里巴巴等互联网企业也在凭借其强大的数据积累和客户基础，进军金融业，开拓新的盈利点，这也成为金融产品在线销售的一大推动力。

2012 年 11 月，中共“十八大”提出将金融体制改革列为未来十年发展的重中之重。当前，中国各金融企业都制定了“十二五”发展规划，“科技引领创新”是最为核心的指导思想。未来中国的金融企业将依靠构建智慧型的数据分析体系，充分挖掘数据中的规律,以支持业务创新与服务创新。

从未来几年看，中国金融行业在“十二五”时期面临发展方式转型的挑战，转型主要集中在三大方面：

一，中国金融行业将根据巴塞尔协议 III 和第二代偿付能力等的要求，建立全面的风险管理体制，向严监管转型。大数据能够加强风险的可审性和管理力度。

二，从粗放式管理向精细化管理转型，信息化重点也将从业务信息化向管理信息化转变。大数据能够支持精细化管理。当前中国银行业利率市场化改革已经起步，利率市场化必然会对银行业提出精细化管理的新要求。

三，从“利润为中心”和“保单为中心”向“客户为中心”转型。大数据支持服务创新，能够更好地实现“以客户为中心”理念，通过对客户消费行为模式进行分析（比如事件关联性分析），提高客户转化率，开发出不同的产品以满足不同客户的市场需求，实现差异化竞争。

2.6 健康医疗与大数据

伴随着中国医疗卫生服务的信息化进程推进，必将产生大量的数据。这些数据主要来源于医疗业务活动、健康体检、公共卫生 9 项服务等医疗卫生服务。数据内容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。其中大量充斥着非结构化/半结构化的数据，包括图像，office 文档，以及 XML 结构文档等。

随着国家积极倡导的 3-5-2-1 区域医疗系统的建设，预计在全国会出现上百个医疗数据中心，每个数据中心都将承载近 1000 万人口的医疗卫生数据并提供各种类型的服务。根据估算，中国一个中等城市（一千万人口）50 年所积累的医疗数据量就会达到 10PB 级[28]。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。

从目前各地的医疗信息系统的建设来看，尽管数据已经电子化，但绝大多数

[28]数据来源 CCID：http://www.ccidnet.com

的医疗数据是处于归档状态。如果要想快速检索是十分复杂的，这些数据仍然分散的存储于不同的业务系统中。过去不是没有整合这些数据的需求，而是缺乏适合的技术手段。单纯的使用传统的存储技术和分析处理方法，存在两个问题：一个是多样的和变化的数据格式，另一个是大数据量的数据获取速度。

大数据解决方案在医疗行业的应用场景众多，以下是目前在中国医疗行业中，大数据的几个主要应用场景：

1) 居民健康档案数据管理和服务

“健康档案是个人全生命周期的医疗/健康数据的管理”，其应用场景包括：

- 从医生及卫生行政管理人员的角度来看，全生命周期的健康档案调阅有着现实意义。例如，对于慢性病患者，以往病程的变化、治疗的过程都对医生诊断和处置有着重要的辅助作用。过敏史、不良反应这些数据对避免出现医疗差错和事故也有着积极的作用。
- 对于海量的医疗及健康数据进行统计和分析，为管理决策、监管实施等提供了更为科学的依据。
- 传统的临床科研往往基于抽样调查进行，而随着健康档案数据的丰富可以大幅减轻工作量，同时提高科研数据的质量和数量以及数据处理的效率。

2) 医院的大数据管理和服务

伴随着医疗技术的发展，医院积累了大量不同类型的数据，比如医疗数据、音频、视频和图片等数据信息，这些数据已经成为医院宝贵的财富。医疗数据大量是非结构化、半结构化的数据，包括文本信息、图片、影像、多媒体信息等。麦肯锡的预测表明，未来非结构化和半结构化的数据将大幅增长，其中影像和电子病历的数据量占到医院整体数据量的一半以上，电子病历相关的数据将持续大幅增长，而影像依然是医疗数据中数据量最大的部分。

医院的大数据管理和服务，将主要集中在临床诊断和临床科研，并且为医院管理层的决策支持提供实时有效的数据服务。与健康档案数据相比较，医院院内的数据具有更好的数据质量和实时性。基于院内各不同的信息系统的数据，可以进行更为多样的大数据分析和处理。

3) 大数据在医疗其他领域的应用

来自赛迪顾问的数据表明，智慧医疗相关的大数据应用规模年度复合增长率将达到 111.3%（2013-2015）。未来的大数据的利用前景十分广阔，不仅用于临床诊断、临床科研，而且为政府公共卫生决策及个人管理健康都会发挥积极的作用。在医疗保险、生物制药、生命科学等相关计算领域中，大数据的技术也将广泛应用。

2.7 生物信息、制药与大数据

在生物信息、制药方面，大数据应用在数据规模、多样性、处理速度等维度均出现了巨大的变化。

首先，数据规模呈指数型增长。医疗数据规模将由 2005 年的 130 艾字节（10^{18} 字节）上升至 2015 年的 7,910 艾字节，并将于 2020 年达到 35 泽字节(10^{21} 字节)。然而，全世界医疗数据仅有五分之一为适于计算机处理的结构化数据，其余五分之四为非结构化数据，包括手写病历、未归类文档及音、视频文件等，以 15 倍于结构化数据的增长速度增加。现有的医疗数据形式包括病历、放射影像、临床试验数据、遗传学及人口学数据、基因组序列等。新的大数据形式包括 3D 影像、基因组及生物传感器输入数据。

为应对数据规模的增长，人们应用虚拟化及云计算等技术实现有效的数据管理。知名厂商如 IBM、思科、谷歌及 DNAnexus，Appistry，NextBio 等医疗数据专业厂商都在开发相应产品和服务，以帮助客户管理庞大的数据。

其次，医疗数据的多样化也给管理和应用带来了很大困难。大多数医疗数据为非结构化数据，包括医疗档案、手写医嘱、出入院记录、纸质处方、放射、核磁共振和 CT 影像等。目前，来源于遗传学和基因组学研究、社交媒体、康复健身设备等的结构和非结构化数据流混杂在一起，很难直接利用计算机进行存储与管理。

为解决上述问题，IBM 拟利用 Watson 的自然语言处理能力融合多种数据集；Health Fidelity 也采用自然语言处理技术实现非结构化数据到结构化的转化；包括 Explorys、PracticeFusion、athenahealth Inc.、Humedica 在内的其它厂商也在开展此类业务。大数据医疗应用的潜力主要在于结合多种数据形式，从个人及群体两个层次开展研究，探索发现数据规律。多种数据源结合可使相关研究更为迅速

和可靠地开展。例如，医药厂商可将基因组数据与群体临床数据结合起来，有助在第一时间面向受众发布精准的治疗药物和治疗方案。

第三，医疗数据正以更快的速度源源不断地产生，因此要求数据的采集、分析、比较和决策从较慢的批处理向实时处理方式转变。未来重症监护室（ICU）中的实时数据分析处理，如尽早发现感染并及时实施包括广谱抗生素在内的多种治疗，可有效降低患者的发病率和死亡率。目前，实时数据流已用于监控 ICU 内的新生儿，用于预测致命的感染。从事此类研发工作的公司包括 Baxter International、Boston Scientific Corporation、HospiraInc、Medtronic Inc、Zoll Medical Corporation 和 Abiomed Inc。

基于以上变化，诞生了下列典型应用：

（1）电子病历的有效发掘与利用。斯坦福大学把所有医院的电子病例及数据库，都转换成斯坦福大学数据中心的数据，从许多不同的来源解析成堆的数据，试图发现那些对于解决问题来说最有用的模式，以便使管理人员更加全面地了解病人的各种需求。斯坦福大学转化医学中心主任研究员 Bruce Ling 称，目前的一些数据已经不能满足行政管理者的需求。例如，对于一般心脏疾病的治疗，各地区诊所分布图与目前病人居住的地方其实是不重叠的。所以，就需要通过大数据来分析各种类型的病人都集中在哪个区域，以此来重新部署各诊所的分布，以满足不同患者的需求。不仅如此，Bruce Ling 称，从新兴的分子诊断方法到行医的最新标准，这些数据也同样可以录入电子病历。在他看来，医疗行业进入大数据时代后，就可以发生质的变化，对医生和病人都会带来更多的实惠。

（2）基因组学数据应用。基因组学可以说是大数据在医疗健康行业最经典的应用。基因测序的成本在不断降低，同时产生着海量数据。许多公司和研究机构正通过高级算法和云计算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。大数据将会给个体化医疗带来很多机遇，对于制药公司来说，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几率。可以说，掌握基因测序技术，挖掘出丰富且有效的信息，对于个性化的诊治非常关键。而用于医疗的各种影像手段，本身的发展同样也是大数据爆炸的过程。中国科学院深圳先进技术研究院研究员郑海荣表示，传统的影像方式看不到分子级的病变，基于基因组学的不断

发展，现如今不仅可以从成像上获得病变分子的存在，甚至还可以获得细胞级的存在。“在医院里发现了肿瘤，原因是源于微观尺度的变异。如果把每个尺度的信息都获取，那我们就可以生活在一个非常立体的和多维度的世界。”郑海荣说。与此同时，斯坦福大学心脏科急救中心主任 Andrew Shin 也指出，建立医疗保健大数据平台，就需要将不同模式结合起来。即将影像学、基因组学等不同数据模式进行整合，使它能够进行结构化及数据一体化处理，并为患者提供更高的价值。

（3）健康应用。国家地理杂志和时代杂志的前摄影师 Rick Smolan 推出了一个名叫 Human Face of Big Data 的移动应用，帮助人们挖掘自身收集、分析甚至可视化大量数据的能力。现在，人们可以通过智能手机、汽车甚至地毯跟踪自身及亲友的健康状况。GE 和 Intel 联合开发的“魔毯”项目，通过在家中地毯内置传感器感应家中老人下床和行走的速度和压力。一旦这些数据发生异常，系统就会给亲人发送一个警报信息。Smolan 称，这个装置没有安装摄像头，也不会侵犯用户隐私。纽约西奈山医学院生物医学信息部主任 Joel Dudley 表示：“这些个人追踪设备让大数据涉及到的医疗保健范围扩大到了诊所和医院之外。它们给了普通人主动探索科技与医疗问题的机会，让用户及时看到自身相关的数据，以及这些数据是如何改变他们的健康和生活。”

（4）医疗数据分析。没有配套的分析工具，收集了再多的数据也是枉然。据悉，美国政府已经意识到此困境，并拨款 2 亿多美金支持大数据分析项目。2012 年 10 月，正在联合进行 8 个大数据研究项目的美国国家科学基金会和美国国立卫生研究院（NIH）就获得了 1500 万美元的科研基金。NIH 院长 Francis Collins 称：“这笔资金将帮助我们更快地开发出适当的方法从大量复杂数据中提取重要的相关信息进行分析，从而早日改善人们的生活。”与此同时，NIH 还将自己的研究数据传上云端分享给其他科学家。当这类跟踪设备逐渐普及，隐私问题必然会随之出现。Joel Dudley 则认为，未来关于隐私的讨论会和现在有着很大的不同。试想，当你拍下某个人的照片，并通过某个能够分析其 DNA 状况的应用发现他有患上某种疾病的风险时，被拍者应该不会纠结于自身隐私被侵犯以至于忽略自己的健康状况。

（5）精神卫生应用。随着社会生活压力不断增大，心理疾病，尤其是抑郁

症发病率不断升高。根据世界卫生组织[29]估计，约有四分之一的世界人口会在一生中出现抑郁症状。据世界卫生组织资料显示，全球病历记录抑郁症患者高达1.22 亿，且仍在迅速增长，很快将超过冠心病成为全球主要健康危害第二大疾病，其在成年人口中的终生患病率达 5%-10%。全世界有 85 万重度抑郁症患者实施自杀。在欧洲，有 8%的女孩和 2%的男孩表现出了重度抑郁症状。据中国心理卫生协会[30]数据显示，目前中国有病例记录抑郁症患者已超过 3000 万。

为实现及时有效的精神疾病发现与预测，许多研究者结合生物传感器与数据建模，通过对多种生物、心理数据的特征筛选与建模预测或诊断精神疾病，并进行及时干预。由于用户需要一周七天，一天 24 小时穿戴便携式生物数据采集设备，因此对于数据的获取、处理与分析提出了新的要求，包括：（1）数据采集的连续性、可靠性与稳定性；（2）实时高效的数据处理方式；（3）可更有效地融合多种类型及格式的生物、心理数据，在已建立的多模态数据模型下通过发掘多指标体系及其关联演化发现精神疾病模型。

[29]http://www.who.int
[30]http://www.camh.org.cn

第三章 大数据技术体系现状

根据大数据处理的生命周期，大数据的技术体系通常可以分为大数据采集与预处理，大数据存储与管理，大数据计算模式与系统，大数据分析与挖掘，大数据可视化计算以及大数据隐私与安全等几个方面。

3.1 大数据采集与预处理

3.1.1 问题与挑战

根据MapReduce产生数据的应用系统分类，大数据的采集主要有四种来源：管理信息系统、Web 信息系统、物理信息系统、科学实验系统。

1. **管理信息系统**是指企业、机关内部的信息系统，如事务处理系统、办公自动化系统，主要用于经营和管理，为特定用户的工作和业务提供支持。数据的产生既有终端用户的原始输入，也有系统的二次加工处理。系统的组织结构上是专用的，数据通常是结构化的。
2. **Web 信息系统**包括互联网上的各种信息系统，如社交网站、社会媒体、搜索引擎等，主要用于构造虚拟的信息空间，为广大用户提供信息服务和社交服务。系统的组织结构是开放式的，大部分数据是半结构化或无结构的。数据的产生者主要是在线用户。电子商务、电子政务是在 Web 上运行的管理信息系统。
3. **物理信息系统**是指关于各种物理对象和物理过程的信息系统，如实时监控、实时检测，主要用于生产调度、过程控制、现场指挥、环境保护等。系统的组织结构上是封闭的，数据由各种嵌入式传感设备产生的，可以是关于物理、化学、生物等性质和状态的基本测量值，也可以是关于行为和状态的音频、视频等多媒体数据。
4. **科学实验系统**，实际上也属于物理信息系统，但其实验环境是预先设定的，主要用于研究和学术，数据是有选择的、可控的，有时可能是人工模拟生成的仿真数据。

在物理信息系统中，对于一个具体的物理对象，可采用不同观测手段，对其

不同的属性（方面）进行测量，如测量一辆行驶汽车的尺寸、速度、路线、尾气、外观等，其观测结果为具有不同形式的数据，这些数据代表实体不同的模态，称为多模态(multi-modal)。对于一个实体的多模态原始数据，需要做融合处理(data fusion)。在融合处理中，需要减少误差，保证数据的完整性和正确性。在高级的嵌入式系统或数据采集系统中，通常具有数据质量控制和数据融合处理功能[2]。

从人-机-物三元世界观点看，管理信息系统和 Web 信息系统属于人与计算机的交互系统，物理信息系统属于物与计算机的交互系统。关于物理世界的原始数据，在人-机系统中，是通过人实现融合处理的；而在物-机系统中，需要通过计算机等装置做专门的处理。融合处理后的数据，被转换为规范的数据结构，输入并存储在专门的数据管理系统中，如文件或数据库，形成专门的数据集。

对于不同的数据集，可能存在不同的结构和模式，如文件、XML 树、关系表等，表现为数据的异构性（heterogeneity）。对多个异构的数据集，需要做进一步集成处理（data integration）或整合处理（data consolidation），将来自不同数据集的数据收集、整理、清洗，转换后，生成到一个新的数据集，为后续查询和分析处理提供统一的数据视图。

通常大数据描述了一个对象(物理的或逻辑的)或一个过程的全景式的和全周期的状态，因此，其来源必然是多源的，其形式是多模态的。数据的多源和多模态的不确定性和多样性，必然导致数据的质量存在差异，严重影响到数据的可用性。由于数据量的大规模性，即使错误数据的相对比例不大，而绝对的错误数据量也是非常可观的。据国际咨询机构调查，全球财富 1000 强企业中 25%以上的企业信息信息系统存在不正确的数据，美国企业信息系统中 1%-30%的数据存在各种错误，美国工业企业由于数据错误而引起的生产事故和决策错误，每年造成 6000 多亿美元的损失[3]。

数据的可用性取决于数据质量。数据质量的定义有很多说法。按照文献[4]的定义，数据质量包含 5 种特性：精确性、一致性、完整性、同一性和实效性。精确性指数据符合规定的精度，不超出误差范围；一致性指数据之间不能存在相互矛盾；完整性指数据的值不能为空；同一性指实体的标识是唯一的；时效性指数据的值反映了实际的状态。此外，考虑到人为因素，还可以要求第 6 个性质，真实性，即数据不能是人工伪造的。

3.1.2 主要进展

针对管理信息系统中异构数据库集成技术、Web 信息系统中的实体识别技术和 DeepWeb 集成技术、传感器网络数据融合技术已经有很多研究工作，取得了较大的进展，已经推出了多种数据清洗和质量控制工具[31]，例如，美国 SAS 公司的 Data Flux，美国 IBM 公司的 Data Stage，美国 Informatica 公司的 InformaticaPower Center。

但是，针对各种类型、各种应用的大数据的特点，如何保证一致性、精确性、完整性、统一性、时效性、真实性等六个性质，并且保证可行的处理效率，还缺乏全面系统的研究，许多新问题有待于发现和解决。

3.1.3 发展趋势

为了保证大数据的可用性，首先必须在数据的源头上把好质量关，做好从原始数据到高质量信息的预处理。具体的关键技术有：

（1） **数据源的选择和高质量原始数据的采集方法。**用于从可靠的高质量数据源里，获得高质量的原始数据。为了确保数据源的质量，需要建立数据源的质量评估理论模型，包括数据源的综合质量评估和高质量数据源的选择方法。然后，针对各种模态数据的特点，建立高质量多模态数据的获取方法，包括：有效的数据采集方法、多模态数据融合算法、数据的保质转换算法、数据精确性和一致性方面的错误校验和纠错、数据完整性方面的缺失值估计、数据的时效性检测、数据的真实性验证等。

（2） **多源数据的实体识别和解析方法。**用于识别和合并相同的实体，区分不同的实体。为了高质量的数据集成奠定基础，必须保证数据的实体同一性，解决来自多个数据源的多模态数据的实体识别问题。需要建立多源数据的实体关联模型和识别模型、多源多模态数据的实体自动识别方法、实体识别效果的评估模型等。

（3） **数据清洗和自动修复方法。**根据正确性条件和数据约束规则，清除不合理和错误的数据，对重要的信息进行修复，保证数据的完整性。需要建立数据正确性语义模型、关联模型和数据约束规则、数据错误模型和错误识别学习框架、针

[31] Gartner 研究报告, Magic Quadrant for Data Quality Tools, 2012 年 8 月，http://useready.com/wp-content/uploads/2013/07/Gartner-Data-Quality-2012.pdf

对不同错误类型的自动检测和修复算法、错误检测与修复结果的评估模型和评估方法等。

（4） **高质量的数据整合方法。**在数据采集和实体识别的基础上，进而实现数据到信息的高质量整合。需要建立多源多模态信息集成模型、异构数据智能转换模型、异构数据集成的智能模式抽取和模式匹配算法、自动的容错映射和转换模型及算法、整合信息的正确性验证方法、整合信息的可用性评估方法等。

（5） **数据演化的溯源管理。**用于对数据的演化过程进行跟踪和记录，以保证和控制数据的质量。需要建立世系模型及其追踪技术，主要包括时空、多粒度、多路径和不确定的海量信息演化的演化模型和演化描述方法、演化模式的正向性评估模型与方法、演化的可逆性判定与近似求解算法、分布式、多粒度、概率化的世系追踪技术等。

总之，大数据的采集和预处理是大数据的源头，在源头上把好质量关，对大数据的后续处理和分析至关重要。因此，对大数据的使用者、研究者、开发者以及上级主管部门，提出如下建议：

（1） 提高用户对大数据可用性的重要性的认识，切实开展大数据质量控制，确保大数据处理和分析结果的正确性；

（2） 针对大数据质量控制面临的挑战性问题，学术界应加强对大数据可用性评估和保证的关键技术的研究和开发；

（3） 大数据的质量控制具有广泛的需求和巨大的市场前景，工业界应注重大数据可用性的评估，加强数据质量保证软件的开发和推广。

（4） 建议政府有关部门尽快建立关于大数据可用性（数据质量）的标准，保证大数据的统一质量，有效保证大数据的利用价值。

3.2 大数据存储与管理

3.2.1 问题与挑战

大数据给存储系统带来了 3 个方面的挑战：1）存储规模大，通常达到 PB（1,000 TB）甚至 EB（1,000 PB）量级。2）存储管理复杂，需要兼顾结构化、非结构化和半结构化的数据。3）数据服务的种类和水平要求高，换言之，上层

应用对存储系统的性能、可靠性等指标有不同的要求，而数据的大规模和高复杂度放大了达到这些指标的技术难度。这些挑战在存储领域并不是新问题，但在大数据背景下，解决这些问题的技术难度成倍提高，数据的量变终将引起存储技术的质变。

大数据环境下的存储与管理软件栈，需要对上层应用提供高效的数据访问接口，存取 PB 甚至 EB 量级的数据，并且能够在可接受的响应时间内完成数据的存取，同时保证数据的正确性和可用性；对底层设备，存储软件栈需要充分高效的管理存储资源，合理的利用设备的物理特性，以满足上层应用对存储性能和可靠性的要求。在大数据带来的新挑战下，要完成以上这些要求，需要更进一步的研究存储与管理软件技术[5]。

3.2.2 主要进展

根据为上层应用提供的访问接口和功能侧重不同，存储与管理软件主要包括文件系统和数据库；在大数据环境下，目前最适用的技术是分布式文件系统、分布式数据库以及以及访问接口和查询语言。

- **分布式文件系统**

分布式文件系统所管理的数据存储在分散的设备或节点上，存储资源通过网络连接。用分布式文件系统对大数据进行存储与管理，目前的研究，主要涉及到以下几个关键的技术：

（1） **高效元数据管理技术。**大数据应用下，元数据的规模也非常大，元数据的存取性能是整个分布式文件系统性能的关键。常见的元数据管理可以分为集中式和分布式元数据管理架构。集中式元数据管理架构采用单一的元数据服务器，优点是实现简单，但存在单点故障等问题。分布式元数据管理架构[6]则将元数据分散在多个结点上，从而解决了元数据服务器性能瓶颈问题，提高了可扩展性，但实现复杂，并引入了元数据一致性的问题。此外，还有一种无元数据服务器的分布式架构，使用在线算法组织数据，不需要专用的元数据服务器。但是该架构对数据一致性的保证很困难，实现复杂。文件目录遍历操作的效率低下，并且缺乏文件系统全局监控管理功能。

（2） **系统弹性扩展技术。**大数据环境下，数据规模和复杂度的增加往往非

常迅速，所以按需的扩展系统规模，是十分必要的。实现存储系统的高可扩展性首先要解决两个方面的重要问题，元数据的分配和数据的透明迁移。前者主要通过静态子树划分和动态子树划分技术实现，后者则侧重数据迁移算法的优化。此外，大数据存储系统规模庞大，结点失效率高，因此还需要实现一定程度上的自适应管理功能。系统必须能够根据数据量和计算的工作量估算所需要的结点个数，并动态地将数据在结点间迁移，以实现负载均衡；同时，结点失效时，数据必须可以通过副本等机制进行恢复，不能对上层应用产生影响。

（3） **存储层级内的优化技术。**构建存储系统时，需要基于成本和性能来考虑，因此存储系统通常采用多层不同性价比的存储器件组成存储层次结构。大数据的规模大，因此构建高效合理的存储层次结构，可以在保证系统性能的前提下，降低系统能耗和构建成本。利用数据访问局部性原理，可以从两个方面对存储层次结构进行优化。从提高性能的角度，可以通过分析应用特征，识别热点数据并对其进行缓存或与预取，通过高效的缓存预取算法和合理的缓存容量配比，以提高访问性能。从降低成本的角度，采用信息生命周期管理方法，将访问频率低的冷数据迁移到低速廉价存储设备上，可以在小幅牺牲系统整体性能的基础上，大幅降低系统的构建成本和能耗。

（4） **针对应用和负载的存储优化技术。**传统数据存储模型需要支持尽可能多的应用，因此需要具备较好的通用性。大数据具有大规模、高动态及快速处理等特性，通用的数据存储模型通常并不是最能提高应用性能的模型，而大数据存储系统对上层应用性能的关注远超过对通用性的追求。针对应用和负载来优化存储，就是将数据存储与应用耦合，放宽 POSIX 接口，简化或扩展分布式文件系统的功能，根据特定应用、特定负载、特定的计算模型对文件系统进行定制和深度优化，使应用达到最佳性能。这类优化技术在 Google[7]、Facebook[8]等互联网公司的内部存储系统上，管理超过 PB 级的大数据，能够达到非常高的性能。

（5） **针对存储器件特性的优化技术。**随着新型存储器件的发展和成熟，Flash、PCM 等逐渐开始在存储层级中占据一席之地，存储软件栈也随之开始逐渐发生变化。以 Flash 为例，起初各厂商通过闪存转换层 FTL 对新型存储器进行封装，以屏蔽存储器件的特性，适应存储软件栈的现有接口。但是随着 Flash 的普及，产生了许多针对应用对 FTL 进行的优化[9][10]，以及针对 Flash 特性进行

定制的文件系统，甚至有去掉FTL这层冗余直接操作Flash的存储解决方案[11]。传统的本地文件系统，包括分布式文件系统，是否能够与新型存储器件耦合，最大程度地利用这些存储器件新特性上的优势，需要存储软件开发者重新审视存储软件栈，去除存储软件栈的冗余，甚至需要修复一些不再合适的部分[12]。

➢ 分布式数据库

大数据时代企业对数据的管理、查询及分析的需求变化促生了一些新的技术的出现。需求的变化主要集中在数据规模的增长，吞吐量的上升，数据类型以及应用多样性的变化。数据规模和吞吐量的增长需求对传统的关系型数据库管理系统在并行处理，事务特性的保证，互联协议的实现，资源管理以及容错等各个方面带来了很多挑战。而数据类型以及应用的多样性带来了为了支持不同应用的数据管理系统。

● 事务性数据库

这类数据库主要包括 NoSQL 和 NewSQL。NoSQL（“Not Only SQL”或者“Not Relational”）系统往往通过放松对事务 ACID 语义的方法来增加系统的性能以及可扩展性（CAP 定理）。NoSQL[32]系统往往具有以下几个特征：

（1） 非关系数据模型，比如键值存储等

（2） 对简单操作比如键值查询的水平可扩展性，往往不支持 SQL 全集。

（3） 在多个节点中分割和复制数据的能力

（4） 弱并发一致性语义（比如最终一致性）

（5） 充分利用分布式索引和内存

根据管理数据的模式分类，NoSQL 系统可以分为三类：键值系统，文档存储系统以及图数据库。键值系统的代表性系统包括 BigTable，Dynamo, HBase，Gemfire, Redis,, Cassandra, 文档存储系统的代表包括 MongoDB 和 Couchbase，图数据库的代表是 Neo4j 等等。

NoSQL 系统通过对事务语义的放松达到系统的可扩展性，但是把一致性的维护交由用户来管理，这对很多对一致性要求不高的应用来说是足够的。但是如果应用需要保证一致性，这对开发人员来说就很困难了。NewSQL 是在这样的背景下诞生的。NewSQL 系统可以在提供类似 NoSQL 的可扩展性的同时保证事务

[32] NoSQL. http://en.wikipedia.org/wiki/NoSQL

ACID 属性，并且提供 SQL 用户接口。NewSQL 系统通常可以分为两类。

（1） 通用数据库：这类系统保持传统分布式数据库的功能，但是在设计分布式体系架构时充分考虑了大规模高吞吐系统的特性。这类系统的典型代表是 Spanner 和 NuoDB。

（2） 基于内存的数据库：这类系统基本上针对的是高吞吐短小事务，不再采用传统的关系型数据库设计。这类数据库的典型代表是 SQLFire 和 VoltDB。

- **分析型数据库**

分析型数据库在大数据时代也呈现了一种百家争鸣的局势。自从 MapReduce[13]被提出以及 Hadoop[33]的流行，出现了多家针对 Hadoop 的 SQL 分析引擎。代表性系统包括 Hive，HAWQ， Impala 和 Hadapt。

Hive[34]是一个基于 MapReduce 的 SQL 引擎。基本原理是接受 SQL，解析 SQL，然后把 SQL 语句翻译成多个 MapReduce 的任务，通过 MapReduce 来实现基本的 SQL 操作。因为 Hive 基于 MapReduce，所以它把容错，执行以及资源管理的工作都交给了 MapReduce 框架，特点是简单与易于实现。但是它也有一些不可避免的缺陷，包括对标准 SQL 以及实时查询的支持，难于优化带来的查询性能低下，并且很难充分利用整个集群的资源，从而导致并发吞吐量较低。

HAWQ[35]（Hadoop with Query）是 Hadoop 领域与 SQL 兼容的大规模数据分析引擎。HAWQ 继承了 Hadoop 与 MPP 大规模数据库分析引擎的优点，实现了 HDFS 分布式存储与 MPP 执行引擎的结合。HAWQ 实现了 MPP 基于统计的优化器，支持数百万连接的网络互联协议，数据的多级划分与存储和高效的执行引擎。特点是与各种 BI 工具的兼容，实时查询的支持，以及与基于 MapReduce 系统相比的性能优势。

Impala 和 Hadapt 是另外两个基于 Hadoop 的 SQL 引擎。基本的出发点也是把 MPP 的技术引入 Hadoop。但目前还不是很成熟。

➢ 访问接口和查询语言

大数据系统的访问接口和查询语言取决于系统的存储模型。传统的 MPP 数据库都使用关系模型，其查询语言为标准的 SQL。而图数据库有自己的查询语

[33] Apache Hadoop. http://hadoop.apache.org/
[34] Apache Hive. http://hive.apache.org/
[35] EMC Pivotal Initiative. Pivotal HD: HAWQ. 2013. http://www.gopivotal.com/sites/default/files/Hawq_WP_042313_FINAL.pdf

言，可以实现子图匹配、路径查询等功能。Hadoop 本身使用的是 HDFS，MapReduce 编程接口可以作为其访问接口。构建在 Hadoop 之上的类数据库系统则提供各自存储模型所对应的查询语言和访问接口。例如，HBase 提供 API，用于对数据表进行 key-value 形式的查询和增删改。Hive 则提供称为 HiveQL 的查询语言，用于对关系表进行查询，HiveQL 同 SQL 非常相似，并附带一些 SQL 未提供的功能。为了方便对 hadoop 的使用，一系列的查询语言和附加访问接口被提出。Pig 是一种基于 MapReduce 的编程平台，它的访问语言 Pig Latin 是介于 SQL 和过程式程序设计语言之间的语言，结合了 SQL 申明式（declarative）语言的优势以及过程式程序设计的灵活性，被众多程序设计者所热衷。Sqoop 是一种用于在关系数据库和 hadoop 之间进行数据迁移的命令式语言。Mahout 则是构建在 hadoop 之上的机器学习引擎，也拥有自己的一套访问接口。

3.2.3 发展趋势

（1） 大数据索引和查询技术

索引和查询技术是数据处理系统的重要入口之一，近年来随着数据量（Volume）、数据处理速度（Velocity）和数据多样性（Variety）的快速发展，大数据相关的索引和查询技术做为大数据的主要入口之一也变得更为重要。传统的索引和查询技术虽然不能很好的解决大数据带来的挑战，然而其核心技术例如数据库和数据挖掘系统中使用的经典索引例如哈希索引、B 树索引、位图索引和 R 树索引，信息检索系统中的倒排索引等依然是大数据索引和查询系统的基石。

大数据带来的主要挑战是其庞大的数据量，单个节点不能或者无法有效地处理这种数量级的数据。此外数据增长速度非常快，这要求系统不但能处理已有的大数据，还要能快速的处理新数据。这些特征使得我们需要考虑很多在大数据环境中独有的因素来开发和选择大数据索引和查询技术。

分布式是处理大数据的一个基本思路，这同样适用于大数据索引和查询系统。分布式索引把全部索引数据水平切分后存储到多个节点上，这可以很好的解决两个问题：1）单个节点无法存储庞大的索引数据；2）单个节点构建索引的效率瓶颈。当业务增长，需要索引更多的数据或者更快的索引数据时，可以通过水平扩展增加更多的节点来解决。切分数据的方式有多种，常见的方法有随机方法、哈

希方法和区间方法。随机方法将所有数据随机分布到不同的节点，这种方法不支持更新操作。哈希方法根据某个列或者某些列（称为分布键）的哈希值将数据分布到不同的节点。区间方法将所有的数据按照不同区间分布到不同的节点。区间到节点的映射信息需要保存下来。不管使用什么样的切分方法，都需要注意数据分布的均匀性，避免大量数据分布到一个或者几个节点上，这样就失去了分布式计算的优势，因而对算法的选择和设计有一定要求。另外分布键的选择也很重要，好的分布键能将数据相对均匀的分布到不同的节点，从而达到负载均衡的目的。

由于索引数据是分布在不同的节点上，因而查询也是分布式的。所有节点或者部分节点的查询结果由主节点（主从架构）或者查询节点（点对点架构）进行汇总，然后得到最终结果。不同的分布式系统支持不同类型的查询语言和查询能力。分布式数据库系统支持 SQL 查询； NoSQL 产品类型和功能各异，有的仅支持主键查询，有的支持范围查询，有的还支持有限的 JOIN；全文检索系统的查询语法最为灵活，但通常不支持 JOIN 或者有限的支持 JOIN。

当一个节点故障时，将无法访问该节点上的数据。为了提高可用性，防止单点故障，通常使用镜像技术或者保存多个副本到不同的节点上。副本可以使用不同的分布策略，例如基于 Hadoop 的系统通常有 2 个副本，一个副本在同机架上的其他节点，另一个副本在其他机架的节点上。这样一方面可以有效利用数据局部性原理改进性能，另一方面可以最大化的保证数据的可用性。有些系统副本仅仅起数据备份的作用，这种类型的副本不能接受查询请求，主要目的是提高系统的可靠性。有的系统的副本还可以处理用户查询请求，从而实现负载均衡以最大化的利用系统资源。然而副本的引入也大大增加了系统的复杂性，因为分布式环境下任何一个节点可能在任何时刻出错：网络可能故障，磁盘可能故障，系统可能崩溃。多数系统采取保证数据高度一致性的策略：只有主副本接受写请求，然后通过文件块复制或者写管道将数据写入到其他副本。也有一些 NoSQL 系统采用最终一致性策略，这种策略中在某一个时刻数据在不同的副本上可能是不一致的，但是当没有对该数据的更新时，最终的访问将返回该数据的最新值。

当系统不能适应业务的需求时，需要对系统进行动态扩容，这通常需要进行数据的再分布，即根据新系统中节点的个数按照数据分布策略重新对数据进行分布。当数据量庞大时，扩容可能需要较多的时间。为了降低需要移动的数据量，

可以采取某些算法来实现，例如一致性哈希算法。

目前各大数据库厂商例如 Oracle、IBM、Greenplum 都已经有支持分布式索引和查询的产品，很多 NoSQL 数据库例如 Mongodb、HBase、Cassandra 也支持分布式索引和查询。还有很多面向全文检索的产品例如 Solr、ElasticSearch、Sphinx 均支持分布式全文索引和查询，且这些产品都是开源的。值得一提的是 Greenplum 的 GPText 将 Solr 的全文检索能力引入到了 Greenplum 数据库之中，使得它可以同时支持 SQL 和 Solr 的全文检索。

（2） 实时/流式大数据存储与处理

随着业务的增长，业界对大数据的速度（Velocity）维度越来越关注，过去需要几天或者几个小时才能回答的问题现在期望在几分钟、几秒甚至毫秒内得到解决。实时流数据存储和处理技术将会越来越多的被研究和开发。实时流式大数据的处理在很多方面和分布式系统在原理上有很多相似之处，然而也有其独特需求。

实时流数据处理系统包括流数据的实时存储和流数据的实时计算。流数据存储指的是快速高效的存储流式数据到数据库、数据仓库或者数据湖中；流数据的实时计算注重对流数据的快速高效处理、计算和分析。

（1） 数据流加载：实时流式大数据系统中，数据通常以流的方式进入系统。如何高效且可靠地将数据加载到大数据存储系统中成为流式大数据系统实现低延迟处理的基础。此外能够重新处理数据流中的数据也是一个很有价值的特性。

（2） 复杂事件处理（CEP）：数据流中的数据源是多种多样的，数据的格式也是多种多样，而数据的转换、过滤和处理逻辑更是千变万化，因而需要强大而又灵活的复杂事件处理引擎来适应各种场景下的需求。

（3） 高可用性：数据通过复杂处理引擎和流计算框架时，通常会经过很多步骤和节点，而其中任何一步都有出错的可能，为了保证数据的可靠性和精准投递，系统需要具有容错和去重能力。

（4） 流量控制和缓存：整个流系统可能有若干个模块，每个模块的处理能力和吞吐量差别很大，为了实现总体高效的数据处理，系统需要对流量进行控制和动态节点增加和删除的能力。当数据流入大于流出的速度时，还需要有一定的缓存能力，如果内存不足以缓存快速流入的数据时，需要能够持久化到存储层。

目前市场上已经出现了多种大数据实时处理技术，它们各有不同的侧重点，

例如数据传输技术有 Flume, Scribe, Kafka, Sqoop 等，计算框架有 Storm，S4，Spark 等。基于 Hadoop 的 SQL 处理引擎有 Impala， HAWQ 等。另外还有一些产品在大数据流计算框架之上提供分析即服务，例如 Cetas。大数据的实时存储与处理还有很多需要研究和解决的问题。

3.3 大数据计算模式与系统

3.3.1 问题与挑战

计算模式的出现有力推动了大数据技术和应用的发展，使其成为目前大数据处理最为成功、最广为接受使用的主流大数据计算模式。然而，现实世界中的大数据处理问题复杂多样，难以有一种单一的计算模式能涵盖所有不同的大数据计算需求。研究和实际应用中发现，由于 MapReduce 主要适合于进行大数据线下批处理，在面向低延迟和具有复杂数据关系和复杂计算的大数据问题时有很大的不适应性。因此，近几年来学术界和业界在不断研究并推出多种不同的大数据计算模式。

所谓大数据计算模式，即根据大数据的不同数据特征和计算特征，从多样性的大数据计算问题和需求中提炼并建立的各种高层抽象（Abstraction）或模型（Model）。例如，MapReduce 是一个并行计算抽象[13]，Bekerley 大学著名的 Spark 系统中的“分布内存抽象 RDD（RDD, a distributed memory abstraction[14]）”，CMU 著名的图计算系统 GraphLab 中的“图并行抽象”（Graph Parallel Abstraction[15]）等。传统的并行计算方法主要从体系结构和编程语言的层面定义了一些较为底层的并行计算抽象和模型，但由于大数据处理问题具有很多高层的数据特征和计算特征，因此大数据处理需要更多地结合这些高层特征考虑更为高层的计算模式。

为了能更清晰地理解不同的大数据计算模式，首先需要梳理出大数据处理中主要的数据特征和计算特征维度，在此基础上进一步梳理目前出现的各种重要和典型的大数据计算模式。大数据处理包括以下典型的特征和维度：

（1） **数据结构特征**：根据数据结构特征大数据可分为结构化/半结构化数据处理与非结构化数据处理。

（2） **数据获取处理方式**：按照数据获取方式，大数据可分为批处理与流式计算（streaming）方式。

（3） **数据处理类型**：从数据处理类型来看，大数据处理可分为传统的查询分析计算和复杂的数据挖掘分析计算。

（4） **实时性或响应性能**：从数据计算响应性能角度看，大数据处理可分为实时/准实时与非实时计算，或者是联机（online）计算与线下（offline）计算。流式计算通常属于实时计算，查询分析类计算通常也要求具有高响应性能，而批处理和复杂数据挖掘计算通常属于非实时或线下计算。

（5） **迭代计算**：现实的数据处理中有很多计算问题需要大量的迭代计算（如一些机器学习算法），为此需要提供具有高效的迭代计算能力的计算模式。

（6） **数据关联性**：MapReduce 适用于处理数据关系较为简单的计算任务，但社会网络等具有复杂数据关系的计算任务则需要研究和使用图数据计算模式。

（7） **并行计算体系结构特征**：由于需要支持大规模数据的存储计算，大数据处理通常需要使用基于集群的分布式存储与并行计算体系结构和硬件平台。此外，为了克服传统的 MapReduce 框架在计算性能上的缺陷，人们从体系结构层面上提出了内存计算模式。

3.3.2 主要进展

根据大数据处理多样性的需求和以上不同的特征维度，目前出现了多种典型和重要的大数据计算模式。与这些计算模式相适应，出现了很多对应的大数据计算系统和工具[16]。由于单纯描述计算模式比较抽象和空洞，因此，在描述不同计算模式时，将同时给出相应的典型计算系统和工具，这将有助于对计算模式的理解以及对技术发展现状的把握，并进一步有利于在实际大数据处理应用中对合适的计算技术和系统工具的选择使用。

表3-1： 典型大数据计算模式与系统

典型大数据计算模式	典型系统
大数据查询分析计算	HBase，Hive，Cassandra，Impala，Shark，Hana 等
批处理计算	Hadoop MapReduce，Spark 等
流式计算	Scribe，Flume，Storm，S4，Spark Steaming 等
迭代计算	HaLoop，iMapReduce，Twister，Spark 等

图计算	Pregel，Giraph，Trinity，PowerGraph，GraphX 等
内存计算	Dremel，Hana，Spark 等

（1）大数据查询分析计算模式与典型系统

由于行业数据规模的增长已大大超过了传统的关系数据库的承载和处理能力，因此，目前需要尽快研究并提供面向大数据存储管理和查询分析的新的技术方法和系统，尤其要解决在数据体量极大时如何能够提供实时或准实时的数据查询分析能力，满足企业日常的经营管理需求[17]。然而，大数据的查询分析处理具有很大的技术挑战，在数量规模较大时，即使采用分布式数据存储管理和并行化计算方法，仍然难以达到关系数据库处理中小规模数据时那样的秒级响应性能。

大数据查询分析计算的典型系统包括 Hadoop[36]下的 HBase 和 Hive，Facebook 开发的 Cassandra，Google 公司的 Dremel[18]，Cloudera 公司的实时查询引擎 Impala；此外为了实现更高性能的数据查询分析，还出现了不少基于内存的分布式数据存储管理和查询系统，如 UC Berkeley AMPLab 的基于内存计算引擎 Spark[14]的数据仓库 Shark[19]，SAP 公司的 Hana 等。

（2）批处理计算模式与典型系统

最适合于完成大数据批处理的计算模式是 MapReduce[13]。MapReduce 是一个单输入、两阶段（Map 和 Reduce）的数据处理过程。首先，MapReduce 对具有简单数据关系、易于划分的大规模数据采用“分而治之”的并行处理思想；然后将大量重复的数据记录处理过程总结成 Map 和 Reduce 两个抽象的操作；最后 MapReduce 提供了一个统一的并行计算框架，把并行计算所涉及到的诸多系统层细节都交给计算框架去完成，以此大大简化了程序员进行并行化程序设计的负担。

MapReduce 的简单易用性使其成为目前大数据处理最为成功、最广为接受使用的主流并行计算模式。在开源社区的努力下，开源的 Hadoop 系统目前已发展成为较为成熟的大数据处理平台，并已发展成一个包括众多数据处理工具和环境的完整的生态系统。目前几乎国内外的各个著名 IT 企业都在使用 Hadoop 平台进行企业内大数据的计算处理。Spark[14]也是一个批处理系统，其性能方面比 Hadoop MapReduce 有很大的提升，但是其易用性方面目前仍不如 Hadoop MapReduce。

[36] Apache Hadoop [EB/OL],http://hadoop.apache.org/ 2013,08,26

（3）流式计算模式与典型系统

流式计算是一种高实时性的计算模式，需要对一定时间窗口内应用系统产生的新数据完成实时的计算处理，避免造成数据堆积和丢失。很多行业的大数据应用，如电信、电力、道路监控等行业应用、以及互联网行业的访问日志处理，都同时具有高流量的流式数据和大量积累的历史数据，因而在提供批处理数据模式的同时，系统还需要能具备高实时性的流式计算能力。流式计算的一个特点是数据运动、运算不动，不同的运算节点常常绑定在不同的服务器上。

Facebook 的 Scribe 和 Apache 的 Flume 都提供了机制来构建日志数据处理流图。而更为通用的流式计算系统是 Twitter 公司的 Storm[37]、Yahoo 公司的 S4[20]、以及 UCBerkeley AMPLab 的 Spark Steaming[21]。

（4）迭代计算模式与典型系统

为了克服 Hadoop MapReduce 难以支持迭代计算的缺陷，业界和学术界对 Hadoop MapReduce 进行了不少改进研究。HaLoop[22]把迭代控制放到 MapReduce 作业执行的框架内部，并通过循环敏感的调度器保证前次迭代的 Reduce 输出和本次迭代的 Map 输入数据在同一台物理机上，以减少迭代间的数据传输开销；iMapReduce[23]在这个基础上保持 Map 和 Reduce 任务的持久性，规避启动和调度开销；而 Twister[24]在前两者的基础上进一步引入了可缓存的 Map 和 Reduce 对象，利用内存计算和 pub/sub 网络进行跨节点数据传输。

目前，一个具有快速和灵活的迭代计算能力的典型系统是 UC Berkeley AMPLab 的 Spark，其采用了基于分布式内存的弹性数据集模型实现快速的迭代计算。

（5）图计算模式与典型系统

社交网络、Web 链接关系图等都包含大量具有复杂关系的图数据，这些图数据规模常达到数十亿的顶点和上万亿的边数。这样大的数据规模和非常复杂的数据关系，给图数据的存储管理和计算分析带来了很大的技术难题。用 MapReduce 计算模式处理这种具有复杂数据关系的图数据通常不能适应，为此，需要引入图计算模式。

大规模图数据处理首先要解决数据的存储管理问题，通常大规模图数据也需

[37] Storm[EB/OL], http://storm-project.net/ 2013.07.21

要使用分布式存储方式。但是，由于图数据的数据关系很强，分布存储就带来了一个重要的图划分问题（Graph Partitioning）。在有效的图划分策略下，大规模图数据得以分布存储在不同节点上，并在每个节点上对本地子图进行并行化处理。与任务并行和数据并行的概念类似，由于图数据并行处理的特殊性，人们提出了一个新的“图并行[4]”（Graph Parallel）的概念。目前已经出现了很多分布式图计算系统，其中较为典型的系统包括 Google 公司的 Pregel[25]，Facebook 对 Pregel 的开源实现 Giraph，微软的 Trinity[26]，Berkeley AMPLab 的 GraphX[27]，以及 CMU 的 GraphLab 以及由其衍生出来的目前性能最快的图数据处理系统 PowerGraph[15]。

（6）内存计算模式与典型系统

Hadoop MapReduce 为大数据处理提供了一个很好的平台。然而，由于 MapReduce 设计之初是为大数据线下批处理而设计的，随着很多需要高响应性能的大数据查询分析计算问题的出现，MapReduce 其在计算性能上往往难以满足要求。随着内存价格的不断下降以及服务器可配置的内存容量的不断提高，用内存计算完成高速的大数据处理已经成为大数据计算的一个重要发展趋势。Spark 则是分布内存计算的一个典型的系统，SAP 公司的 Hana 则是一个全内存式的分布式数据库系统。

3.3.3 发展趋势

近几年来，随着大数据处理和应用需求急剧增长，同时也由于大数据处理的多样性和复杂性，针对以上的典型的大数据计算模式，学术界和业界不断研究推出新的或改进已有的计算模式和系统工具平台，目前主要有以下三方面的重要发展趋势和方向。

（1）主流的 Hadoop 平台改进后将与其他计算模式和平台共存

由于 MapReduce 当初的设计目标主要是针对具有简单数据关系的大数据线下批处理，使得它在系统构架和计算性能上存在不少不足之处，难以适用于那些具有复杂数据关系和复杂计算模式（如迭代计算、图计算等）的大数据处理计算任务。但尽管如此，由于 Hadoop 生态系统已发展成为目前最主流的大数据处理

平台、并得到广泛的使用。考虑到兼容性，目前业界和学术界并不会完全抛弃 Hadoop 平台，而是试图不断改进和发展现有的平台，增加其对各种不同大数据处理问题的适用性。Hadoop 社区正努力扩展现有的计算模式框架和平台，以便能解决现有版本在计算性能、计算模式、系统构架和处理能力上的诸多不足，这正是目前 Hadoop 2.0 新版本“YARN”的努力目标。目前不断有新的计算模式和计算系统出现，预计今后相当长一段时间内，Hadoop 平台将与各种新的计算模式和系统共存，并相互融合，形成新一代的大数据处理系统和平台。

（2）混合计算模式将成为满足多样性大数据处理和应用需求的有效手段

现实世界大数据应用复杂多样，可能会同时包含不同特征的数据和计算，在这种情况下单一的计算模式多半难以满足整个应用的需求，因此需要考虑不同计算模式的混搭使用。

混合计算模式可体现在两个层面上：一是传统并行计算所关注的体系结构与低层并行程序设计语言层面计算模式的混合，例如，在体结构层，可根据大数据应用问题的需要搭建混合式的系统构架，如 MapReduce 集群+GPU-CUDA 的混合，或者 MapReduce 集群+基于 MIC（Intel Xeon Phi 众核协处理系统）的 OpenMP/MPI 的混合模型。

混合模式的另一个层面是大数据处理高层计算模式的混合。比如，一个大数据应用可能同时需要提供流式计算模式以便接受和处理大量流式数据，提供基于 SQL 或 NoSQL 的数据查询分析能力以便进行日常的数据查询分析，提供线下批处理和迭代计算已完成基于机器学习的深度数据挖掘分析；一些大数据计算任务可能还涉及到复杂图计算或者间接转化为图计算问题。因此，很多大数据处理问题将需要混合使用多种计算模式。此外，为了提高计算性能，各种计算模式还可以与内存计算模式混合，实现高实时性的大数据查询和计算分析。

混合计算模式之集大成者当属 UCBerkeley AMPLab 的 Spark 系统，其涵盖了几乎所有典型的大数据计算模式，包括迭代计算、批处理计算、内存计算、流式计算（Spark Streaming）、数据查询分析计算（Shark）、以及图计算（GraphX）。Spark 提供了一个强大的内存计算引擎，实现了优异的计算性能，同时还保持与 Hadoop 平台的兼容性。因此，随着系统的不断稳定和成熟，Spark 有望成为与 Hadoop 共存的新一代大数据处理系统和平台。

（3）内存计算将成为高实时性大数据处理的重要技术手段和发展方向

Hadoop 在处理大数据时计算性能不高、难以满足实时性或高响应性计算任务的要求，为此，人们一直努力改进 Hadoop 的计算性能。但是，在现有 Hadoop 平台面向大数据线下处理的基本构架和工作机制下，性能的改进和提升空间非常有限，难以逾越计算性能低下的障碍；而随着大数据的规模不断扩大，这个问题将越来越为突出。为此，目前已经逐步形成一个基本共识，即随着内存成本的不断降低，内存计算将成为最终跨越大数据计算性能障碍、实现高实时高响应计算的一个最有效技术手段。因此，目前越来越多的研究者和开发者在关注基于内存计算的大数据处理技术，不断推出各种基于内存计算的计算模式和系统。

内存计算是一种在体系结构层面上的解决方法，因此，它可以与各种不同的计算模式相结合，从基本的数据查询分析计算，到批处理和流式计算，再到迭代计算和图计算，都可以基于内存计算加以实现，因此我们可以看到各种大数据计算模式下都有基于内存计算实现的系统，比较典型的系统包括 SAP 的 Hana 内存数据库，微软的图数据计算系统 Trinity，UC Berkeley AMPLab 的 Spark 等。

由于优异的计算性能，内存计算将成为今后高实时性大数据处理的重要技术手段和发展方向。

3.4 大数据分析与挖掘

3.4.1 问题与挑战

大数据时代，不同领域不同格式的数据从生活的各个领域涌现出来。大数据往往含有噪声，具有动态异构性，是相互关联和不可信的。尽管含有噪声，大数据往往比小样本数据更有价值。这是因为从频繁模式和相关性分析得到的一般统计量通常会克服个体的波动，会发现更多可靠的隐藏的模式和知识。另一方面，互相连接的大数据形成大型异构信息网。通过信息网，冗余的信息可用于弥补数据缺失所带来的损失，可用于交叉核对数据的不一致性，进一步验证数据间的可信关系，并发现数据中隐藏的关系和模型。数据挖掘需要集成的、经过清洗的、可信的、可高效访问的数据，需要描述性查询和挖掘界面，需要可扩展的挖掘算法以及大数据计算环境。与此同时，数据挖掘本身也可以用来提高数据质量和可

信度，帮助理解数据的语义，提供智能的查询功能。只有能够鲁棒地进行大数据分析，大数据的价值才能发挥出来。另一方面，从大数据得出的知识有助于纠正错误，并消除歧义[38]。

大数据环境下的分析和挖掘方法与传统的小样本统计分析有着根本的不同，具有如下挑战：

（1） **数据量的膨胀**。随着数据生成的自动化以及数据生成速度的加快，数据分析需要处理的数据量急剧膨胀。一种处理大数据的方法是使用采样技术，通过采样，可以把数据规模变小，以便利用现有的技术手段进行数据管理和分析。然而在某些应用领域，采样将导致信息的丢失，比如 DNA 分析等。在明细数据上进行分析，意味着需要分析的数据量将急剧膨胀和增长[28]。如何对 TB 级的大数据进行分析是一大挑战；

（2） **数据深度分析需求的增长**。为了从数据中发现知识并加以利用进而指导人们的决策，必须对大数据进行深入的分析，而不是仅仅生成简单的报表。这些复杂的分析必须依赖于复杂的分析模型，很难用 SQL 来进行表达，统称为深度分析。人们不仅需要通过数据了解现在发生了什么，更需要利用数据对将要发生什么进行预测，以便在行动上做出一些主动的准备。比如通过预测客户的流失预先采取行动，对客户进行挽留。这里，典型的 OLAP 数据分析操作(对数据进行聚集、汇总、切片和旋转等)已经不够用，还需要路径分析、时间序列分析、图分析、What-if 分析以及由于硬件/软件限制而未曾尝试过的复杂统计分析模型等；

（3） **自动化、可视化分析需求的出现**。因为数据规模很大，要对大数据进行有效分析，分析过程需要按照完全自动化的方式进行。这就要求计算机能够理解数据在结构上的差异，明白数据所要表达的语义，然后“机械”地进行分析。对大数据分析来说，设计一个好的适于分析的数据表示模式是非常重要的。此外，大数据也使下一代可实时应答的交互式数据分析成为可能。将来，系统应该能够根据网站的内容自动构造查询，自动提供热门推荐，自动分析数据的价值并决定是否需要保存。目前，在保证交互式响应的同时如何进行 TB 级的复杂查询处理已成为一个重要的研究课题。

[38] Challenges and Opportunities with Big Data，A community white paper developed by leading researchers across the United States, www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

3.4.2 主要进展

针对上面提到的挑战，研究者提出了一些试验性的解决方法和途径，其中的许多方法具有一定的实际应用价值。例如，针对传统分析软件扩展性差以及 Hadoop 分析功能薄弱的特点，IBM 公司的研究人员致力于对 R 和 Hadoop 进行集成[29]。R 是开源的统计分析软件，通过 R 和 Hadoop 的深度集成，把计算推向数据并且并行处理，使 Hadoop 获得了强大的深度分析能力。另有研究者实现了 Weka（类似于 R 的开源的机器学习和数据挖掘工具软件）和 MapReduce 的集成[30]。标准版 Weka 工具只能在单机上运行，并且不能超越 1GB 内存的限制。经过算法的并行化，在 MapReduce 集群上，Weka 不仅突破了原有的可处理数据量的限制，轻松地对超过 100GB 的数据进行分析，同时利用并行计算提高了性能。经过改造的 Weka，赋予 MapReduce 技术深度分析的能力。另有开发者发起了 Apache Mahout 项目的研究，该项目是基于 Hadoop 平台的大规模数据集上的机器学习和数据挖掘开源程序库，为应用开发者提供了丰富的数据分析功能。

针对频繁模式挖掘、分类和聚类等传统的数据挖掘任务，研究人员也提出了相应的大数据解决方案。如，Iris Miliaraki 等人提出了一种可扩展的在 MapReduce 框架下进行频繁序列模式挖掘的算法[31]，Alina Ene 等人用 MapReduce 实现了大规模数据下的 K-center 和 k-median 聚类方法[32]，Kai-Wei Chang 等人提出了针对线性分类模型的大数据分类方法[33]。U Kang 等人使用“Belief Propagation 算法（简称 BP）”处理大规模图数据发掘异常模式[34]。

另有一些研究针对大规模图数据进行分析。Jayanta Mondal 等人[35]提出了一个基于内存的分布式数据管理系统来管理大规模动态变化的图以支持低延迟的查询处理方法，提出了一种混合的复制（replication）策略来检测结点读写的频率从而动态的决定哪些数据需要复制（replication）。Shengqi Yang 等人[36]对基于集群上的大规模图数据管理和局部图的访问特征（广度优先查询和随机游走等）进行研究，为了在图查询处理中减少机器间通讯，提出来分布式图数据环境，同时提出了两级别划分管理架构。Jiewen Huang 等人提出了一个多节点的可扩展 RDF 数据管理系统，比目前系统的效率高出 3 个数量级。

3.4.3 发展趋势

（1） 更加复杂、更大规模的分析和挖掘。在大数据新型计算模式上实现更加复杂和更大规模的分析和挖掘是大数据未来发展的必然趋势。例如，需要进行更细粒度的仿真、时间序列分析、大规模图分析和大规模社会计算等等。另一方面，在大数据上进行复杂的分析和挖掘，需要灵活的开发、调试、管理等工具的支持。

（2） 大数据的实时分析和挖掘。面对大数据，分析和挖掘的效率成为此类大数据应用的巨大挑战。尽管可以利用大规模集群并行计算，以 MapReduce 为代表的并行计算模型并不适合高性能的处理结构化数据的复杂查询分析。在数十 TB 以上的数据规模上，分析和发掘的实时性受到了严峻的挑战，是目前尚未彻底解决的问题。而查询和分析的实时处理能力，对于人们及时获得决策信息，做出有效反应是非常关键的前提。

（3） 大数据分析和挖掘的基准测试。各种大数据分析和挖掘系统各有所长，其在不同类型分析挖掘下，会表现出非常不同的性能差异。目前迫切需要通过基准测试，了解各种大数据分析和挖掘系统的优缺点，以明确能够有效支持大数据实时分析和挖掘的关键技术，从而有针对的进行深入研究。

3.5 大数据可视化分析

3.5.1 问题与挑战

在大数据时代，数据的数量和复杂度的提高带来了对数据探索，分析，理解和呈现的巨大挑战。除了直接的统计或者数据挖掘的方式，可视化通过交互式视觉表现的方式来帮助人们探索和解释复杂的数据。一个典型的可视化流程是首先将数据通过软件程序系统转化为用户可以观察分析的图像。利用人类视觉系统高通量的特性，用户通过视觉系统，结合自己的背景知识，对可视化结果图像进行认知，从而理解和分析数据的内涵与特征。同时，用户还可以交互地改变可视化程序系统的设置，改变输出的可视化图像，获得对数据的不同侧面的理解。因此，可视化是一个交互与循环往复的过程。

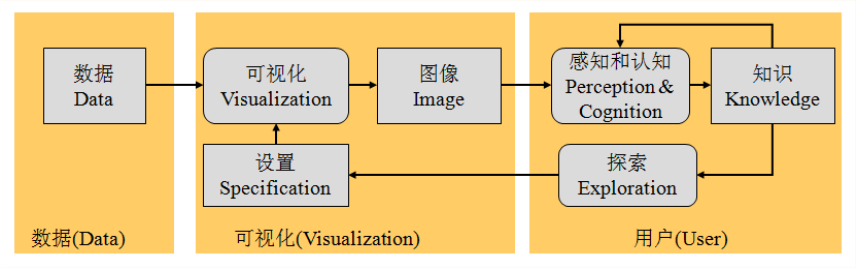

图3-1： 可视化流程（源自[37]）

可视化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，可视化所提供的洞察力有助于使用者更快更好地从复杂数据中得到新的发现，这使得可视化成为数据科学中不可或缺的重要部分

人类对于数据对象通过作图的方式帮助理解分析古已有之。例如古人的地图和星图，早期物理学家对实验结果的绘图。现代意义上的可视化源自于计算机技术的发展，首先是对于科学数据的可视化，其后扩展到更广泛的信息可视化。二十一世纪开始后，随着反恐等需求，对于海量、复杂数据的分析进一步催生了可视分析，通过可视界面，结合人机交互和背景自动数据分析挖掘，对海量复杂数据开展分析。

3.5.2 主要进展

在可视化的发展中，首先面对大规模数据挑战的是在科学可视化方向。高通量仪器设备、模拟计算以及互联网应用等都在快速产生着庞大的数据，对 TB 乃至 PB 量级数据的分析和可视化成为现实的挑战。大规模数据的可视化和绘制主要是基于并行算法设计的技术，合理利用有限的计算资源，高效地处理和分析特定数据集的特性。很多情况下，大规模数据可视化的技术通常会结合多分辨率表示等方法，以获得足够的互动性能。在科学大规模数据的并行可视化工作中，主要涉及数据流线化(Data Streaming)，任务并行化(Task Parallelism)，管道并行化(Pipeline Parallelism)和数据并行化(Data Parallelism)四种基本技术[38]：。

数据流线化将大数据分为相互独立的子块后依次处理。在数据规模远大于计算资源时是主要的一类可视化手段。它能够处理任意大规模的数据，同时也可能提供更有效的缓存使用效率，并减少内存交换。但通常这类方法需要较长的处理时间，难以提供对数据的交互挖掘。离核渲染是数据流线化的一种重要形式。在

另外一些情况下，数据则是以流的形式实时逐步获得，必须要有能够适应数据涌现形式的可视化方法。

任务并行化是把多个独立的任务模块平行处理。这类方法要求将一个算法分解为多个独立的子任务，并需要相应的多重计算资源。其并行程度主要受限于算法的可分解粒度以及计算资源中节点的数目。管道并行化则是同时处理各自面向不同数据子块的多个独立的任务模块。以上任务并行化和管道并行化两类方法，如何达到负载的平衡是实现高效分析的关键难点。

数据并行化是将数据分块后进行平行处理，通常称为单程序多数据流(SPMD)模式。这类方法能达到高度的平行化，并且在计算节点增加的时候可以达到较好的可扩展性。对于非常大规模的并行可视化，节点之间的通讯往往是制约因素，提供合理的通讯模式是高效结果的关键，而提高数据的本地性也可以大大提高效率。以上这些技术往往在实践中相互结合，从而构建一个更高效的解决方法。

在信息可视化和可视分析方面，相对对大规模数据的处理出现的相应要晚得多。很多技术，例如多维数据可视化中的平行坐标，多尺度分析，散点图矩阵，层次数据可视化中的树图，图可视化中的多种布局算法，文本可视化的一些基本方法，并不是都有很好的可扩展性。在面对大数据挑战的可视化中，需要做出相应的调整。

传统对网络数据的可视化可以通过图的形式实现，这是将网络中的每个节点简化为图中的节点，网络中的联系可视化为图中的边，这样网络数据的可视化可以通过经典的节点－边的形式表现。这类可视化方法的难点主要在于图的排布算法。有效的图布局应该能够直观地揭示节点之间的联系，类似的、相互联系紧密的节点会聚集在一起。但是现在大规模的网络数据的节点可能高达数百万，其边可能高达数亿，这样的网络数据难以使用传统的图可视化方法可视化。

高维信息可以通过维度压缩、平行坐标等手段实现可视化。但是在数据达到一定规模以后，这样的方法并不能很好扩展。一些可能的方案包括提供一些子空间的选择，用户可以根据分析需要，在高维度空间选择适合问题解决的子空间，从而缩小数据规模。

图形硬件对于大规模数据可视化具有重要意义。最新的超级计算机大量地应用 GPU 作为计算单元。如何更好发掘最新的图形硬件潜力，提供更加灵活的大

数据可视化和绘制的解决方法是具有重大意义的课题。

3.5.3 发展趋势

面对大数据，结合国际学者的各种观点[39]，相应的大数据可视化与分析也面临着各种挑战：

（1） 原位分析（In Situ Analysis）。传统的可视化方式是先将数据存储于磁盘、然后根据可视化的需要进行读取分析。这一种处理方式对于超过一定量级的数据来说并不适合。最初是对应与 exascale 规模的超级计算机计算获得的大量科学数据产生的挑战，I/O 几乎成为无法克服的困难。科学家提出了原位可视分析的概念，在数据仍在内存中时就会做尽可能多的分析。数据在进行了一定的可视化（同时也是数据规模的简化），能极大地减少 I/O 的开销，只有极少数的视觉投影后的次生数据需要转移到显示平台。这个方法可以实现数据使用与磁盘读取比例的最大化，从而最大限度地克服 I/O 的瓶颈限制。然而，它也带来了一系列设计与实现上的挑战，包括交互分析、算法、内存、I/O、工作流和线程的相关问题。原位分析要求可视化方案和计算紧密结合，这样很多传统的可视化方法需要进行修改或者筛选才可以用于这样的可视化模式。由于可视化的一部分处理在计算核点上进行，那样就会对可以进行的处理方案有所限制。

（2） 大数据可视化中的人机交互。在可视化和可视分析中用户界面与交互设计扮演着越来越重要的角色。用户必须通过合理的交互方式，才可以有效地探索发现数据中的隐含信息，进行可视推理，通过意义构建，获得新的认知。然而尽管数据规模和机器的计算能力都在持续快速地增长，千百年来人的认知能力却是始终不变的。以人为中心的用户界面与交互设计面临的挑战是复杂和多层次的，并且在不同领域都有交叠。机器自动处理系统对于一些需要人类参与判断的分析过程往往表现不佳。其他的挑战则源于人的认知能力，现有技术不足以让人的认知能力发挥到极限。我们需要提供更好的人机交互界面和设计，方便使用者，特别是专家用户能够最大程度发挥其背景知识，在数据的分析中扮演更加积极的角色。从更广泛的一个意义上说，可视化可以建立一个可视的交互界面，提供人和数据的对话。

（3） 协同与众包可视分析。在大数据时代，个人或者少数几个分析用户可

能无法面对数据规模和复杂度带来的挑战。大数据分析中往往设计多种不同来源甚至领域的数据。利用众人的智慧，通过众包等模式进行有效的复杂可视化成为一种必然的选择。在众包可视化工作中，如何设计合理高效的可视化平台，承载相应的复杂高难度的可视化系统工作；如何设计交互的中间模式，支持多用户的协调工作；如何反映多用户的差别，都是可以研究的课题。和协同的可视分析方式比较，协同可视化趋于少数的几个领域专家交互合作开展对数据的可视分析，众包可视化则更趋向不特定多数的使用者，规模也更大。如何开展有效的众包和协同可视化，是非常重要的研究课题。

（4） 可扩展性与多级层次问题。在大规模数据可视分析的可扩展性问题上，建立多级层次是主流的解决办法。这种方法可以通过建立不同大小的层面，提供用户在不用解析度下的数据浏览分析能力。但是当数据量增大时，层级的深度与复杂性也随之增大。在继承关系复杂且深度大的层次关系中巡游与搜索最优解是可扩展性分析的主要挑战。

（5） 不确定性分析和敏感性分析。不确定性的量化问题可以追溯到由实验测量产生数据的时代。如今，如何量化不确定性已经成为许多领域的重要问题。了解数据中不确定性的来源对于决策和风险分析十分重要。随着数据规模增大，直接处理整个数据集的能力也受到了极大的限制。许多数据分析任务中引入数据的不确定性。不确定性的量化及可视化对未来的大数据可视分析工具而言极端重要，我们必须发展可应对不完整数据的分析方法，许多现有算法必须重新设计，进而考虑数据的分布情况。一些新兴的可视化技术会提供一个不确定性的直观视图，来帮助用户了解风险，从而帮助用户选择正确的参数，减少产生误导性结果的可能。从这个方面来看，不确定性的量化与可视化将成为绝大多数可视分析任务的核心部分。另一方面，对于可视化而言，用户的交互或者新的参数的输入，都会导致不同可视化结果的出现。在大数据的情况下，向用户提供背景知识，告知预期的操作可能引发的可视化结果的变化程度，或者用户当前所在参数空间的周边状况，这一些都属于对可视分析结果的敏感性分析，对于高效的可视化交互是极端重要的。

（6） 可视化与自动数据计算挖掘的结合。可视化提供了用户对数据的直观分析，用户可以通过交互界面对数据进行分析了解。同时，我们要注意到很多的

数据分析是批量的。我们如何能够将一些比较确定的分析任务利用机器自动完成，同时引导用户来进行更具有挑战性的可视分析工作，是可视分析发展中的核心课题。

（7） 面向领域和大众的可视化工具库。提供相应的工具库可以大大提高不同领域分析数据的能力。

大数据时代涌现并推动了很多可视化商业化的机会。Tabealu 的成功上市反映了市场对可视化工具的需求。类似 IBM Manyeyes 这样在线可视化工具的流行，则表明在一定程度上满足了广大普通用户对可视化方法的需求。国际的几个大公司也在开展相应的研究，企图把可视化引入其不同的数据分析和展示的产品中。各种可能相关的商品也将会不断出现，对可视化服务的商业需求将是未来的一个最大方向。

3.6 大数据隐私与安全

3.6.1 问题与挑战

隐私[40]是当事人不愿意被他人知道或他人不便知道的敏感信息，它与公共利益、群体利益无关，具有隐藏特性。安全[41]是指不受威胁，没有危险、危害、损失。信息安全[42]是指采取技术和管理的安全保护手段，保护软硬件与数据不因偶然的或恶意的原因而遭到破坏、更改、显露。

在大数据时代，传统的隐私数据内涵与外延有了巨大突破与延伸，隐私数据保护不力所造成的恐慌已不能由个人或团体承受，隐私数据保护技术面临更多的挑战。大数据时代下的隐私数据保护与安全体系除涉及技术、管理外，还涉及法律、人伦、生物、道德、商业利益、生活方式等；不只是团体或区域，还涉及到国家安全与国际秩序[43, 44]。隐私数据泄露影响的波及面很可能会突破个人、团体或区域的限制，发展到全球性影响。

从本质上来说，大数据的安全与隐私问题就是我们要能够在大数据时代兼顾安全与自由，个性化服务与商业利益，国家安全与个人隐私的基础上，从数据中挖掘其潜在的巨大商业价值和学术价值，并使其研究成果真正的服务于社会。

在大数据时代，随着我们对大数据的进一步认识和研究，呈现出的安全隐私

挑战有以下几个方面[44]：

（1） 大数据时代的安全与传统安全相比，变得更加复杂：一方面，大量的数据汇集，包括大量的企业运营数据、客户信息、个人的隐私和各种行为的细节记录。这些数据的集中存储增加了数据泄露风险，而这些数据不被滥用，也成为人身安全的一部分。另一方面，大数据对数据完整性、可用性和秘密性带来挑战，在防止数据丢失、被盗取和被破坏上存在一定的技术难度，传统的安全工具不再像以前那么有用。

（2） 使用数据过程中的安全问题：用数据挖掘和数据分析获取商业价值[43]的时候，黑客也可以利用大数据分析向企业发起攻击。黑客可能会打限度地收集有用信息，如社交网络、邮件、微博、电子商务、电话和家庭住址等，使得数据安全异常严重。

（3） 对大数据分析较高的企业和团体，面临更多的安全挑战：对于电子商务、金融、天气预报的分析预测、复杂网络计算和广域网感知等领域，恶意性攻击会造成更会严重的后果。

（4） 基于位置的隐私数据暴露严重[46]：个体用户的移动设备的广泛使用，如手机，移动 GPS 设备等，以及一些网站获取用户位置信息等可以很容易得到用户的移动轨迹。而根据研究发现，用户的移动模式和用户身份识别之间有着强烈的对应关系，使得用户的隐私很容易暴露。同时，用户的位置信息保护比用户的身份信息保护更具有挑战性，因为我们在获取数据时要保证较高的精度。

（5） 缺乏相关的法律法规保证：目前为止，还没有严格的法律法规来保证用户的数据隐私安全[47]。特别是一些涉及用户敏感数据的一些记录，而这些数据也容易被一些非法和不道德组织或个体使用，对用户和社会造成严重的影响和损失，例如，频繁发生的互联网公司数据库泄露事件，特别是 2013 年曝光的美国国家安全局“棱镜计划”监听项目。

（6） 大数据的共享问题：共享问题的主要本质是数据的加密性和数据的有效性之间的矛盾。从社会应用角度考虑，我们会尽可能的提高数据的获取技术，以保证数据的有效性，而从保护用户隐私的角度考虑，我们有必要对数据进行相关操作以降低获取数据的敏感性，从而造成了两者之间的矛盾，两者之间如何进行最佳折中确实非常困难。

（7） 真实数据的动态性变化：真实性的大数据随着时间呈现出动态变化性，使得我们对于大数据的分析计算提出了一些新的方法和技术[48,49]，因而在处理时将面对更为复杂的形式，加大了大数据安全隐私保护的困难。

（8） 多元数据的融合挑战：大数据来自于生活，学术，商业等各个方面，而数据之间的彼此相关性，使得数据的安全隐私保护更为复杂，如何在多元数据融合的大趋势下保证用户的隐私不被泄露是一项重大挑战。

3.6.2 主要进展

数据的安全与隐私问题近年来一直是国内外学者关注的重大研究课题，并且针对不同的应用和数据类型都有相关的研究成果，总的来说，目前所拥有的方法有：

（1） 文件访问控制技术：通过文件访问控制来限制呈现对数据的操作，在一定程度解决数据安全问题。

（2） 基础设备加密：其本质是对大数据的存储设备进行安全防护，但不能解决大数据安全的本质问题。

（3） 匿名化保护技术：匿名化技术适用于各类数据和众多应用，并且算法通用性高，能保证发布数据的真实性，实现简单。匿名化过程不可逆，如决策分类器的构建，聚类等应用，如 k-匿名模型，m-invariance 等。但匿名化技术对隐私保护效果并不明显，使得隐私泄露可能性很大，

（4） 加密保护技术：加密保护技术能够保证数据的真实性，可逆性和无损性，对隐私保护程度很高，主要应用与分布式下的数据挖掘和操作，如 SMC 模型，分布式关联规则挖掘算法[50]，差分隐私等。但是该技术的计算开销很大，对大数据的支持不大适用。

（5） 基于数据失真的技术：该技术可应用与关联规则的挖掘和隐藏等，如随机干扰，随机化，阻塞，凝聚等。数据失真技术的实现比较简单，但会造成数据的偏差，可能造成数据价值的丧失，

（6） 基于可逆的置换算法：可逆的置换算法可以保证数据的真实性，并且效率比较高，常用于数据中心的大规模系统隐私保护，如位置变换，映射变化等。但该技术对于安全隐私保护力度仍然不够充分。

3.6.3 发展趋势

随着大数据的不断发展和研究，其巨大价值在被不断挖掘的过程中，数据的安全和隐私发展呈现出新的发展趋势和挑战：

（1） NoSQL 有待进一步完善：迎合了大数据的时代，适合非结构化数据的存储和分析，有灵活、可扩展性强、降低复杂性等特点，但是在安全保护上有待进一步提高。

（2） 针对于 APT 的攻击：在大数据时代，我们在利用数据来获取价值，APT 的攻击隐藏在数据内部，很难被我们发现，所以专门针对于 APT 攻击的研究是非常很总要的。

（3） 大数据的迅速发展和数据量的急剧增加及急速的动态变化，使得我们在对数据的操作时所面临的的安全问题更加严重。

（4） 数据的多元化与彼此的关联性进一步发展，深度挖掘技术，分析方法，算法模型的进一步优化和提高，使得对单一数据的安全隐私保护方法变得极其脆弱，需要针对多元数据融合提出新的安全隐私保护技术。

（5） 针对目前的大数据计算，主要采取的是分布式计算方法。而采用分布式计算的时候必然面临着数据传输，信息交互等过程，如何在这个过程中保护数据价值不泄露，信息不丢失，保护所有站点的安全与分布式系统的隐私是大数据发展面对的重大挑战。

（6） 目前，社交网络成为现代生活不可或缺的部分，一般来说，社交网络都会获取个体用户的位置信息（如 facebook，新浪微博等），以及此网络的迅速动态变化和实时交互等性质，使得我们对网络的安全加密与数据保护更为困难，而作为目前迅速发展起来的社交网络，我们需要进一步加强此方面的安全隐私保护。

（7） “三权分立”的模式应成为一种趋势，即数据的采集过程保护，存储管理保护，以及数据的分析使用过程的安全保护需要由不同的管理决策者来执行，这样可以在一定程度上保护大数据的安全隐私。

最后，大数据的保护需要学术界、商业界以及政府部门的共同参与，需要形成有效的安全机制和国家法律法规来约束和保护大数据的安全隐私，从而保证大数据时代的健全、安全发展。

第四章 大数据 IT 产业链与生态环境

4.1 大数据国内外相关产业现状

大数据的需求和应用的发展需要完整的 IT 产业链，并构建良好的产品体系和上下游生态环境。本部分内容将具体介绍我国的大数据相关产业情况。

4.1.1 大数据产业链全景图

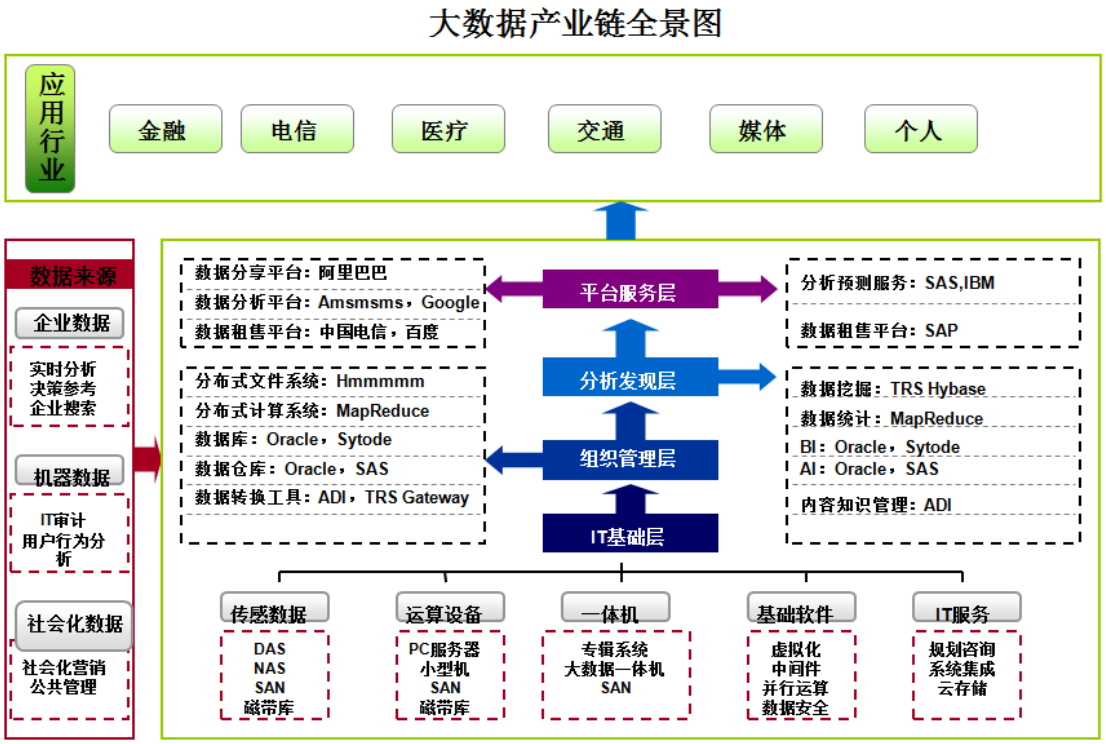

图4-1： 大数据产业链全景图网络生活

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

4.1.2 国内外发展呈现的四个趋势

➢ 开源软件&产业垂直整合

需求驱动致使越靠近消费端的企业，在整个产业链上会拥有越来越多的发言权。大数据时代开源技术的发展已经可以和商用软件分庭抗礼，传统的操作系统、中间件、数据库等平台级软件的同质化趋势已经渐趋明显。最终用户的关注焦点集中如何解决企业的业务问题，而不是购买谁的数据库或者操作系统。因此，越靠近最终用户的企业，将在产业链中拥有越大的发言权。

开源软件加剧了基础软件的同质化趋势，而软、硬件一体化的趋势，进一步弱化了产业链上游的发言权。垂直整合推动大数据产业集约化的发展道路，从而最大限度的获得商业利润。

➢ 非结构化大数据处理分析成为难点和重点

随着互联网和通信技术的迅猛发展，数据类型早已不是单一的以的结构化数据，还充斥着广泛存在于社交网络、物联网、电子商务等之中网络日志、音频、视频、图片、地理位置信息等等多类型的数据。这些数据被命名为非结构化数据。据统计， 85%的数据属于非结构化数据。这些非结构化数据的产生往往伴随着社交网络、移动计算和传感器等新的渠道和技术的不断涌现和应用。但是现有的数据处理方法仅适用于结构化数据，无法将大量的非结构化数据与结构化数据进行统一、整合，从而就无法发掘数据中的价值。

目前国内在非结构化大数据挖掘分析方面，在社会化计算领域，针对于微博数据取得一定的实用性进展，并形成了一定的市场规模。代表产品，如 TRS SMAS（Social Media Analyze Service）社会化媒体云服务平台，它是建立在 TRS 大数据分析挖掘系统基础上的大型在线服务平台，该服务面向政府、企事业单位和个人，以在线云服务的方式提供信息监测、统计分析、关系挖掘、传播效果评估等一系列服务，范围涵盖微博 SNS、网络媒体、论坛博客等全媒体，功能支持事前预警、事中分析和事后处理，为互联网信息的全面分析构建了完整的生态链条。TRS SMAS 平台在大数据的智能挖掘、热点分析方面具有领先的技术优势和数据优势，可从复杂的社会关系中挖掘出有价值的知识，并通过在线方式，即买即开通，为客户提供必要、关键和友好的应用服务。

➢ 机器数据挖掘成为一个重要的发展方向

大数据中，机器数据是最大且增长最快的一部分。每个现代企业机构，无论规模大小，都会产生海量的机器数据，利用这些数据是目前机构或企业的关键任务。

机器数据是由机器产生的数据，也是大数据最原始的数据类型，包括了日志文件、各种历史记录、Web 服务器日志等。它们会由网络交换机、企业应用系统、网络以及安全设备等产生。

面向物联网、电子商务、医疗，电信、金融领域，机器数据结合 IT 运维、系统安全、搜索引擎、电子商务等特定应用的需求实现大数据环境下的机器数据的存储、管理、检索和分析。

目前国外有代表性机器数据挖掘厂商为 Splunk，Splunk 针对 IT 运维、信息安全、交易分析等方面提供业界领先解决方案与产品。通过运用专利数据分析技术，提供多种产品以满足各行各业用户在关键业务的运营保障、安全确保及业务分析方面的需求。目前国内一些厂商也在开发类似的机器挖掘产品，填补国内空白。

➢ 大企业的定制化解决方案

大数据软件技术起源于以下国外 Google、Yahoo 等巨头的分布式计算平台，并随着这些技术的开源基础架构，在国内互联网公司中得到广泛定制化应用。所以目前大数据软件和应用的特点体现出开源和多样性的特点。一些拥有海量数据的大企业业，并没有互联网公司那样的大数据系统部署能力，因此这一需求推动了大数据标准化和产品化解决方案市场的发展。

在国外市场，已经出现以提供企业级大数据软件产品的公司，如 Cloudera。Cloudera 于 2008 年创建，Cloudera 通过提供基于 Hadoop 企业版大数据解决方案，四年的时间里，Cloudera 已经从一家默默无闻的公司发展为解决大数据问题不得不依靠的公司。其在一定程度上与将近一半的财富 50 强企业合作，其中包括 AOL、哥伦比亚广播公司、eBay、Expedia、摩根大通、Monsanto、诺基亚、RIM 和迪斯尼等。

在国内，一些厂商也把海量非结构化信息处理技术和 Hadoop 架构进行有效结合集成，并结合企业在大数据采集、存储、分析挖掘、可视化方面的具体需求，

开发企业级大数据分析挖掘系统。推动大数据分析系统在企业的落地。

4.2 大数据产学研合作相关社区、开源组织、行业协会

4.2.1 大数据相关社区及开源组织

大数据的相关社区及开源组织主要涉及大数据共享、大数据技术交流等方面。其中，大数据共享是产学研更为关注的焦点。

为了促进大数据共享，众多国家和地区正在建立数据开放网站，政府支持的网站如：美国的 www.data.gov 网站[39]，香港的公共数据开放网站[40]，肯尼亚的公共数据开放平台[41]。由企业建立的数据共享网站如：DataMarket[42]、InfoChimps[43]、Import.IO[44]、Factual[45]等。

与欧美等发达国家相比，国内的数据共享起步较晚。之前，在科技部的支撑下，国内已经建立了许多领域的数据共享网站，如气象科学数据平台，地震科学数据共享中心，林业科学数据平台，农业科学数据共享中心，海洋科学数据平台，人口与健康科学数据共享平台，地球系统科学数据共享平台等。国家统计局也于 2013 年 9 月正式开放了新版数据库“国家数据”[46]，向普通公众共享了各项统计数据。

除了政府支持的数据共享网站外，国内企业也开始建设开源的数据共享网站，如数据堂[47]是一个由企业建设的开源数据共享网站，通过产学研合作，已搜集并免费公开了 4 万多组以计算机学科为主的数据集，合作的高校、科研机构及企业数量有上百家。

为了促进产学研的紧密结合，部分国内研发企业也搭建了自己的数据共享平台，向外界公开自己的工业数据，如搜狗实验室[48]定期向外界免费公开中文信息处理方面的数据。百度公司也创建了一个融合技术资源的开放技术社区——百度

[39] Data.gov. www.data.gov
[40]香港政府一站通. data.one.gov.hk
[41] Open Kenya. www.opendata.go.ke
[42] DataMarket. www.datamarket.com
[43] InfoChimps. www.infochimps.com
[44] Import.IO. www.import.io
[45] Factual. www.factual.com
[46]国家数据. data.stats.gov.cn
[47]数据堂. www.datatang.com
[48]搜狗实验室. www.sogou.com/labs

开放研究社区[49]，这个平台上提供了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，并为科研人员提供一个互动交流的平台，促进研发人员间的技术沟通。

现在的数据共享仍然是“各自为政”，如何建立全学科的国家级数据共享平台具有重要的战略意义。目前，数据共享联盟的模式已经被学者们提出并正在筹建，将成为未来数据共享的核心环节。

此外，为了促进大数据的技术交流，国内企业及高校也开始搭建大数据的交流社区。如IBM公司的大数据学院社区[50]主要针对IBM公司的大数据产品提供技术交流，北京理工大学的大数据论坛[51]是一个综合性的大数据讨论社区。

目前，大数据的交流社区仍处于起步阶段，用户量及活跃度不高，尚缺乏一个权威、有影响力的大数据交流社区。

4.2.2 大数据行业协会

IT 技术的迅速发展使大数据成了各行各业共同面对的问题。为了有效应对大数据引起的挑战，同时充分利用大数据带来的机遇，从2012年9月起，中国计算机学会、中国通信学会等纷纷成立了相应的大数据专家委员会。

中国计算机学会大数据专家委员会于2012年10月正式成立。委员会宗旨包括三个方面：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。大数据专家委员会目前下设有五个工作组，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战略材料的编写工作。

中国通信学会大数据专家委员会已于2012年9月成立，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，中

[49]百度开放研究社区. openresearch.baidu.com
[50] IBM 大数据学院. bigdata.db2china.net
[51]北理工大数据论坛. www.bigdatabbs.com

国工程院院士邬贺铨、倪光南、李国杰、何新贵和中国科学院院士怀进鹏以及来自政府部门、学术界、研究机构和企业的知名专家学者担任委员。

4.3 数据生产、数据共享与隐私保护等相关政策与法规

4.3.1 大数据政策法规概述

在大数据时代背景下，越来越多的企业希望借助数据存储、数据分析等为自身带来更多利益，由此引发包括数据版权纠纷、用户隐私泄露等在内的一系列问题。一直以来，大数据在隐私方面都存在巨大的挑战。一方面，随着基于云计算的应用以及社交网络的普及，无论是围绕企业销售，还是个人的消费习惯，身份特征等，都变成了以各种形式存储的数据；另一方面，立法滞后、网络监管不力也使各类数据泄露事件处于无法可依的状态。如何激励数据的生产、共享和利用，又尽可能避免引发的一系列问题，需要从国家层面建设完善健全的大数据政策和法规。

4.3.2 数据生产的相关政策与法规

在互联网数据生产上，主要遵循《互联网信息服务管理办法》[52]，主要限制互联网信息服务提供者不得制作、复制、发布和传播危害国家、社会秩序、不健康信息或侵害他人权益的信息。

针对企业或机构自行采集或生产数据，在数据未公开或利用前，没有明确的政策及法规。相关政策法规主要集中在数据的利用上。当数据被公开、泄露或不正当利用时，需要符合相关的政策法规。

4.3.3 数据共享的相关政策与法规

在数据共享上，国外发达国家的各政府部门都制定了相应的“研究数据政策”，对科学数据的保存与管理等作了明确具体的规定。美国对政府拥有和政府资助生产的数据采用“完全与开放”的共享政策，《信息自由法》和《版权法》规定不

[52]互联网信息服务管理办法（国务院令第 292 号）

允许联邦政府拥有版权，信息服务的收费最多不超过服务本身的成本，对信息的二次开发利用没有限制。欧洲国家主要采取“成本回收模式”发布共享数据，有关部门将持有的信息采取有偿共享或商业化运作方式，从市场上收回数据创建和收集的成本，相关法规有 1995 年通过的《数据库法律保护指令》。

我国政府也非常重视科学数据的管理与共享，编制了“科学数据共享工程建设规划”，制定了《科学数据共享条例》、《国家科技计划项目科学数据汇交办法》、《科学数据共享工程管理办法》、《科学数据共享工程试点遴选和检查评估办法》和《科学数据分类分级共享及其发布策略》等一系列数据共享的政策法规。然而，与国外发达国家相比，我国科学数据共享的政策还不够完善，已制定的相关条例法规缺少相应的法律效力，限制了科学数据的广泛共享[51]。

在数据的版权方面，根据《世界知识产权组织版权条约》及我国著作权法规定，数据内容不管是否受版权保护，只要其内容的选取或编排构成智力创造，便享有版权。但在大数据时代下，许多数据的版权鉴定依然模糊，缺乏清晰完善的法规作为指导。

4.3.4 隐私保护的相关政策与法规

Web 技术的发展使得搜集各种用户个人信息变得更加容易，用户经常很难控制他们个人信息的收集、存储、利用甚至出售。隐私问题是大数据的主要挑战。

欧洲国家的隐私法律相对健全，德国在 1997 年通过了一项与隐私有关的法律是《信息和通信服务法》，用来处理电讯业的个人数据保护。英国在 1998 年颁布了《数据保护法》，对个人权利进行了加强，如禁止不经个人允许将个人数据用于出售。欧盟在 1995 年颁布了《欧盟数据保护指令》，标准化了数据隐私的保护。

国内还没有个人数据保护方面的立法，也没有专门的个人隐私法。我国现行法律对隐私权的保护较为滞后，还没有形成完整的法律体系，仅在一些相关的法律中有些零散的规定。《宪法》第 38、39、40 条分别规定了公民的人格尊严、住宅、通信秘密权不受侵犯。我国《民法通则》也未将隐私权作为一项独立的人格权，而是将其作为名誉权的一部分来进行保护[53]。2001 年 2 月 26 日最高人民法

[53]最高人民法院《关于贯彻执行〈中华人民共和国民法通则〉的若干意见》（试行）

院《关于确定民事侵权精神损害赔偿责任若干问题的解释》则将隐私权作为一项独立的人格利益，但是，该解释仍未从法律上确立隐私权作为一项独立民事权利的地位。针对网络隐私权，仅在一些相关法律中涉及到了这一问题，如《计算机信息网络国际联网管理暂行规定实施办法》[54]、《计算机信息网络国际联网安全保护管理办法》[76][55]、《中华人民共和国电信条例》[56]。但从总体上看，隐私权还没有成为我国法律体系中一个独立的人格权，对隐私权的保护以及侵害隐私权的诉讼也没有形成专门的法律制度，在执行上经常难以具体操作。

4.4 大数据产业链的创新与瓶颈

4.4.1 大数据产业的创新发展

大数据产业的创新发展,主要靠两股力量的推动，如图 4-2 所示，第一以政府主导的社会公共管理和服务，涉及到医疗、交通、环保、科技服务等与社会公共管理和民生有关的行业；第二是以互联网为代表的自由消费数据市场，涉及到金融、证券、电力、电子商务、互联网等行业，这一部分以自由市场化驱动为主体。

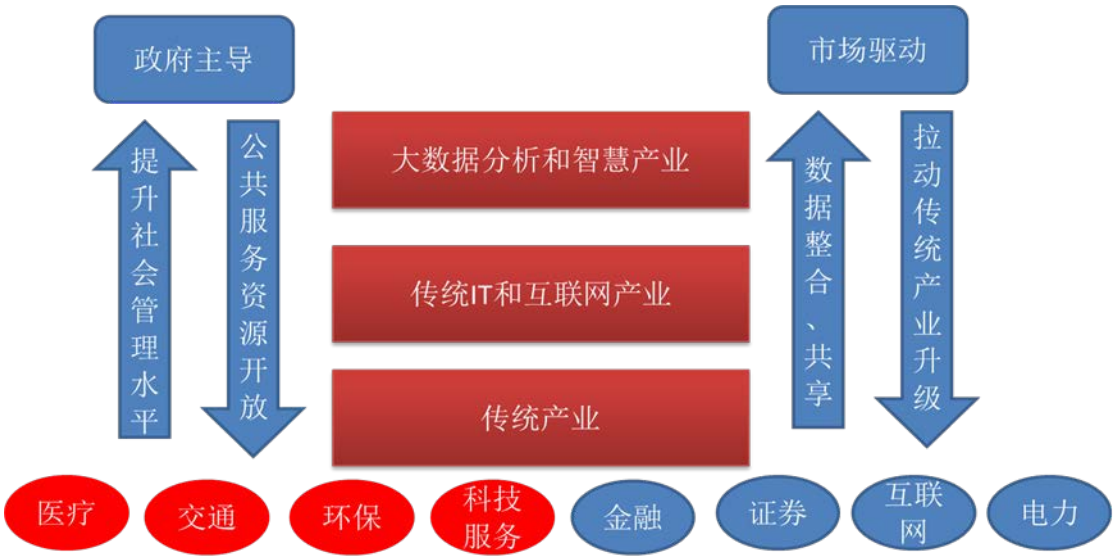

图4-2： 推动大数据产业发展的两股力量

[54]《中华人民共和国计算机信息网络国际联网管理暂行规定》（国务院令第 195 号）
[55]《计算机信息网络国际联网安全保护管理办法》（公安部令第 33 号）
[56]《中华人民共和国电信条例》中华人民共和国国务院令（第 291 号）

➢ 产业优化升级和新型经济模式

在这两股力量的推动下，将会在传统产业、传统 IT 和互联网产业的基础上形成新的产业层级：大数据分析和智慧产业，大数据分析和智慧产业的发展将推动传统产业、传统信息产业的升级和改造，并产生新型的经济模式。

如：目前基于手机和可佩戴计算的反馈经济模式逐渐萌芽，在医疗领域三个斯坦福大学生创立的脊椎病防治中心，通过皮带传感器监测坐姿，传输到云中心，通过比较分析，给出调整信号到手机。智慧社区建设中，通过监测老人血压和脉搏心跳，反馈吃药和预警信息。

基于大数据的新型经济模式，将逐步改变社会经济形态，带来一场新的生产革命，主要表现为以下几点：

（1） 一切以数据为中心，基于数据的挖掘分析；

（2） 以最小的成本，创造更快更好的产品；

（3） 试验性思维，微创新，降低风险；

（4） 反复迭代与用户紧密联系，精益求精。

➢ 提升社会公共管理水平

从公共管理和服务的角度看，随着我国城镇化的发展和大城市的不断涌现，资源向少数大城市集中，“大城市病”不可避免，“大城市病”的主要特征就是公共资源十分紧张，人口与资源的矛盾极其尖锐。使得相关的配套设施难以承受人口压力，教育、卫生、交通、水电气等公共事业，会因为人口的迅猛增长而日益相对紧缺。大数据产业的创新发展必将和解决“大城市病”紧密联系，如果将智慧城市比喻为人，将组成智慧城市感知功能的传感器比作人的五官，将连接传感器的网络比作神经，将控制和存储信息的云技术比作中枢，那么依据大数据分析和决策就是智慧城市的大脑。

4.4.2 大数据产业发展的主要瓶颈

➢ 大数据生态河流学说

国外一个公认的大数据生态河流学说：人类历史上最早的一批城市都诞生在

河边，河流为人类提供了食物、水和交通。当今社会，大数据犹如环绕在城市中的一条数据河流，滋养着城市信息经济的发展，数据将成为社会发展重要的生产资料、创新的动力。城市作为大数据的主要载体，保证数据河流源源不断的滋养着城市，就必须做到数据河流的流通、开放，同时必要的堤坝围挡也是保障数据河流有序、流畅的关键。

➢ 我国大数据产业需要走产业环境推动信息环境发展的路径

美国走在了大数据开放的前列，我国由于行政体制的特点，政府长期以来只进行数据的采集，缺乏对数据的开放、分析和整合。政府掌握的公共数据是一个社会的基础数据，没有它的开放，整个社会的数据整合就难以谈起，大数据产业的发展就会大打折扣。

数据的开放将创造出更大的社会价值。如果一个城市公开所有的交通事故的数据，包括时间和地点，人们将会根据这些数据提高警觉，改善城市的交通安全。人们对数据的需求增长，催生了产业的发展。例如现在有很多商品比价手机应用系统，可以帮助使用者买到性价比最高的商品。城市路况手机应用，提供路况拥堵状况为公众出行提供实时建议，而且为地铁系统在客流高低峰时段、热点站和普通站之间的调配提出更优的方案。由需求推动的大数据产业环境已经形成。最终产业环境的发展必然带动信息环境的发展。走产业环境推动信息环境发展的路径将是我国大数据产业发展的最佳方向。

➢ 数据开放与隐私保护

数据开放之后，我们将面临隐私保护的巨大挑战。政府、银行、医院等各种组织收集了大量的个人信息，这些数据和资料可能与个人的生活方式、消费方式密切相关，在对其进行分析的时候，可能根据各种信息整合就能准确知道是谁在做什么，这就涉及到个人隐私的问题。近年来侵犯个人隐私案件时有发生，如谷歌泄露个人隐私事件、盛大云数据丢失事件、2011 年韩国三大门户网站之一 Nate 和社交网络“赛我网”遭到黑客攻击，致使 3500 万用户信息泄露等事件，这些严重侵犯了用户的合法权益。隐私保护已经成为大数据产业发展的主要瓶颈之一。

大数据的发展给个人带来生活、工作上的便利，给企业带来了更多机遇，为社会创造出更多的价值。大数据产业的创新发展需要政府开放公共数据，以此推动整个社会数据的开放，鼓励个体创新，同时做好隐私保护工作，保障大数据产

业的健康发展。

第五章 大数据人才资源

大数据来势迅猛，仿佛一夜之间大家都在谈大数据、做大数据、用大数据。于是，大数据人才变成了紧缺人才。根据麦肯锡报告，“到 2018 年，美国在‘深度分析’人才方面将面临 14 万至 19 万的人才缺口；在‘能够分析数据帮助公司做出商业决策’方面将面临 150 万的人才缺口”。另一方面，《哈佛商业评论》声称，21 世纪最富挑战的工作是数据科学家。很多企业开始设立数据科学家岗位。美国社交媒体公司 DataSift 的创始人兼 CTONick Halstead 认为：“大数据的真正价值就在于‘数据科学家’这一提法的传播”。据 Gartner 预测，到 2015 年，全球将新增 440 万个与大数据相关的工作岗位，且会有 25%的组织设立首席数据官职位。其中有 190 万个工作岗位将在美国。而每一个与大数据有关的 IT 工作，都将在技术行业外部再创建 3 个工作岗位，这将在美国再创建将近 600 万个工作岗位。但是，Gartner 也同时指出，拥有大数据技能的 IT 专业人员严重短缺，只有 1/3 的新的工作岗位将雇佣到人员。

从广义上讲，大数据人才就是数据科学家和数据工程师，因此，大数据人才的培养就是数据科学家和数据工程师的培养。这从国际上开设的《数据科学》课程、数据科学学位计划、数据科学短期培训班可以看出这一点。

在中国，香港中文大学从 2008 年起设立“数据科学商业统计”科学硕士学位；复旦大学从 2007 年起开设数据科学讨论班，2010 年开始招收数据科学博士研究生，并从 2013 年起开设研究生课程《数据科学》；北京航空航天大学于 2012 年设立大数据工程硕士学位。在美国，加州大学伯克利分校（UC Berkeley）从 2011 年起开设《数据科学导论》课程，并从 2012 年起开设《数据科学和分析》课程；伊利诺伊大学香槟分校（University of Illinois at Urbana-Champaign，UIUC）从 2011 年起举办“数据科学暑期研究班”（Data Sciences Summer Institute program）；哥伦比亚大学（Columbia University）从 2011 年起开设《数据科学导论》课程，2013 年起开设《应用数据科学》课程，并将从 2013 年秋季起开设“数据科学专业成就认证”（Certification of Professional Achievement in Data Sciences）培训项目，并计划从 2014 年起设立硕士学位，2015 年起设立博士学位；芝加哥大学（University of Chicago）开设 3 个月的夏季培训课程；纽约大学（New York

University）将从 2013 年秋季起设立“数据科学”硕士学位；南加州大学（South California University）设立“数据科学”硕士学位；华盛顿大学（University of Washington）从 2013 年 5 月起开设《数据科学导论》课程，并对修满数据科学相关课程学分的学生颁发数据科学证书（Certificate in Data Science）；雪城大学（Syracuse University）也提供数据科学高级研究证书（Certificate Advanced Studies in Data Science）培训项目。在英国，邓迪大学从 2013 年起设立“数据科学”科学硕士学位。

从上述人才的培养计划来看，数据科学家应该系统地掌握数据分析相关的技能，主要包括数学、统计学、数据分析、商业分析和自然语言处理等，具有较宽的知识面，具有独立获取知识的能力，具有较强的实践能力和创新意识。其中，只有复旦大学的课程设置强调了数据科学家是研究数据的科学家，而不仅仅是一个数据工程师或者数据分析师。

目前，大数据人才培养可以分成两个方面：一个是学位培养，另一个职业培训。

5.1 数据科学学位人才培养

该类型人才的培养主要包括本科生、硕士生、博士生，向其颁发数据科学学士、硕士和博士学位,为政府和公司输送数据科学家。而培养数据科学家，除了需要很好掌握数学、计算机科学和应用统计学等基础知识点外，还需深入学习经济、生物、物理、化学等交叉学科业务课程，并在数据获取、数据存储、数据检索等数据工程方面做深入的了解和亲身实践。IBM 公司的“全球大学关系项目”总监、同时也是计算机科学家的吉姆•斯伯热表示，从学术角度看，在一些本来跟数据无缘的学科里，比如社会科学和人文学科的一些分支，大数据也正在发挥重要作用。数据科学学位的人才培养需要关注如下几点：

- **数据科学家储备欠缺**

大数据职位相关的技能主要包括数学、统计学、数据分析、商业分析和自然语言处理，数据科学家是复合型人才，是对数学、统计学、机器学习等多方面知识的综合掌控。

大数据最关键的部分是数据分析和挖掘数据价值，要获得这些，就需要大量

的数据科学家。数据科学家是复合型人才，是对数学、统计学、机器学习等多方面知识的综合掌控。初级的分析人员只能是对数据进行报表、描述性分析，真正高级的数据科学家需要对数据做出预测性的、有价值的分析。从目前的人才储备来看，这部分的储备欠缺。

- **掌握机器学习和知识图谱很重要**

从计算机学界的理解来看，大数据的核心技术是机器学习和知识图谱。这是一种框架性的知识，介于基础设施和应用之间的技术。例如大数据应用的代表谷歌公司就有两个大的开发方向，一个是机器学习，另一个是由搜索团队负责的知识图谱。

任何一种大数据方案都不可能适合所有的行业，因此，大数据的核心业务必然是一种扎根于特定行业，综合运用已有的存储、分析、挖掘、展现技术，根据用户需求并融入行业特色技术模型的一站式大数据平台业务。正是由于大数据具有这样的业务特点，所以企业最需要两种人才：一类是综合型人才，另一类是技术专家。一方面，大数据具有强烈的行业特点，这就需要综合型人才，这种人才需要了解行业，了解技术的各个层面，以综合的视角制定确实可行的方案为目的，还必须具有统计学背景，并对数据管理有丰富经验，他们是目前最急缺的人才；另一方面，大数据方案的实现，必须由技术专家来完成，技术专家的能力也直接决定了企业所能制定大数据方案的深度和广度，传统的数据库应用开发，特别是商业智能应用开发人才，以及熟悉 Hadoop 等分布式存储的人才，也都是必不可少的。

- **大数据人才培养需要校企合作**

企业可以与学校联合培养人才，或建立专门的数据科学家团队，或与专业的数据处理公司合作，以解人才之急。对于企业来说，虽然人才储备有缺口，但是大数据业务还是得做。虽然目前大数据应用比较少，人才也比较少，但是中国的知识积累并不少，例如中国的学术界和产业界在机器学习上也有积累，现在的问题是如何将这方面知识和大数据结合起来。

由企业和大学合作来培养自己所需要的大数据人才，是考虑到大数据的解剖对象是大量的数据，这些数据只有企业才有，而学校并不生产数据。在企业的支持下，学校就能通过针对性的实践训练来培养学生的技能。

5.2 数据科学职业人才培训

该类型人才的培训主要针对大数据在商业和数据分析中的应用、商业智能的管理者、数据库专家和在校欲将数据科学作为未来职业的研究生，提供中短期培训项目，培训合格后向其颁发数据科学培训证书，从而培养符合国民经济发展战略急需的数据工程师和数据分析师。该类型人才的培养除了要初步掌握数学、计算机科学和应用统计学等基础知识点，需着重学习数据获取、数据存储和数据检索等数据工程方面的知识，并根据其所在领域参与商业大数据项目的分析和处理。

大数据分析的广泛应用要求对现有岗位进行再培训，同时也会出现新职位和新技能，比如“首席数据官”就会成为大大小小的企业里司空见惯的职位。这些岗位也并不一定就非要资深数据分析专家才能胜任，那些在各自行业中受过良好教育、经验丰富的非数据专家只要能善于使用大数据工具就能担任。数据科学职业人才就业的行业和岗位需求以及发展趋势如下：

- **大数据职业人才就业的主要行业包括：**

（1）零售、保险、电子商务

（2）政府数据中心

（3）医药和银行

（4）研究性大学

（5）金融机构

（6）互联网企业

- **典型的大数据专业岗位需求：**

（1）大数据系统研发工程师：负责大数据系统研发工作，包括大规模非结构化数据业务模型构建、大数据存储、数据库架构设计以及数据库详细设计、优化数据库构架、解决数据库中心建设设计问题。他们还负责集群的日常运作、系统的监测和配置、Hadoop 与其他系统的集成。

（2）大数据应用开发工程师：负责搭建大数据应用平台、开发分析应用程序。他们熟悉工具或算法、编程、包装、优化或者部署不同的 MapReduce 事务。他们以大数据技术为核心，研发各种基于大数据技术的应用程序及行业解决方案。

（3）大数据分析师：运用算法来解决分析问题，并且从事数据挖掘工作。他们最大的本事就是能够让数据道出真相；此外，他们还拥有某个领域的专长，

帮助开发数据产品，推动数据解决方案的不断更新。

（4）数据可视化工程师：具备良好的沟通能力与团队精神，责任心强，拥有优秀的解决问题的能力。他们负责在收集到的高质量数据中，利用图形化的工具及手段的应用，一目了然地揭示数据中的复杂信息，帮助企业更好的进行大数据应用开发，发现大数据背后的巨大财富。

- **大数据领域从业人员的十个趋势。**

（1）薪金将继续增长

（2）大数据人才供不应求

（3）雇佣外包

（4）人才团队内出现分歧

（5）大数据专业人士需要不断进步

（6）精通大数据的专业人才将成为最重要的业务角色

（7）大数据领域需要数据科学家

（8）高校应对大数据人才缺口

（9）数据驱动的工作令人满意并充满挑战

（10）大数据专业人士将拥抱未来

第六章 大数据发展趋势与建议

6.1 大数据科学问题与学科发展趋势

6.1.1 大数据的科学问题

尽管大数据的涌现为人们提供了前所未有的宝贵机遇，但同时也提出了重大的挑战。首先，大数据规模巨大、分布广泛、动态演变、模态多样、关联复杂、真伪难辨等这一系列特性带来了数据复杂性的挑战。特别地，大数据模式多样，内容难于理解；大数据关联关系复杂，数据难以有效识别；数据的质量良莠不齐，真伪难以判定。因此，需要揭示、度量并刻画数据复杂性，并厘清其中的内在关联机理。其次，大数据的数据复杂性又不可避免地带来了关于大数据是否可以计算以及计算复杂性的挑战。即便大数据可以计算，当前处理有限规模数据的计算体系已然失效，因此需要寻找大数据的稳定内核及计算边界，在此基础上提出新型的不以数据规模为变量的计算范式。最后，大数据种种特性的综合呈现还造成大数据处理在系统层面的系统复杂性的挑战。因此，又需要提出面向不同大数据模式（如离线历史数据与在线流式数据等）的新型处理系统架构以及相应的评价体系与优化策略。

- **大数据复杂性的内在机理**

由于大数据的出现，人们处理计算问题时获得了前所未有的大规模样本，但同时也不得不面对更加复杂的数据对象，其典型的特性是类型和模式多样、关联关系繁杂、质量良莠不齐。大数据内在的复杂性使得数据的感知、表达、理解和计算等多个环节面临着巨大的挑战，导致了传统全量数据计算模式下时空维度上计算复杂度的激增，很多传统的数据分析与挖掘任务如检索、主题发现、语义和情感分析等变得异常困难。然而目前，人们对大数据复杂性的内在机理及其背后的物理意义缺乏理解，对大数据的分布与协作关联等规律认识不足，对大数据的复杂性和计算复杂性的内在联系缺乏深刻理解，加上缺少面向领域的大数据处理知识，极大地制约了人们对大数据高效计算模型和方法的设计能力。有鉴于此，如何量化定义大数据复杂性的本质特征及其外在度量指标，进而研究数据复杂性

的内在机理是个基础问题。需要建立多模态关联关系下的数据分布理论和模型，厘清数据复杂度和时空计算复杂度之间的内在联系，通过对数据复杂性内在机理的建模和解析，阐明大数据按需约简、降低复杂度的原理与机制，使之成为大数据计算的理论基石。

- **大数据的可计算性及新型计算范式**

大数据规模巨大等特性使得传统的计算方法已经不能有效地支持大数据计算和处理。在求解大数据的问题时，需要重新审视和研究它的可计算性、计算复杂性和求解算法。特别地，大数据计算不能像小样本数据集那样依赖于对全局数据的统计分析和迭代计算，需要突破传统计算对数据的“独立同分布”和“采样充分性”的假设。因此，研究面向大数据计算的高效新型范式，改变人们对数据计算的本质看法，提供处理和分析大数据的基本方法，支持价值驱动的特定领域应用，是大数据研究的一个核心问题。而大数据样本量充分，内在关联关系密切而复杂，价值密度分布极不均衡，这些特征对研究大数据的可计算性及建立新型计算范式提供了机遇，同时也提出了挑战。研究大数据的获取和表达，研究数据空间中各种关联关系及其语义特征与表示，基于数据内在结构、关联关系、产生规律及演化特点，发现针对大数据的稳定计算模型，特别是针对计算问题的大数据稳定内核和计算边界，建立局部近似整体的非确定性增量学习理论和方法，进而提出不以样本规模为变量的新型计算范式。

- **大数据处理系统的效能评价与优化**

大数据处理系统是支持大数据科学研究的基础平台。对于规模巨大、价值稀疏、结构复杂、变化迅速的大数据，其处理亦面临计算复杂度高、任务周期长、实时性要求强等难题。大数据及其处理的这些难点不仅对大数据处理系统的系统架构、计算框架、处理方法提出了新的挑战，更对大数据处理系统的运行效率及单位能耗提出了苛刻要求，要求大数据处理系统必须具有高效能的特点。对于以高效能为目标的大数据处理系统的系统架构设计、计算框架设计、处理方法设计和测试基准设计研究，其基础是大数据处理系统的效能评价与优化问题研究。大数据处理系统的效能评价与优化问题具有极大的研究挑战性，其解决不但要求厘清大数据的复杂性、可计算性与系统处理效率、能耗间的关系，还要综合度量系统中如系统吞吐率、并行处理能力、作业计算精度、作业单位能耗等多种效能因

素，更涉及实际负载情况及资源分散重复情况的考虑。大数据处理系统的效能评价与优化问题的解决可奠定大数据处理系统设计、实现、测试与优化的基本准则，是构建能效优化的分布式存储和处理的硬件及软件系统架构的重要依据和基础，因此是大数据科学研究必须解决的关键问题。

6.1.2 大数据的学科发展趋势

美国政府 6 个部门启动的大数据研究计划中，除了自然科学基金会的研究内容提到要“形成一个包括数学、统计基础和计算机算法的独特学科”外，绝大多数研究项目都是应对大数据带来的技术挑战，重视的是数据工程而不是数据科学，主要考虑大数据分析算法和系统的效率。例如，国防部高级研究计划局(DARPA)的大数据研究项目包括：多尺度异常检测项目旨在解决大规模数据集的异常检测和特征化；网络内部威胁计划旨在通过分析传感器和其他来源的信息，进行网络威胁和非常规战争行为的自动识别；Machine Reading 项目旨在实现人工智能的应用和发展学习系统，对自然文本进行知识插入。其实，技术上解决不了的问题越来越多，就会逐步凝练出共性的科学挑战问题。因此，在条件还不成熟的时候，计算机科学家应虚心地甘当一段时期的“助手”，虚心与各应用领域的科研人员合作，首先努力解决各领域大数据处理提出的技术挑战问题。在网络大数据方面可能计算机学者的主动性会较早发挥出来。从学科的角度来说，大数据的发展将主要关注以下的问题。

- **“数据科学”研究的对象是什么？**

计算机科学是关于算法的科学，数据科学是关于数据的科学。从事数据科学研究的学者更关注数据的科学价值，试图把数据当成一个“自然体(data nature)”来研究，提出所谓“数据界（data universe）”的概念，颇有把计算机科学划归为自然科学的倾向。但脱离各个领域的“物理世界”，作为客观事物间接存在形式的“数据界”究竟有什么共性问题还不清楚。物理世界在网络空间中有其数据映像，目前一些学者在研究的数据界的规律其本质可能是物理世界的规律（还需要在物理世界中测试验证）。除去各个领域的规律，作为映像的“数据界”的共同规律是什么？这是一个数据科学需要探索的根本问题。

任何领域的研究，若要成为一门科学，一定是研究共性的问题。针对非常狭

窄的领域的某个具体问题，主要依靠该问题涉及的特殊条件和专门知识做数据挖掘，不大可能使大数据成为一门科学。数据研究能成为一门科学的前提是，在一个领域发现的数据相互关系和规律具有可推广到其他领域的普适性。抽象出一个领域的共性科学问题往往需要较长的时间，提炼“数据界”的共性科学问题还需要一段时间的实践积累。至少未来 5-10 年内计算机界的学者还需要多花精力协助其他领域的学者解决大数据带来的技术挑战问题。通过分层次的不断抽象，大数据的共性科学问题才会逐步清晰明朗。

当前数据科学的目标还不很明确，但与其他学科一样，科学研究的道路常常是先做“白盒研究”，知识积累多了就有可能抽象出通用性较强的“黑盒模型”和普适规律。数据库理论是一个很好的例子。在经历了层次数据库、网状数据库多年实践以后，Codd 发现了数据库应用的共性规律，建立了有坚实理论基础的关系模型。在这之前人们也一直在问数据库可不可能有共性的理论。现在大数据研究要做的事就是提出像关系数据库这样的理论来指导海量非结构化数据的处理。

信息技术的发展使我们逐步进入“人-机-物”融合的三元世界，未来的世界可以做到“机中有人，人中有机，物中有机，机中有物”。所谓“机”就是联系人类社会（包括个人身体与大脑）与物理世界的网络空间，其最基本的构成元素是不同于原子和神经元的 bit。物理空间和人类社会（包括人的大脑）都有共性的科学问题和规律，与这两者有密切联系的网络空间会不会有不同的共性科学问题？从“人-机-物”三元世界的角度来探讨大数据科学的共性问题，也许是一个可以尝试的突破口。

➢ 数据背后的共性问题—关系网络

观察各种复杂系统得到的大数据，直接反映的往往是一个个孤立的数据和分散的链接，但这些反映相互关系的链接整合起来就是一个网络。例如，基因数据构成基因网络，脑科学实验数据形成神经网络，Web 数据反映出社会网络。数据的共性、网络的整体特征隐藏在数据网络中，大数据往往以复杂关联的数据网络这样一种独特的形式存在，因此要理解大数据就要对大数据后面的网络进行深入分析。网络有不少参数和性质，如平均路径长度、度分布、聚集系数、核数、介数等，这些性质和参数也许能刻画大数据背后的网络的共性。因此，大数据面临

的科学问题本质上可能就是网络科学问题，复杂网络分析应该是数据科学的重要基石。

网络数据研究应发现网络数据产生、传播以及网络信息涌现的内在机制，还要研究隐藏在数据背后的社会学、心理学、经济学的机理，同时利用这些机理研究互联网对政治、经济、文化、教育、科研的影响。基于大数据对复杂系统内在机理进行整体性的研究，也许将为研究复杂系统提供新的途径。从这种意义上看，数据科学是从整体上研究复杂系统的一门科学。

发现 Scale-free 网络的 Albert-László Barabási 教授在 2012 年 1 月的 Nature Physics 上发表一篇重要文章“The network takeover”。文章认为：20 世纪是量子力学的世纪，从电子学到天文物理学，从核能到量子计算，都离不开量子力学；而到了 21 世纪，网络理论正在成为量子力学的可尊敬的后继，正在构建一个新的理论和算法的框架。

➢ 大数据研究中的关联关系与因果关系

大数据研究不同于传统的逻辑推理研究，而是对数量巨大的数据做统计性的搜索、比较、聚类、分类等分析归纳，因此继承了统计科学的一些特点。统计学关注数据的相关性或称关联性，所谓“相关性”是指两个或两个以上变量的取值之间存在某种规律性。“相关分析”的目的是找出数据集里隐藏的相互关系网（关联网），一般用支持度、可信度、兴趣度等参数反映相关性。严格来讲，统计学无法检验逻辑上的因果关系。

也许正是因为统计方法不致力于寻找真正的原因，才促使数据挖掘和大数据技术在商业领域广泛流行。企业的目标是多赚钱，只要从数据挖掘中发现某种措施与增加企业利润有较强的相关性，采取这种措施就是了，不必深究为什么能增加利润，更不必发现其背后的内在规律和模型。一般而言，企业收集和处理大数据，不是按学者们经常描述的“从数据到信息再到知识和智慧”的研究思路，而是走“从数据直接到价值”的捷径。Google 广告获得巨额收入经常被引用作为大数据相关分析的成功案例，美国 Wired 杂志主编 Chris Anderson 在他的著名文章“The End of Theory”的结尾发问：“现在是时候问这一句了：科学能从谷歌那儿学到什么？”。

因果关系的研究曾经引发了科学体系的建立，近代科学体系获得的成就已经

证明，科学是研究因果关系最重要的手段。相关性研究是可以替代因果分析的科学新发展还只是因果分析的补充，不同的学者有完全不同的看法。对于简单的封闭的系统，基于小数据的因果分析容易做到。当年开普勒发现行星三大定律，牛顿发现力学三大定律都是基于小数据。但对于开放复杂的巨系统，传统的因果分析难以奏效，因为系统中各个组成部分之间相互有影响，可能互为因果，因果关系隐藏在整个系统之中。现在的“因”可能是过去的“果”，此处的“果”也可能是别处的“因”，因果关系本质上是一种相互纠缠的相关性。在物理学的基本粒子理论中，颇受重视的欧几里德量子引力学（霍金所倡导的理论）本身并不包括因果律。因此，对于大数据的关联分析是不是“知其然而不知其所以然”，其中可能包含深奥的哲理，不能贸然下结论。

➢ 社会科学的大数据研究

根据数据的来源，大数据可以初略地分称两大类：一类来自物理世界，另一类来自人类社会。前者多半是科学实验数据或传感数据，后者与人的活动有关系，特别是与互联网有关。这两类数据的处理方式和目标差别较大，不能照搬处理科学实验数据的方法来处理 Web 数据。

科学实验是科技人员设计的，如何采集数据、处理数据事先都已经想好了，不管是检索还是模式识别，都有一定的科学规律可循。美国的大数据研究计划中专门列出寻找希格斯粒子（被称为“上帝粒子”）的大型强子对撞机（LHC）实验。这是一个典型的基于大数据的科学实验，至少要在一万亿个事例中才可能找出一个希格斯粒子。2012 年 7 月 4 日，CERN 宣布发现新的玻色子，标准差为 4.9，被认为可能是希格斯玻色子（承认是希格斯玻色子粒子需要 5 个标准差，即 99.99943%的可能性是对的）。设计这一实验的激动人心之处在于，不论找到还是没有找到希格斯粒子，都是物理学的重大突破。从这一实验可以看出，科学实验的大数据处理是整个实验的一个预定步骤，发现有价值的信息往往在预料之中。

Web 上的信息（譬如微博）是千千万万的人随机产生的，从事社会科学研究的学者要从这些看似杂乱无章的数据中寻找有价值的蛛丝马迹。网络大数据有许多不同于自然科学数据的特点，包括多源异构、交互性、时效性、社会性、突发性和高噪声等，不但非结构化数据多，而且数据的实时性强，大量数据都是随

机动态产生。科学数据的采集一般代价较高，LHC 实验设备花了几十亿美元。因此对采集什么数据做过精心安排。而网络数据的采集相对成本较低，网上许多数据是重复的或者没有价值，价值密度很低。一般而言，社会科学的大数据分析，特别是根据 Web 数据做经济形势、安全形势、社会群体事件的预测，比科学实验的数据分析更困难。

未来的任务主要不是获取越来来越多的数据，而是数据的去冗分类、去粗取精，从数据中挖掘知识。几百年来，科学研究一直在做“从薄到厚”的事情，把“小数据”变成“大数据”，现在要做的事情是“从厚到薄”，要把大数据变成小数据。要在不明显增加采集成本的条件下尽可能提高数据的质量。要研究如何科学合理地抽样采集数据，减少不必要的数据采集。两三岁的小孩学习识别动物和汽车等，往往几十张样本图片就足够了，研究清楚人类为什么具有小数据学习能力，对开展大数据分析研究具有深远的指导意义。

近十年来增长最快的数据是网络上传播的各种非结构化或半结构化的数据。网络数据的背后是相互联系的各种人群，网络大数据的处理能力直接关系到国家的信息空间安全和社会稳定。从心理学、经济学、信息科学等不同学科领域共同探讨网络数据的产生、扩散、涌现的基本规律，是建立安全和谐的网络环境的重大战略需求，是促使国家长治久安的大事。我国拥有世界上最多的网民和最大的访问量，在网络大数据分析方面已经有较强的基础，有可能做出世界领先的原始创新成果，应加大网络大数据分析方面的研究力度。

- **数据处理的复杂性研究**

计算复杂性是计算机科学的基本问题，科学计算主要考虑时间复杂性和空间复杂性。对于大数据处理，除了时间和空间复杂性外，可能需要考虑解决一个问题需要多大的数据量，暂且称为“数据量复杂性”。数据量复杂性和空间复杂性不是一个概念，空间复杂性要考虑计算过程中产生的空间需求。

基于唯象假设的数据处理常出现增量式进步，数据多一点，结果就好一点点。这类问题的数据科学价值可能不大。反过来，可能有些问题的数据处理像个无底洞，无论多少数据都不可能解决问题。这种问题有些类似 NP 问题。我们需要建立一种理论，对求解一个问题达到某种满意程度（对判定问题是有多大把握说“是”或“否”，优化问题是接近最优解的程度）需要多大规模的数据量给出理论上的

判断。当然，目前还有很多问题还没有定义清楚，比如，对于网络搜索之类的问题，如何定义问题规模和数据规模等。

对从事大数据研究的学者而言，最有意思的问题应该是，解决一个问题的数据规模有一个阈值。数据少于这个阈值，问题解决不了；达到这个阈值，就可以解决以前解决不了的大问题；而数据规模超过这个阈值，对解决问题也没有更多的帮助。这类问题可称为“预言性数据分析问题”，即在做大数据处理之前，我们可以预言，当数据量到达多大规模时，该问题的解可以达到何种满意程度。

与社会科学有关的大数据问题，例如舆情分析、情感分析等，遇到许多理论问题过去没有考虑过，才刚刚开始研究。迫切需要计算机学者与社会科学领域的学者密切合作，共同开拓新的疆域。借助大数据的推力，社会科学将脱下“准科学”的外衣，真正迈进科学的殿堂。

➢ 科研第四范式是思维方式的大变化

已故图灵奖得主吉姆•格雷（Jim Gray）提出的数据密集型科研“第四范式”（the fourth paradigm），将大数据科研从第三范式（计算科学）中分离出来单独作为一种科研范式，是因为其研究方式不同于基于数学模型的传统研究方式。Google 公司的研究部主任 Peter Norvig 的一句名言可以概括两者的区别：“所有的模型都是错误的，进一步说，没有模型你也可以成功”（All models are wrong, and increasingly you can succeed without them）。PB 级数据使我们可以做到没有模型和假设就可以分析数据。将数据丢进巨大的计算机机群中，只要有相互关系的数据，统计分析算法可以发现过去的科学方法发现不了的新模式、新知识甚至新规律。实际上，Google 的广告优化配置、战胜人类的 IBM 沃森问答系统都是这么实现的，这就是“第四范式”的魅力！

美国 Wired 杂志主编 Chris Anderson 2008 年曾发出“理论已终结”的惊人断言：“数据洪流使（传统）科学方法变得过时”（The Data Deluge Makes the Scientific Method Obsolete）。他指出获得海量数据和处理这些数据的统计工具的可能性提供了理解世界的一条完整的新途径。Petabytes 让我们说：相互关系已经足够（Correlation is enough）。我们可以停止寻找模型，相互关系取代了因果关系，没有具有一致性的模型、统一的理论和任何机械式的说明，科学也可以进步。

Chris Anderson 的极端看法并没有得到科学界的普遍认同，数据量的增加能否引起科研方法本质性的改变仍然是一个值得探讨的问题。对研究领域的深刻理解（如空气动力学方程用于风洞实验）和数据量的积累应该是一个迭代累进的过程。没有科学假设和模型就能发现新知识究竟有多大的普适性也需要实践来检验，我们需要思考：这类问题有多大的普遍性？这种优势是数据量特别大带来的还是问题本身有这种特性？所谓从数据中获取知识要不要人的参与，人在机器自动学习和运行中应该扮演什么角色？也许有些领域可以先用第四范式，等领域知识逐步丰富了再过渡到第三范式。

6.2 大数据的技术挑战与发展趋势

6.2.1 大数据的技术挑战

现有的数据中心技术很难满足大数据的需求，需要考虑对整个 IT 架构进行革命性的重构。而存储能力的增长远远赶不上数据的增长，因此设计最合理的分层存储架构已成为 IT 系统的关键。数据的移动已成为 IT 系统最大的开销，目前传送大数据最高效也最实用的方式是通过飞机或地面交通工具运送磁盘而不是网络通信。在大数据时代，IT 系统需要从数据围着处理器转改变为处理能力围着数据转，将计算推送给数据，而不是将数据推送给计算。大数据也导致高可扩展性成为对 IT 系统最本质的需求，并发执行（同时执行的线程）的规模要从现在的千万量级提高到 10 亿级以上。

在应对处理大数据的各种技术挑战中，以下几个问题值得高度重视。

- **大数据的去冗降噪技术**

大数据一般都来自多个不同的源头，而且往往以动态数据流的形式产生。因此，大数据中常常包含有不同形态的噪声数据。另外，数据采样算法缺陷与设备故障也可能会给导致大数据的噪声。大数据的冗余则通常来自两个方面：一方面，大数据的多源性导致了不同源头的数据中存在有相同的数据，从而造成数据的绝对冗余；另一方面，就具体的应用需求而言，大数据可能会提供超量特别是超精度的数据，这又形成数据的相对冗余。降低噪声、消除冗余是提高数据质量、降低数据存储成本的基础。

➢ 大数据的新型表示方法

目前表示数据的方法，不一定能直观地展现出大数据本身的意义。要想有效利用数据并挖掘其中的信息或知识，必须找到最合适的数据表示方法。在一种不合适的数据表示中寻找大数据的固定模式、因果关系和关联关系时，可能会落入固有的偏见之中。数据表示方法和最初的数据产生者有着密切关系。如果原始数据有必要的标识，就会大大减轻事后数据识别和分类的困难。但标识数据会给用户增添麻烦，所以往往得不到用户认可。研究既有效又简易的数据表示方法是处理网络大数据必须解决的技术难题之一。

➢ 高效率低成本的大数据存储

大数据的存储方式不仅影响其后的数据分析处理效率也影响数据存储的成本。因此，就需要研究高效率低成本的数据存储方式。具体则需要研究多源多模态数据高质量获取与整合的理论和技术、流式数据的高速索引创建与存储、错误自动检测与修复的理论和技术、低质量数据上的近似计算的理论和算法等。

➢ 大数据的有效融合

数据不整合就发挥不出大数据的大价值。大数据的泛滥与数据格式太多有关。大数据面临的一个重要问题是个人、企业和政府机构的各种数据和信息能否方便的融合。如同人类有许多种自然语言一样，作为网络空间中唯一客观存在的数据难免有多种格式。但为了扫清网络大数据处理的障碍，应研究推广不与平台绑定的数据格式。大数据已成为联系人类社会、物理世界和网络空间的纽带，需要通过统一的数据格式构建融合人、机、物三元世界的统一信息系统。

➢ 非结构化和半结构化数据的高效处理

据统计，目前采集到的数据 85%以上是非结构化和半结构化数据，而传统的关系数据库技术无法胜任这些数据的处理，因为关系数据库系统的出发点是追求高度的数据一致性和容错性。根据 CAP (Consistency, Availability, tolerance to network Partitions) 理论，在分布式系统中，一致性、可用性、分区容错性三者不可兼得，因而并行关系数据库必然无法获得较强的扩展性和良好的系统可用性。系统的高扩展性是大数据分析最重要的需求，必须寻找高扩展性的数据分析技术。以 MapReduce 和 Hadoop 为代表的非关系数据分析技术，以其适合非结构数据

处理、大规模并行处理、简单易用等突出优势，在互联网信息搜索和其他大数据分析领域取得了重大进展，已成为大数据分析的主流技术。MapReduce 和 Hadoop 在应用性能等方面还存在不少问题，还需要研究开发更有效、更实用的大数据分析和管理技术。

- **适合不同行业的大数据挖掘分析工具和开发环境**

不同行业需要不同的大数据分析工具和开发环境，应鼓励计算机算法研究人员与各领域的科研人员密切合作，在分析工具和开发环境上创新。当前跨领域跨行业的数据共享仍存在大量壁垒，海量数据的收集，特别是关联领域的同时收集还存在很大挑战。只有跨领域的数据分析才更有可能形成真正的知识和智能，产生更大的价值。

另外，整个大数据产业也越来越需要一个公开的、标准化的、高度整合的基础平台，方便各行各业上层业务的研发和创新，并能够平衡整体方案的投资成本（CapEx）和运营成本（OpEx），降低大数据业务的经济及技术准入门槛，从而真正推动整个大数据产业的发展。

- **大幅度降低数据处理、存储和通信能耗的新技术**

大数据的获取、通信、存储、管理与分析处理都需要消耗大量的能源。在能源问题日益突出的今天，研究创新的数据处理和传送的节能方法与技术是个重要的研究方向。

6.2.2 大数据的技术发展趋势

CCF 大数据专家委员会希望通过对于大数据发展趋势的年度预测，将最受关注的科学、技术、产业、应用、政策等相关变化趋势发掘出来，以便大数据领域相关的各界人士能够从中获得启迪，或顺势而为，或依势而起，或引领潮头。

与大数据的热点问题有所不同，大数据的发展趋势是什么，这是个相对开放的问题。因此，我们从大数据专家委员会委员的反馈中总结出了多达 37 个发展趋势的候选项。经过整理，把他们归结成 7 个不同的方面，具体包括大数据的整体态势与发展、大数据与学术、大数据与人、大数据的安全与隐私、大数据应用、大数据系统与处理以及大数据对产业的影响。在投票过程中，我们采用了 2+10

的模式，其中 2 是请委员们从提供的选项中分别选择出大数据最令人瞩目的子学科与应用，然后再选出 10 个最可能的发展趋势。

经过全体委员的积极投票，我们得到如下的结果。对于最令人瞩目的子学科，我们提供了 8 个选项：机器学习、计算机视觉、复杂网络、社会计算、数据挖掘、分布式计算、高性能计算与统计学。在收集到的选票中，数据分析与预测、分布式计算以及社会计算当选为最令人瞩目的子学科。而对于最令人瞩目的应用，我们提供的选项包括了城市管理、国家部委大数据、犯罪侦查、电信、金融（股市预测、金融分析等）、电子商务、电力、能源石化与医疗（流行病监控和预测等）。在所收集到的选票中，医疗、金融、城市管理与电子商务是委员们最为关心的大数据应用。最后，我们从其他的候选项中，根据投票数，选出了 10 个主要的发展趋势。

- **数据资源化**

这一候选发展趋势得到了委员们最多的关注。数据的资源化是指大数据在企业、社会和国家层面成为重要的战略资源。往后大数据将成为新的战略制高点，是大家抢夺的新焦点；大数据将不断成为机构的资产，成为提升机构和公司竞争力的有力武器。

- **大数据隐私问题**

大数据对于隐私将是一个重大挑战，现有的隐私保护法规和技术手段难于适应大数据环境，个人隐私越来越难以保护，有可能会出现有偿隐私服务，数据“面罩”将会流行，而且预计 2013 年将会颁布关于大数据隐私的标准和条例。

- **大数据与云计算等深度融合**

大数据处理离不开云计算技术，云计算为大数据提供弹性可扩展的基础设施支撑环境以及数据服务的高效模式，大数据则为云计算提供了新的商业价值，因此大数据技术与云计算技术必然进入更完美的结合期。总体而言，云计算、物联网、移动互联网等新兴计算形态，既是产生大数据的地方，也是需要大数据分析方法的领域。

- **基于海量数据（知识）的智能**

不久的将来将会有更多基于海量数据（知识）的智能成果出现，甚至有可能

产生人工大脑。至少类似于 Chinese Room 这样的问题将得到彻底解决。因为所有人们能想到的问题，在问之前就都已经被人回答过了，所以，即便在没有思考和逻辑的情况下，也可以利用前人的经验，同样可以起到脑的功能，甚至也可能通过大数据直接进行推理。

- **大数据分析的革命性方法**

在大数据分析上，将出现革命性的新方法。就像计算机和互联网一样，大数据可能是新一波的技术革命。基于大数据的数据挖掘、机器学习和人工智能可能会改变小数据/小世界里的很多算法和基础理论，这方面很可能会产生理论级别的突破。

- **大数据安全**

大数据的安全令人担忧，大数据的保护越来越重要---大数据的不断增加，对数据存储的物理安全性要求会越来越高，从而对数据的多副本与容灾机制提出更高的要求。进入新的世纪，网络和数字化生活使得犯罪分子更容易获得关于人的信息，也有了更多不易被追踪和防范的犯罪手段，可能会出现更高明的骗局，也就是说大数据已经把你出卖。

- **数据科学兴起**

从 2013 年开始，数据科学作为一个与大数据相关的新兴学科出现，将有专门针对数据科学的专业形成，有博士、硕士甚至本科生出现。同时，有大量数据科学的专著被出版。

- **数据共享联盟**

数据共享联盟将在 2013 年逐渐壮大成为产业的核心一环。数据是基础，之前在科技部的支持下，已建立了多个领域的数据共享平台，包括气象、地震、林业、农业、海洋、人口与健康、地球系统科学数据共享平台等。之后，数据共享将扩展到企业层面。

- **大数据新职业**

大数据将催生一批新的就业岗位，如数据分析师、数据科学家等。具有丰富经验的数据分析人才成为稀缺资源，数据驱动型工作机会将呈现出爆炸式的增长。大数据领域最优秀的科学家们纷纷转行股票、期货、甚至赌博（能比别人多看远

一秒钟，就是效益）。

- **更大的数据**

现在的大数据，将来都不够大。大数据将获得更多的关注、研究、开发和应用，所引起的后果是，体现大数据特征的体量大、速度快、模态多、价值密度低等几个 V 的特性将变得更加极致。尤其是大数据的价值密度会越来越低——数据不断地增长，如何去除大数据中的噪声等垃圾数据，进而从中挖掘和提取出有价值信息的难度也随之增大。

6.3 大数据产业的发展重点

6.3.1 构建大数据产业生态环境

目前关心大数据研究、开发、生产和应用的主要是五类人群：一是网络信息服务企业；二是其他各行业的有关领导和信息化工作者，特别是金融、电信和制造业；三是政府部门负责智慧城市和信息化建设的官员；四是信息领域的软硬件制造商和科研开发人员；五是基础研究领域的科研人员。从战略布局考虑，只有充分调动这五方面人员的积极性，才能形成健康发展的产业生态环境。既要重视企业提高经济效益的短期需求，又要重视科研人员的长期基础研究；既要发挥企业的主动性，又要体现政府的宏观规划和政策指引作用。

发展大数据要重视基础设施建设。传统的基础设施是“铁公机（铁路、公路、机场）”，看重有形资源。数据像土地和矿产等有形资产一样也是巨大的财富，大数据时代新的基础设施看重的是无形的数据。中国未来10年不能照搬前10年的“铁公机”那一套建设模式。发展大数据产业必须要有先进的信息化基础设施，大数据与云计算是相伴而行的孪生技术，大数据是云计算的杀手级应用，在发展云计算的过程中一定要重视大数据的获取、分析和应用。

对于大多数希望利用大数据改善管理和提高效率的企业，首先不是购买新设备，扩大数据中心，而是要审查现有的信息资产,发现未处理的“黑暗数据”并确定是否有商业价值,这是大数据战略的第一步目标。2013年3月IBM商业价值研究院和牛津大学赛德商学院共同发表的大数据白皮书提出五项关键建议：以客户

为中心推动初始举措；制订整个企业的大数据蓝图；从现有数据开始，实现近期目标；根据业务优先级逐步建立分析能力；基于可衡量的指标制定业务投资回报分析[7]。这些建议值得重视。

6.3.2 大数据产业的发展重点

并非每个企业都需要具备管理大数据的能力，发展大数据产业要有重点企业、重点产品和重点应用。但是，通过获取和和分析数据提高管理能力和决策水平是对每一个单位的普遍要求，应在全社会提倡数据意识，真正把数据当成宝贵的财富。

在最近3-5年内，我国发展大数据产业应重点抓好以下几件事：

（1）发展基于互联网的大数据公共服务

依托搜索、电商、社交等互联网龙头企业的平台，实行数据开放策略，积极发展面向公众的大数据公共服务。政府应主要着力于建设良好的外部环境，通过法律法规、行业监管和技术标准等手段解决好公平竞争、数据隐私保护等共性问题。

（2）推动大数据行业应用

发展大数据产业应坚持应用为先的原则，可优先考虑以下行业。

- 金融证券业。基于大数据交易的挖掘分析方法可实现系统性的金融风险管控，有针对性地整合分析证券、银行客户资产、上市公司披露信息等大规模数据，综合分析客户的资产负债、支付等状况，帮助评估客户信用等级，探测潜在的交易欺诈和违法行为，从而提高金融风险的可审性和管理力度。
- 医疗卫生：建立覆盖全国的电子病历数据库，促进个性化疾病预防与医疗服务产业的发展，提高医疗质量，降低医疗差错，优化工作流程，改善医患关系，实现全面的疫情监测和快速响应。对于制药行业而言，可以通过药效的比对分析加速新药研制。
- 公共服务和社会管理：在政府开放数据的基础上，积极推进大数据在政务和公共服务领域的应用，特别是在智慧城市建设中要大力推广大数据

技术，惠及大众，提升政府的管理效率和服务水平。

- 智能制造。采用大数据技术可以减少20％至50％的产品开发时间，促进我国制造业的转型升级。

（3）设立大数据重大科技专项

发展大数据产业需要突破大数据存储、处理和应用的关键技术，大幅度提高从大数据中发现价值的能力，力争大数据系统的性价比和性能功耗比均提高100倍以上，摆脱垄断商业软件和硬件的制约。应启动“大数据创新实验平台及示范应用”重大科技项目，可首先在国家网络信息安全、金融、医疗健康、生物信息学等基础研究和政府开放数据的语义互联等领域构建实验平台。

6.4 大数据未来发展的思考与建议

尽管大数据意味着大机遇，但同时也意味着工程技术、管理政策、人才培养等方面的大挑战。只有解决了这些基础性的挑战问题，才能充分利用这个大机遇，得到大数据的大价值。因此，我国亟需在国家层面对大数据给予高度重视，特别需要从政策制定、资源投入、人才培养等方面给予强有力的支持；另一方面，建立良性的大数据生态环境是有效应对大数据挑战唯一出路，需要科技界、工业界以及政府部门在国家政策的引导下共同努力，通过消除壁垒、成立联盟、建立专业组织等途径，建立和谐的大数据生态系统。

6.4.1 促进大数据基础研究的建议

就大数据研究计划与措施，有如下的建议：

（1）优先支持网络大数据研究

大数据涉及物理、生物、脑科学 、医疗、环保、经济、文化、安全等众多领域。网络空间中的数据是大数据的重要组成部分，这类大数据与人的活动密切相关，因此也与社会科学密切相关。而网络数据科学和工程是信息科学技术与社会科学等多个不同领域高度交叉的新型学科方向，对国家的稳定与发展有独特的作用，因此应特别重视与支持网络大数据的研究。大数据涉及应用领域很广，当前大数据的研究应与国计民生密切相关的科学决策、环境与社会管理、金融工程、

应急管理（如疾病防治、灾害预测与控制、食品安全与群体事件）以及知识经济为主要应用领域。

（2）大数据科学的基础研究

无论是外国政府的大数据研究计划，还是国内外大公司的大数据研发，当前最重视的都是大数据分析算法和大数据系统的效率。因此，当工业界把主要精力放在应对大数据的工程技术挑战的时候，科技界应该开始着手关注大数据的基础理论研究。大数据科学作为一个新兴的交叉学科方向，其共性理论基础将来自多个不同的学科领域，包括计算机科学、统计学、人工智能、社会科学等。因此，大数据的基础研究离不开对相关学科的领域知识与研究方法论的借鉴。在大数据的基础研究方面，建议研究大数据的内在机理，包括大数据的生命周期、演化与传播规律，数据科学与社会学、经济学等之间的互动机制，以及大数据的结构与效能的规律性（如社会效应、经济效应等）。在大数据计算方面，研究大数据表示、数据复杂性以及大数据计算模型。在大数据应用基础理论方面，研究大数据与知识发现（学习方法、语义解释），大数据环境下的实验与验证方法，以及大数据的安全与隐私等。

（3）大数据研究的组织方式

2012 年 10 月，中国计算机学会和中国通信学会各自成立了大数据专家委员会，从行业学会的层面来组织和推动大数据的相关产学研用活动。但这还不够，建议科学院、科技部、自然科学基金委共同推动成立一个的组织机构，建立一个大数据科学研究平台，更好地组织大数据的协同创新研究与战略性应用；成立国家级的行业大数据共享联盟，使产业界、科技界以及政府部门都能够参与进来，一方面为学术研究提供基本的数据资源，另一方面为大数据的应用提供理论与技术支持。此外，还需成立国家级的面向大数据研究与应用的开源社区，同时也向国际开源社区的核心团队举荐核心成员，使国际顶级的开源社区能够听到来自中国的“声音”。

（4）大数据研究的资源支持

在资源支持方面，建议启动“中国大数据科学与工程研究计划”，从宏观上对我国的大数据产学研用作出系统全面的短期与长期规划。设立自然科学重大研究计划（基金重大）以及重大基础科学研究项目群（973 项目群或 863 重大项目）

等专项资金，有针对性地资助有关大数据的重大科研活动。此外，国家在大数据平台的构建、典型行业的应用以及研发人才的培养等方面应提供相应的财力、物力与人力支持。

6.4.2 发展大数据产业的政策建议

（1）政府带头，实现等级制数据开放共享

数据既然是一种资源，就应当像管理国土资源、矿产资源一样，有一套有法必依的法规。政府拥有大量数据，如果不开放政府数据，大数据研究和应用就会面临“无米之炊”的窘境。应尽快制定“数据政府”创新应用计划，数年内建成data.cn 政府数据服务网站，实现中央政府和各级地方政府数据的开放共享和综合利用。政府数据的公开与共享是开展大数据研究和推广应用的第一步，必须在限定时间内完成。政府开放的数据应该不受版权或第三方所有权的限制，最好符合开放标准，通过搜索引擎容易找到。根据数据的内容性质，可分等级地控制开放范围。

开放政府数据应遵循以下四项原则。（a）价值导向原则：数据资源具有经济价值和社会价值，对数据资源开发利用者有吸引力；（b）质量保障原则：数据格式应当方便使用，内容及时更新；（c）责权利统一原则：政务数据资源拥有部门承担数据开放的责任，依法明确开放数据的范围。数据用户拥有对其下载后数据的使用行为负责的义务:（d）数字连续性原则：开放的政务数据资源应维护其数字连续性，保证可持续再用。

对所有掌控数据资源国家部门，尤其是地理信息、医疗、交通、教育等公共服务部门，国家应出台约束性或鼓励性政策，在不泄露国家秘密和个人隐私的前提下开放数据，真正实现数据共享。对个人信息，也要积极推动有关的立法工作，探索通过技术标准、行业自律等手段解决法律出台前的个人信息保护问题。

（2）尽早制定国家大数据研究与产业发展规划

应借鉴国外的经验，根据国情和技术发展趋势尽快制定务实而前瞻的大数据研究与产业发展规划。为应对国际金融危机，我国中央和地方政府投资了4万亿元，主要用于传统的基础设施建设，信息基础设施的投入很少。为了适应改变经

济发展方式和调整经济结构的需要，一部分基础设施需要腾笼换鸟。笼要换成数据基础设施，鸟要换成数据。

大数据研究与产业发展规划要加强顶层设计，统筹规划大数据与物联网、云计算与智慧城市建设同步发展。大数据是云计算和物联网的重要应用，也是智能城市建设的关键内容之一。要将提高数据的积累、加工和利用能力作为提高综合国力的标志性目标。规划既要务实，强调应用为先，又要有着眼长远，具有前瞻性。

（3）研究和制定科学务实的大数据评价标准

大数据大数据不是一个可精确量化的产品，而是一系列信息技术的集合，其边界并不清楚。对于不同的应用，数据量从TB 级到PB级都可以称为大数据。采用大数据技术，各个产业都可以增加产量或提高效益。因此，我们无法精确统计大数据产业的产值和增加值，不应当用一个简单的GDP 值来衡量大数据产业的效益。应当研究一种科学合理的评价方法评价各地各企业的大数据技术水平和应用规模。要特别关注数据观念的渗透程度，重视生产生、活方式的改变和普惠大众方面的业绩。

（4）建立数据资产化和数据资产流转体系

数据既然是与土地、矿产一样的资源，就要按照资产属性进行管理。公共的数据资源要共享，可交换的企业和个人数据资源就应按市场法则进行有偿交换。要建立数据资产化的基本标准，让不同机构、不同领域的数据形成规范化资产；建立数据资产访问、连接和共享机制，搭建数据资产交易平台，形成数据流转的层次化体系结构；研究数据资产的所有权、使用权以及价值评估体系，通过市场化模式保障数据资产流转的可行性。

（5）完善法律法规，保障数据安全

清晰界定与国家安全相关的数据，通过法规、标准等方式严格规范国家重要数据的备份和迁移，保障数据安全、可靠。积极推动个人信息保护法律的立法工作，探索通过技术标准、行业自律等手段解决法律出台前的个人信息保护问题。大数据时代只靠技术手段难以保护个人隐私，必须要制定保护个人隐私的法律法规，加大对侵害个人隐私行为的打击力度。同时要加强对个人隐私保护的行政监

管，建立对个人隐私保护的测评机制，推动大数据行业的自律和监督。

（6）加强人才培养

扶持高等学校大数据相关专业的发展，培养数据存储、数据挖掘、数据可视化等方面的专门人才。鼓励高校和企业通过建立联合实验室、研发中心等形式，联合培养理论和实践相结合的大数据专业人才。

参考文献

[1] 李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊，2012, 27(6):647-657.

[2] C. Batini, C. Cappiello, C. Francalanci, A. Maurino. Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR), 2009, 41(3):16.

[3] 张化光，刘金海，郁智博，吴振宁. 过程控制的大数据信息获取与处理技术综述与展望.《中国自动化学会通讯》，2012 年第四期专刊，http://www.caa.org.cn/ccaa.php?to=ccaa/indexcontent.action?Iid=10

[4] 李建中，刘显敏. 大数据的一个重要方面:数据可用性. 计算机研究与发展, 2013, 06: 1147-1162.年 06 期

[5] Weil, S.A., Brandt, S.A., Miller, E.L., Long, D.D.E., Maltzahn, C.. Ceph: A scalable, high-performance distributed file system. In Proceedings of the 7th Symp. on Operating Systems Design and Implementation (OSDI). 2006:307–320

[6] 王元卓，靳小龙, 程学旗. 网络大数据：现状与挑战. 计算机学报，2013，36(6): 1-15.

[7] Ghemawat S, Gobioff H, Leung ST. The Google file system. In Proc. of the 19th ACM Symp. on Operating Systems Principles. New York: ACM Press, 2003:29−43.

[8] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel. Finding a needle in Haystack: Facebook’s photo storage. In Proc. 9th USENIX OSDI, 2010.

[9] Gupta, Y. Kim, and B. Urgaonkar. Dftl: a flash translation layer employing demand-based selective caching of page-level address mappings. In Proc.ASPLOS, ASPLOS XIV, 2009:229–240.

[10] D. Ma, J. Feng, and G. Li. Lazyftl: a page-level flashtranslation layer optimized for nand flash memory. InProc. SIGMOD ’11, 2011.

[11] S. Hardock, I. Petrov, R. Gottstein, A. Buchmann. NoFTL: Database Systems on FTL-less Flash Storage. VLDB 2013 (Demonstrations Track), 2013.

[12] X. Ouyang, D. W. Nellans, R. Wipfel, and D. Flynn.Beyond block I/O: Rethinking traditional storageprimitives. In HPCA, 2011:301–311.

[13] Dean, Jeffrey and Ghemawat, Sanjay.MapReduce: simplified data processing on large clusters. Communications of the ACM. 2008, 3(51-1):107-113.

[14] M. Zaharia, M. Chowdhury, T. Das,et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012: 2-16.

[15] J. Gonzalez, Y. Low, H. Gu. PowerGraph: Distributed graph-parallel computation on natural graphs. Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation, 2012:17-30.

[16] 吴甘沙. 大数据计算范式的分野与交融. 程序员,2013, 9.

[17] 程学旗, 王元卓. 大数据计算的技术体系与引擎系统. 高科技与产业化, 2013, 9 (5): 62-65.

[18] S. Melnik, A. Gubarev, J. Long, et al. Dremel: interactive analysis of web-scale datasets. Proceedings of the VLDB Endowment, 2010,3(1-2):330-339.

[19] C. Engle, A. Lupher, R. Xin, et al. Shark: fast data analysis using coarse-grained distributed memory. Proceedings of the 2012 ACM SIGMOD International Conference on Management, 2012: 689-692.

[20] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed Stream Computing Platform. Proceedings of the 10th International Conference on Data Mining Workshops, 2010: 170-177

[21] M. Zaharia, T. Das, H. Li, et al. Discretized streams: an efficient and fault-tolerant model for stream processing on large clusters. Proceedings of the 4th USENIX conference on Hot Topics in Cloud Ccomputing, 2012:10-16.

[22] Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst. HaLoop: efficient iterative data processing on large clusters. Proceedings of the VLDB Endowment, 2010,3(1-2): 285-296.

[23] Y. Zhang, Q. Gao, L. Gao, et al. iMapReduce: A Distributed Computing Framework for Iterative Computation. Proceedings of the 2011 IEEE

International Symposium on Parallel and Distributed Processing Workshops and PhD Forum, 2011:1112-1121.

[24] J. Ekanayake, H. Li, B. Zhang, et al. Twister: a runtime for iterative MapReduce. Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, 2010:21-25.

[25] G. Malewicz, M. Austern, A. Bik, et al. Pregel: a system for large-scale graph processing. Proceedings of the 2010 international conference on Management of data, 2010:135-146.

[26] B. Shao, H. Wang, Y. Li, et al. Trinity: A Distributed Graph Engine on a Memory Cloud. Proceedings of the 2013 ACM SIGMOD International Conference on Management, 2013: 1-12.

[27] R. Xin, J. Gonalez, M. Franklin.GraphX: A Resilient Distributed Graph System on Spark. Proceedings of the First International Workshop on Graph Data Management Experience and System, 2013:12-18.

[28] 覃雄派，王会举，杜小勇，王珊. 大数据分析——RDBMS 与 MapReduce 的竞争与共生. 软件学报, 2012, 23(1):32−45.

[29] Das S, Sismanis Y, Beyer KS, Gemulla R, Haas PJ, McPherson J. Ricardo: Integrating R and Hadoop. In Proc. of the SIGMOD, 2010: 987−998.[doi: 10.1145/1807167.1807275].

[30] Wegener D, Mock M, Adranale D, Wrobel S. Toolkit-Based high-performance data mining of large data on MapReduce clusters. In Proc. of the ICDM Workshop, , 2009: 296−301. [doi: 10.1109/ICDMW.2009.34]

[31] Miliaraki I, Berberich K, Gemulla R, et al. Mind the gap: large-scale frequent sequence mining. SIGMOD '13, 2013: 797-808.

[32] Ene A, Im S, Moseley B. Fast clustering using MapReduce. KDD '11, 2011: 681-689.

[33] Chang K, Roth D. Selective block minimization for faster convergence of limited memory large-scale linear models. KDD '11, 2011: 699-707.

[34] Kang U, Chau D H, Faloutsos C. Mining large graphs: Algorithms, inference, and discoveries. In Data Engineering (ICDE), 2011 IEEE 27th International

Conference on, 2011: 243-254.

[35] Mondal J, Deshpande A. Managing large dynamic graphs efficiently. SIGMOD '12, 2012: 145-156.

[36] Yang S, Yan X, Zong B, et al. Towards effective partition management for large graphs. SIGMOD '12, 2012: 517-528.

[37] Chris Johnson, Robert Moorhead, Tamara Munzner, Hanspeter Pfister, Penny Rheingans, and Terry S. Yoo. “NIH/NSF Visualization Research Challenges Report”. IEEE Computer Society Press, 2006.

[38] E. Wes Bethel, Hank Childs, Charles Hansen. High Performance Visualization. CRC Press. 2012.

[39] Pak Chung Wong，Han-Wei Shen，Christopher R. Johnson，Chaomei Chen，Robert B. Ross. The Top 10 Challenges in Extreme-Scale Visual Analytics. IEEE CG&A, 2012: 63-67.

[40] Blum, A., et al. A learning theory approach to noninteractive database privacy. J. ACM, 2013, 60(2): 1-25

[41] Bainbridge, W. S. Privacy and property on the net: Research questions. Science, 2003, 302(5651): 1686-1687

[42] Weiser, M. Critical Issues and Information Security and Managing Risk. The 9th International Conference on Computing and InformationTechnology (IC2IT2013), Springer, 2013.

[43] Bhatti, R., et al. Emerging trends around big data analytics and security: Panel. Proceedings of the 17th ACM symposium on Access Control Models and Technologies, ACM, 2012.

[44] Schadt, E. E. The changing privacy landscape in the era of big data. Mol Syst Biol, 2012, 8(1).

[45] Anderson, R. and T. Moore The economics of information security. Science, 2006, 314(5799): 610-613.

[46] Howe, D., et al. Big data: The future of biocuration. Nature, 2008, 455(7209): 47-50.

[47] Marx, V. Biology: The big challenges of big data. Nature, 2013, 498(7453):

255-260.

[48] Condie, T., et al. Machine learning for big data. Proceedings of the 2013 international conference on Management of data, ACM, 2013.

[49] Lejun Fan, Yuanzhuo Wang, Xiaolong Jin, Jingyuan Li, Xueqi Cheng, Shuyuan Jin. Comprehensive Quantitative Analysis on Privacy Leak Behavior， PLOS ONE, 2013.[DOI: 10.1371/journal.pone.0073410]

[50] Shim, K. MapReduce algorithms for Big Data analysis. Proceedings of the VLDB Endowment, 2012, 5(12): 2016-2017.

[51] 司莉，邢文明.国外科学数据管理与共享政策调查及对我国的启示. 情报资料工作, 2013, 34(1): 61-66.

中国计算机学会大数据专家委员会

地址：北京市海淀区中关村科学院南路6号　邮编：100190

电话：010-6260-0905

邮箱：bigdata@ccf.org.cn