大数据技术进展与发展趋势

程学旗,靳小龙,杨婧,徐君

中国科学院计算技术研究所,中国科学院网络数据科学与技术重点实验室, 北京 100190

摘要 随着IT技术的高速发展,世界范围内各行各业都在进行信息化变革,几乎每个行业都在努力发现和利用大数据的价值。为了充分利用大数据带来的机遇,同时有效应对大数据带来的挑战,国内外产业界、科学界和政府部门都在积极布局、制定战略规划。本文介绍大数据背景与动态,描述各国大数据政策实践及中国大数据发展的政策环境和产业界生态发展状况;阐述大数据技术的进展,梳理其生态体系和创新特点;提出大数据可视化、多学科融合、安全与隐私、深度分析等发展趋势和相关建议。

关键词 大数据技术;信息技术;大数据生态体系

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IDC(国际数据公司)的监测统计,2011年全球数据总量已经达到1.8 ZB,而这个数值还在以每2年翻一番的速度增长,预计到2020年,全球将总共拥有35 ZB的数据量,比2011年增长了近20倍。换句话说,近2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[1,2]。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。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。

大数据时代的2个特点非常有利于中国信息产业跨越式发展[3]。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,也亟待政府、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。大数据超越信息技术,使人们重新界定国家竞争的主战场,重新审视政府治理水平,重新认识科学研究的新范式,重新审视产业变迁的驱动因素,重新理解投资的决策依据,重新思考公司的战略和组织结构。

1 国内外大数据发展动态

1.1 国际大数据战略决策

纵观世界各国的大数据策略,存在3个共同点:一是推动大数据全产业链的应用;二是数据开放与信息安全并重;三是政府与社会力量共同推动大数据应用[4]。本文以美国、英国、日本、德国4个国家为例具体说明。

1)美国。2009年,美国政府推出公共服务平台(data.gov),全面开放了40万联邦政府原始数据和地理数据。2012年3月,美国白宫科技政策办公室发布《大数据研究和发展计划》,成立“大数据高级指导小组”。通过对海量数据分析萃取信息,提升对社会经济发展的预测能力。美国国家科学基金会、国家卫生研究院、国防部、能源部、国防部高级研究局、地质勘探局6个联邦部门和机构宣布投资2亿美元,共同提高收集、储存、保留、管理、分析和共享海量数据所需核心技术的先进性,并形成合力;加强对信息技术研发投入以推动超级计算和互联网的发展。2013年,美国发布《政府信息公开和机器可读行政命令》,要求公开教育、健康等七大关键领域数据,并对各政府机构数据开放时间提出了明确要求。2013年11月,美国信息技术与创新基金会发布《支持数据驱动型创新的技术与政策》指出,政府不仅要大力培养所需技能劳动力和推动数据相关技术研发,还要制定推动数据共享的法律框架,并提高公众对

收稿日期:2016-05-30;修回日期: 2016-06-28

基金项目:国家重点基础研究发展计划(973计划)项目(2013CB329602,2014CB340401);国家自然科学基金面上项目(61572473);国家杰出青年科学基金项目(61425016);国家自然科学基金青年科学基金项目(61303049)

作者简介:程学旗,研究员,研究方向为网络科学与社会计算、互联网搜索与挖掘、网络信息安全等,电子信箱:cxq@ict.ac.cn;杨婧(通信作者),博士,研究方向为最优化查询、数据挖掘等,电子信箱:jyang@ict.ac.cn

引用格式:程学旗, 靳小龙, 杨婧, 等. 大数据技术进展与发展趋势[J]. 科技导报, 2016, 34(14): 49-59; doi:10.3981/j.issn.1000-7857.2016.14.006

数据共享重大意义的认识。2014年5月,美国发布《大数据:把握机遇,守护价值》白皮书,对美国大数据应用与管理的现状,政策框架和改进建议进行集中阐述。2016年4月,麻省理工学院推出了“数据美国”在线大数据可视化工具,可以实时分析展示美国政府公开数据库(Open Data)。

2)英国。2011年11月,英国政府发布了对公开数据进行研究的战略决策,建立了有“英国数据银行”之称的data.gov.uk网站,希望通过完全公布政府数据,进一步支持和开发大数据技术在科技、商业、农业等领域的发展。2012年5月,英国政府注资10万英镑,支持建立了世界上首个开放数据研究所ODI(Open Data Institute)。ODI研究所将为那些对公众有益的商业企业活动提供数据背景支持,不但释放了新的商业潜力,还推动了经济发展以及个人收入增长的新形式。2013 年5月,英国政府和李嘉诚基金会联合投资9000万英镑,在牛津大学成立全球首个综合运用大数据技术的医药卫生科研中心。中心将通过搜集、存储和分析大量生物医疗数据,与业界共同界定新药物研发方向,处理新药研发过程中的瓶颈,并为发现新的治疗手段提供线索。2013 年8月,英国政府发布《英国农业技术战略》。该战略指出,英国今后对农业技术的投资将集中在大数据上,目标是将英国的农业科技商业化。2014年,英国政府投入7300万英镑进行大数据技术的开发,包括在55个政府数据分析项目中展开大数据技术的应用;以高等学府为依托投资兴办大数据研究中心,如图灵大数据研究院。2015年,英国政府承诺将开放有关交通运输、天气和健康方面的核心公共数据库。

3)日本。2012年6月,日本IT战略本部发布电子政务开放数据战略草案,迈出了政府数据公开的关键一步。2012年7月,日本总务省ICT基本战略委员会发布了《面向2020年的ICT综合战略》,提出“活跃在ICT领域的日本”的目标,将重点关注大数据应用所需的社会化媒体等智能技术开发、传统产业IT创新、新医疗技术开发、缓解交通拥堵等公共领域应用等。2013年6月,日本正式公布新IT战略-创建最尖端IT国家宣言。全面阐述2013—2020年期间以发展开放公共数据和大数据为核心的日本新IT国家战略,提出要把日本建设成为具有世界最高水准的广泛运用信息产业技术的社会。为此,日本政府推出数据分类网站(data.go.jp),目的是提供不同政府部门和机构的数据供使用,向数据提供者和数据使用者开放数据。数据涉及各类白皮书、地理空间信息、人群运动信息、预算、年终财务和流程数据等。2013年7月,日本三菱综合研究所牵头成立了“开放数据流通推进联盟”,旨在由产官学联合,促进日本公共数据的开放应用。2014年8月,日本内阁府决定在每月公布的月度经济报告中采用互联网上累积的“大数据”作为新的经济判断指标。内阁府将根据网络用户对产品和服务的搜索情况和推特网站上所发帖子来分析实时消费动向。日本防卫省也将从2015年开始正式研讨将 “大数据”运用于海外局势的分析。这一举措作为自卫队海外活动扩大背景下的新方案,旨在强化情报收集能力。

4)德国。2010年,德国制定“数字德国2015的ICT 战略”,在能源、交通、保健、教育、休闲、旅游和管理等传统行业采用现代ICT 技术实现智能网络化。2013年4月,德国政府提出了“工业4.0”的概念。该项目德国联邦政府投入2亿欧元,由德国联邦教研部与联邦经济技术部联手资助,在德国工程院、弗劳恩霍夫协会、西门子公司等德国学术界和产业界的建议和推动下形成,并已上升为国家级战略。德国IT行业协会BITKOM 于2014年初发表报告称,大数据业务在德国发展迅速,到2016年有望达到136亿欧元。2014年8月20日,德国联邦政府内阁通过了由德国联邦经济和能源部、内政部、交通与数字基础设施建设部联合推出的《2014—2017年数字议程》,提出在变革中推动“网络普及”“网络安全”“数字经济发展”3个重要进程,希望以此打造具有国际竞争力的“数字强国”。

由此可见,大数据超越信息技术、使人们重新界定国家竞争的主战场,重新审视政府治理水平,重新认识科学研究的新范式,重新审视产业变迁的驱动因素,重新理解投资的决策依据,重新思考公司的战略和组织结构。

1.2 国际大数据产业变革

2013年6月,美国中央情报局前雇员斯诺登揭开了“数据战争”的冰山一角。美国的“棱镜计划”事实上把所有国家、个人都纳在美国国家安全局(NSA)的监控之下。参与棱镜计划的公司包括谷歌、雅虎、Facebook、微软、苹果、思科、Oracle、IBM等科技巨头。由此可见,在大数据时代,IT产业的强大已经成为直接决定一个大国是否成为强国的最为关键的因素之一。

产业需要变革,行业需要互通互融。所谓“大数据+”,就是将大数据思维嫁接到不同的产业中,推动大数据在各行各业落地。大数据不仅只关系到IT行业,而且众多行业龙头公司都已经意识到了大数据新思维的巨大冲击。互联网、金融、电信、医疗、政府等是大数据运营的重点领域。而大多数领域的大数据发展应用仍处在初级阶段,在大数据应用的实践过程中也遇到了数据资产不明、应用需求不定、平台建设、技术路线、安全隐私问题等方面的挑战,但是大数据应用在各领域还是做出了一些有益的探索,并取得了一定的成绩。

在电信行业,一些发达国家电信运营商对大数据的利用,一方面提升服务质量,改善内部管理,包括客户维系、精准营销和网络运营与管理,代表企业分别为法国电信、英国O2、NTT DoCoMo和沃达丰。法国电信开展针对用户消费的大数据分析评估,借助大数据改善服务水平,提升用户体验;英国O2在英国推出了免费WiFi 服务,以积累更多的用户,从而收集到更多的用户数据,用在精准的媒体广告和营销服务方面;NTT DoCoMo通过制作精细化表格,收集用户详细信息,大大加强了CRM系

统和知识库,准确定位目标客户,提高了业务办理的成功性;沃达丰爱尔兰公司的Tellabs“洞察力分析”服务是将通信网络中的大数据转化为可利用的情报。另一方面确立商业模式,创造外部收益,包括直接出售数据获取收益,以及与第三方公司合作项目给运营商创造盈利,代表企业有AT&T、西班牙电信、Dynamic Insights、Verizon、德国电信和沃达丰。AT&T将与用户相关的数据出售给政府和企业以获利;西班牙电信成立了动态洞察部门;Dynamic Insights开展大数据业务,为客户提供数据分析打包服务,与市场研究机构GFK进行合作,在英国、巴西推出了首款产品名为智慧足迹(Smart Steps);Verizon成立了精准营销部门Precision Marketing Division,提供了精准营销洞察、精准营销、移动商务等服务,包括联合第三方机构对其用户群进行大数据分析,再将有价值的信息提供给政府或企业获取额外价值,数据业务的盈利在其整个业务中占比非常高;德国电信和沃达丰主要尝试通过开放API向数据挖掘公司等合作方提供部分用户匿名地理位置数据,以掌握人群出行规律,有效地与一些LBS应用服务对接。

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

在交通运输方面,美国Inrix公司和新泽西州运输部达成合作伙伴关系。Inrix公司通过汽车和移动电话GPS装置上的信号和数据,采集主干道上的车速数据,然后实时向新泽西州运输部警示任意主干道上的路况险情,同时向司机的车载GPS装置或移动电话发送警示提醒司机注意路况险情。这个项目现已扩展为跨州服务,覆盖范围包括马里兰州和北卡罗来纳州。

在农业方面,美国天气保险公司(Climate Corporation)可以为美国的农民提供天气意外保险,农民朋友可以在电脑上模拟未来可能破坏农业生产的天气,然后选择合适的保险进行投保,这样在未来发生灾害时损失可以降低到最少。该公司通过庞大的传感器网络分析和预测2000万美国农田的气温、降水、土壤湿度和产量。在知晓高温天的天数以及土壤湿度数据后,建立模型帮助其预判农民需要的天气保险金额以及公司需要支付的保费。

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

在外包领域,大数据技术也已成为信息技术行业的“下一个大事件”。目前,一些外包行业巨头也开始进军大数据市场,试图瓜分这一块大蛋糕。印度全国软件与服务企业协会预计,印度大数据行业规模在3年内将达到12亿美元,是目前规模的6倍,同时还是全球大数据行业平均增长速度的2倍。

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

在人与机器的围棋大战中,AlphaGo击败李世石的事实再次展示了大数据应用产业的巨大潜力。通过大数据掌握消费习惯,摸准产业发展脉络,提供有效供给,已成为当前产业转型升级的方式之一。

综上所述,数据资产可以成为任何产业的最核心竞争力。未来几年,随着数据中心等基础设施建设的落地,大数据市场将进一步向软件和服务端拓展,深度融合多个产业。对大数据的价值挖掘也将进入快速发展期,为不同行业的需求提供差异化的服务。

1.3 中国大数据发展态势

1.3.1 中国政府促进大数据发展的措施

随着信息技术的高速发展,世界范围内各行各业都在进行信息化变革,几乎每个行业都在努力发现和利用大数据的价值。为了充分利用大数据带来的机遇,同时有效应对大数据带来的挑战,中国产业界、科技界和政府部门也在积极布局、制定战略规划。

2012年8月,国务院制定了促进信息消费扩大内需的文件,推动商业企业加快信息基础设施演进升级,增强信息产品供给能力,形成行业联盟,制定行业标准,构建大数据产业链,促进创新链与产业链有效嫁接。同时,构建大数据研究平台,整合创新资源,实施“专项计划”,突破关键技术。工业和信息化部为鼓励和推进大数据产业发展也制定了3大措施:一是在已通过促进信息消费扩大内需的意见、软件和信息技术服务业“十二五”规划等政策规划中,对大数据发展进行了部署;二是推动全国信息技术标准化技术委员会开展了大数据标准化的需求分析、标准体系框架研究及相关标准研制工作,并向相关国际标准化组织提交了大数据研究提案;三是利用项目资金等手段进行了前沿部署,支持关键技术产品的研发和产业化。

2015年8月,国务院发布《促进大数据发展行动纲要》(以下简称为《纲要》),这是指导中国大数据发展的国家顶层设计和总体部署。《纲要》明确指出了大数据的重要意义,大数据成为推动经济转型发展的新动力、重塑国家竞争优势的新机遇、提升政府治理能力的新途径。《纲要》清晰地提出了大数据发展的主要任务:加快政府数据开放共享,推动资源整合,提升治理能力;推动产

业创新发展,培育新兴业态,助力经济转型;强化安全保障,提高管理水平,促进健康发展。《纲要》还提出了组织、法规、市场、标准、财政、人才、国际交流等几方面的政策机制要求。《纲要》的出台,进一步凸显大数据在提升政府治理能力、推动经济转型升级中的关键作用。"数据兴国"和"数据治国"已上升为国家战略,将成为中国今后相当长时期的国策。未来,大数据将在稳增长、促改革、调结构、惠民生中发挥越来越重要的作用。

1.3.2 大数据基础研究列为中国战略研究主题

2012年,国家重点基础研究发展计划(973计划)专家顾问组在前期项目部署的基础上,将大数据基础研究列为信息科学领域4个战略研究主题之一。2013年,973计划将"面向网络信息空间大数据计算的基础研究"列为指南的重要支持方向。2014年,科技部基础研究司在北京组织召开"大数据科学问题"研讨会,邀请有关专家围绕973计划大数据研究布局、中国大数据发展战略、国外大数据研究框架与重点、大数据研究关键科学问题、重要研究内容和组织实施路线图等展开研讨,并对数据挖掘与管理、深度学习、大数据智能、大数据与其他学科的交叉等进行了深入交流。大数据对国家科技、经济、社会发展意义重大,应加强顶层设计,研究制定大数据研究的战略布局和实施路线图,推动学科交叉,拓展并提升我国大数据科学的研究能力和应用水平。

近两年,国家重点基础研究和高技术发展计划大力支持大数据重大建设项目,由北京航空航天大学承担的"网络信息空间大数据计算理论"、中国科学院计算技术研究所承担的"网络大数据计算的基础理论及其应用研究"、清华大学承担的"面向城市管理的三元空间大数据计算理论与方法""大数据群体计算的基础理论与关键技术"、上海交通大学承担的"城市大数据三元空间协同计算理论与方法"、山东大学承担的"城市大数据计算理论和方法"等项目获得973计划支持。由上海交通大学承担的"面向大数据的内存计算关键技术与系统"、电子科技大学承担的"初等数学问题求解关键技术及系统"、科大讯飞承担的"基于大数据的类人智能关键技术与系统"、国网上海电力公司承担的"智能配用电大数据应用关键技术"、哈尔滨工业大学承担的"生物大数据开发与利用关键技术研究"、中山大学肿瘤防治中心承担的"常见恶性肿瘤大数据处理分析与应用研究"等项目获得国家高技术研究发展计划(863计划)支持。2014年获批的国家自然科学基金立项项目,"项目主题词"含"大数据"共144条,其中200万以上经费的项目有18个。

2016年,国家发改委正式印发《关于组织实施促进大数据发展重大工程的通知》(以下简称《通知》)。《通知》称,将重点支持大数据示范应用、共享开放、基础设施统筹发展,以及数据要素流通。同时将择优推荐项目进入国家重大建设项目库审核区,并根据资金总体情况予以支持。国家重点支持的项目,包括社会治理大数据应用、公共服务大数据应用,以及产业发展大数据应用、创业创新大数据应用等。《通知》还提到,将组织大数据开放计划,开展大数据全民创新竞赛。建立统一的公共数据共享开放平台体系,以及整合分散的政务数据中心,并首次提到了探索构建国家数据中心体系开展绿色数据中心试点。同时,在最受业界关注的大数据交易方面,《通知》也提到,将重点支持数据要素流通,建立完善国家大数据标准体系,依托已建的大数据交易所,探索建立大数据交易平台,提供丰富的数据产品、交易模式等方面的规范制度[5]。

1.3.3 中国大数据产业强势增长

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

百度、阿里巴巴、奇虎360、京东等互联网企业依靠自身的数据优势,均已将大数据作为公司的重要战略。大数据正在从理论走向实践,从专业领域走向全民应用的阶段。百度在大数据方面让人印象深刻的有百度迁徙这样的公益项目,应用在民生和新闻等领域。百度网盟利用基于大数据的CTR(广告内容匹配)数据使站长的平均收入提升70%。阿里巴巴集团宣布无线开放战略,启动百川计划。该计划将全面分享阿里无线资源,为移动开发者提供技术、数据、商业等全链条基础设施服务。其中,大数据层面则将联合移动应用统计分析平台友盟,帮助开发者完善数据精准挖掘分析及完善个性化推送体系。

奇虎360举办首届数字世界大会,并发布实效平台、聚效平台和来店通等3款产品,把集合了数10亿用户信息的数据免费分享给广告主,帮助广告商利用大数据做更有效的营销。京东也在积极通过大数据技术挖掘用户需求,提供更精准的服务。借助微信能够带来巨大流量的优势,举行京东微信购物的众筹活动,一个月参与人数就达到40万人次。

综上所述,国内大数据产业起步较晚,同时由于互联网技术也有所滞后,使得中国的大数据发展较领先国家还尚有一段距离。但是,中国又有得天独厚的优势——庞大的用户群,每日有庞大的数据量不断生成,同时受惠用户量也极为众多。中国电子信息产业发展研究院赛迪顾问预测,2016年,中国大数据产业还将保持强势增长态势,大数据市场年复合增长率有望达到30%以上。对大数据的价值挖掘将快速渗透到产业的方方面面。从政策环境上看,中国在数据开放的过程中仍然存在安全隐患,需要健全的法律法规以及先进的数据安全技术作保障。研究机构提醒有关部门研究制定网络数据采集、传输、存储、使用管理的标准规范,加大对

隐私信息保护、网络安全保障、跨境数据流动的管理等。在政府数据开放方面也亟需进一步加强。

2 大数据技术进展

目前,大数据领域每年都会涌现出大量新的技术,成为大数据获取、存储、处理分析或可视化的有效手段。大数据技术能够将大规模数据中隐藏的信息和知识挖掘出来,为人类社会经济活动提供依据,提高各个领域的运行效率,甚至整个社会经济的集约化程度。

2.1 大数据生命周期

图1展示了一个典型的大数据技术栈。底层是基础设施,涵盖计算资源、内存与存储和网络互联,具体表现为计算节点、集群、机柜和数据中心。在此之上是数据存储和管理,包括文件系统、数据库和类似YARN的资源管理系统。然后是计算处理层,如Hadoop[6]、MapReduce[7]和Spark[8],以及在此之上的各种不同计算范式,如批处理、流处理和图计算等,包括衍生出编程模型的计算模型,如BSP、GAS等。数据分析和可视化基于计算处理层。分析包括简单的查询分析、流分析以及更复杂的分析(如机器学习、图计算等)。查询分析多基于表结构和关系函数,流分析基于数据、事件流以及简单的统计分析,而复杂分析则基于更复杂的数据结构与方法,如图、矩阵、迭代计算和线性代数。一般意义的可视化是对分析结果的展示。但是通过交互式可视化,还可以探索性地提问,使分析获得新的线索,形成迭代的分析和可视化。基于大规模数据的实时交互可视化分析以及在这个过程中引入自动化的因素是目前研究的热点。

有2个领域垂直打通了上述的各层,需要整体、协同地看待。一是编程和管理工具,方向是机器通过学习实现自动最优化、尽量无需编程、无需复杂的配置。另一个领域是数据安全,也是贯穿整个技术栈。除了这两个领域垂直打通各层,还有一些技术方向是跨了多层的,例如"内存计算"事实上覆盖了整个技术栈。

图1 大数据技术栈

2.2 大数据技术生态

大数据的基本处理流程与传统数据处理流程并无太大差异,主要区别在于:由于大数据要处理大量、非结构化的数据,所以在各处理环节中都可以采用并行处理。目前,Hadoop[6]、MapReduce[7]和Spark[8]等分布式处理方式已经成为大数据处理各环节的通用处理方法。

Hadoop是一个由Apache基金会开发的大数据分布式系统基础架构。用户可以在不了解分布式底层细节的情况下,轻松地在Hadoop上开发和运行处理大规模数据的分布式程序,充分利用集群的威力高速运算和存储。Hadoop是一个数据管理系统,作为数据分析的核心,汇集了结构化和非结构化的数据,这些数据分布在传统的企业数据栈的每一层。Hadoop也是一个大规模并行处理框架,拥有超级计算能力,定位于推动企业级应用的执行。Hadoop又是一个开源社区,主要为解决大数据的问题提供工具和软件。虽然Hadoop提供了很多功能,但仍然应该把它归类为多个组件组成的Hadoop生态圈,这些组件包括数据存储、数据集成、数据处理和其他进行数据分析的专门工具。图2展示了Hadoop的生态系统,主要由HDFS、MapReduce、Hbase、Zookeeper、Oozie、Pig、Hive等核心组件构成,另外还包括Sqoop、Flume等框架,用来与其他企业融合。同时,Hadoop生态系统也在不断增长,新增Mahout、Ambari、Whirr、BigTop等内容,以提供更新功能[9]。

低成本、高可靠、高扩展、高有效、高容错等特性让Hadoop成为最流行的大数据分析系统,然而其赖以生存的HDFS和MapReduce组件却让其一度陷入困境——批处理的工作方式让其只适用于离线数据处理,在要求实时性的场景下毫无用武之地。因此,各种基于Hadoop的工具应运而生。为了减少管理成本,提升资源的利用率,有当下众多的资源统一管理调度系统,例如Twitter的Apache Mesos、Apache的YARN、Google的Borg、腾讯搜搜的Torca、Facebook Corona(开源)等。Apache Mesos是Apache孵化器中的一个开源项目,使用ZooKeeper实现容错复制,使用Linux Containers来隔离任务,支持多种资源计划分配(内存和CPU)。提供高效、跨分布式应用程序和框架的资源隔离和共享,支持Hadoop、MPI、Hypertable、Spark等。YARN又被称为MapReduce 2.0,借鉴Mesos,YARN提出了资源隔离解决方案Container,提供Java虚拟机内存的隔离。对比MapReduce 1.0,开发人员使用ResourceManager、Application Master与NodeManager代替了原框架中核心的JobTracker和TaskTracker。在YARN平台上可以运行多个计算框架,如MR、Tez、Storm、Spark等。

基于业务对实时的需求,有支持在线处理的Storm、Cloudar Impala、支持迭代计算的Spark及流处理框架S4。Storm是一个分布式的、容错的实时计算系统,由BackType开发,后被Twitter捕获。Storm属于流处理平台,多用于实时计算并更新数据库。Storm也可被用于"连续计算"(Continuous Computation),对数据流做连续查询,在计算时就将结果以流的形式输出给用

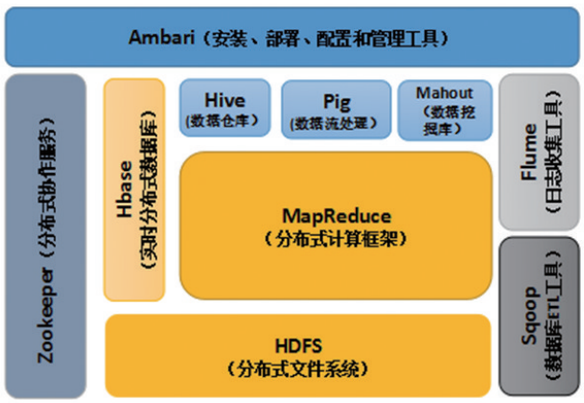

图2　Hadoop生态系统

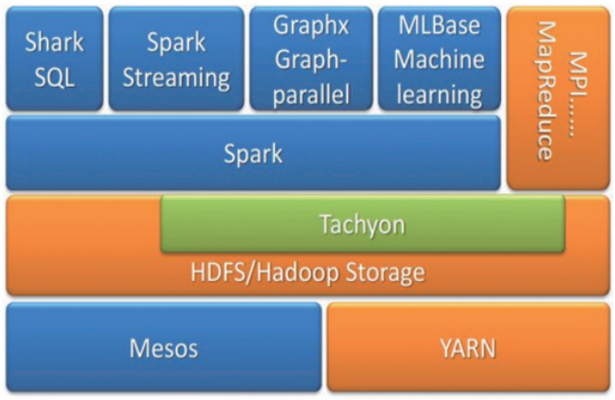

图3　Spark生态系统

户。它还可被用于“分布式RPC”,以并行的方式运行昂贵的运算。Cloudera Impala是由Cloudera开发,一个开源的Massively Parallel Processing(MPP)查询引擎。与Hive相同的元数据、SQL语法、ODBC驱动程序和用户接口(Hue Beeswax),可以直接在HDFS或HBase上提供快速、交互式SQL查询。Impala是在Dremel的启发下开发的,不再使用缓慢的Hive+MapReduce批处理,而是通过与商用并行关系数据库中类似的分布式查询引擎(由Query Planner、Query Coordinator和Query Exec Engine这3部分组成),可以直接从HDFS或者HBase中用SELECT、JOIN和统计函数查询数据,从而大大降低了延迟。

Hadoop社区正努力扩展现有的计算模式框架和平台,以便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Hadoop 2.0版本“YARN”的努力目标。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UC Berkeley AMP Lab开发的Spark生态系统,如图3所示。Spark是开源的类Hadoop MapReduce的通用的数据分析集群计算框架,用于构建大规模、低延时的数据分析应用,建立于HDFS之上。Spark提供强大的内存计算引擎,几乎涵盖了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使用Scala作为应用框架,采用基于内存的分布式数据集,优化了迭代式的工作负载以及交互式查询。与Hadoop不同的是,Spark和Scala紧密集成,Scala像管理本地collective对象那样管理分布式数据集。Spark支持分布式数据集上的迭代式任务,实际上可以在Hadoop文件系统上与Hadoop一起运行(通过YARN、Mesos等实现)。另外,基于性能、兼容性、数据类型的研究,还有Shark、Phoenix、Apache Accumulo、Apache Drill、Apache Giraph、Apache Hama、Apache Tez、Apache Ambari等其他开源解决方案。预计未来相当长一段时间内,主流的Hadoop平台改进后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。

2.3　大数据采集与预处理

在大数据的生命周期中,数据采集处于第一个环节。根据MapReduce产生数据的应用系统分类,大数据的采集主要有4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。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。对多个异构的数据集,需要做进一步集成处理或整合处理,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。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[10],例如,美国SAS公司的Data Flux、美国IBM公司的Data Stage、美国Informatica公司的Informatica Power Center。

2.4　大数据存储与管理

传统的数据存储和管理以结构化数据为主,因此关系数据库系统(RDBMS)可以一统天下满足各类应用需求。大数据往往是半结构化和非结构化数据为主,结构化数据为辅,而且各种大数据应用通常是对不同类型的数据内容检索、交叉比对、深度挖掘与综合分析。面对这类应用需求,传统数据库无论在技术上还是功能上都难以为继。因此,近几年出现了oldSQL、NoSQL与NewSQL并存的局面。总体上,按数据类型的不同,大数据的存储和管理采用不同的技术路线,大致可以分为3类。第1类主要面对的是大规模的结构化数据。针对这类大数据,通常采用新型数据库集群。它们通过列存储或行列混合存储以及粗粒度索引等技术,结合MPP(Massive Parallel Processing)架构高效的分布式计算模式,实现对PB量级数据的存储和管理。这类集群具有高性能和高扩展性特点,在企业分析类应用领域已获得广泛应用;第2类主要面对的是半结构化和非结构化数据。应对这类应用场景,

基于Hadoop开源体系的系统平台更为擅长。它们通过对Hadoop生态体系的技术扩展和封装,实现对半结构化和非结构化数据的存储和管理;第3类面对的是结构化和非结构化混合的大数据,因此采用MPP并行数据库集群与Hadoop集群的混合来实现对百PB量级、EB量级数据的存储和管理。一方面,用MPP来管理计算高质量的结构化数据,提供强大的SQL和OLTP型服务;另一方面,用Hadoop实现对半结构化和非结构化数据的处理,以支持诸如内容检索、深度挖掘与综合分析等新型应用。这类混合模式将是大数据存储和管理未来发展的趋势。

2.5 大数据计算模式与系统

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[7],加州大学伯克利分校著名的Spark系统中的"分布内存抽象RDD"[8],CMU著名的图计算系统GraphLab中的"图并行抽象"(Graph Parallel Abstraction[11])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。

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[12]。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,如表1[13~22]所示,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[23]。

表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

2.6 大数据分析与可视化

在大数据时代,人们迫切希望在由普通机器组成的大规模集群上实现高性能的以机器学习算法为核心的数据分析,为实际业务提供服务和指导,进而实现数据的最终变现。与传统的在线联机分析处理OLAP不同,对大数据的深度分析主要基于大规模的机器学习技术,一般而言,机器学习模型的训练过程可以归结为最优化定义于大规模训练数据上的目标函数并且通过一个循环迭代的算法实现,如图4所示。因而与传统的OLAP相比较,基于机器学习的大数据分析具有自己独特的特点[24]。

1)迭代性:由于用于优化问题通常没有闭式解,因而对模型参数确定并非一次能够完成,需要循环迭代多次逐步逼近最优值点。

2)容错性:机器学习的算法设计和模型评价容忍非最优值点的存在,同时多次迭代的特性也允许在循环的过程中产生一些错误,模型的最终收敛不受影响。

3)参数收敛的非均匀性:模型中一些参数经过少数几轮迭代后便不再改变,而有些参数则需要很长时间才能达到收敛。

这些特点决定了理想的大数据分析系统的设计和其他计算系统的设计有很大不同,直接应用传统的分布式计算系统应用于大数据分析,很大比例的资源都浪费在通信、等待、协调等非有效的计算上。

传统的分布式计算框架MPI(message passing interface,信息传递接口)[25]虽然编程接口灵活功能强大,但由于编程接口复杂且对容错性支持不高,无法支撑在大规模数据上的复杂操作,研究人员转而开发了一系列接口简单容错性强的分布式计算框架服务于大数据分析算法,以MapReduce[7]、Spark[8]和参数服务器Parameter Server[26]等为代表。

分布式计算框架MapReduce[7]将对数据的处理归结为Map和Reduce两大类操作,从而简化了编程接口并且提高了系统的容错性。但是MapReduce受制于过于简化的数据操作抽象,而且不支持循环迭代,因而对复杂的机器学习

目标函数: $\vec{\theta}^{*}=\mathrm{argmax}_{\vec{\theta}}\mathcal{L}\left(\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^{N};\vec{\theta}\right)+\Omega(\vec{\theta})$

迭代优化:

$\vec{\theta}^{*}$ ←随机值;
for$(t=1\ \mathbf{TO}\ T)$
{
　　其他操作;
　　$\vec{\theta}^{(t)}\leftarrow g\left(\vec{\theta}^{(t-1)},\frac{\partial\mathcal{L}}{\partial\vec{\theta}}\Big|_{\vec{\theta}=\vec{\theta}^{(t-1)}}\right)$;
　　其他操作;
}
return $\vec{\theta}^{(T)}$;

图4 基于机器学习的大数据分析算法目标函数和迭代优化过程

算法支持较差,基于MapReduce的分布式机器学习库Mahout需要将迭代运算分解为多个连续的Map和Reduce操作,通过读写HDFS文件方式将上一轮次循环的运算结果传入下一轮完成数据交换。在此过程中,大量的训练时间被用于磁盘的读写操作,训练效率非常低效。为了解决MapReduce上述问题,Spark[8]基于RDD定义了包括Map和Reduce在内的更加丰富的数据操作接口。不同于MapReduce的是Job中间输出和结果可以保存在内存中,从而不再需要读写HDFS,这些特性使得Spark能更好地适用于数据挖掘与机器学习等需要迭代的大数据分析算法。基于Spark实现的机器学习算法库MLLIB已经显示出了其相对于Mahout的优势,在实际应用系统中得到了广泛的使用。

近年来,随着待分析数据规模的迅速扩张,分析模型参数也快速增长,对已有的大数据分析模式提出了挑战。例如在大规模话题模型LDA中,人们期望训练得到百万个以上的话题,因而在训练过程中可能需要对上百亿甚至千亿的模型参数进行更新,其规模远远超出了单个节点的处理能力。为了解决上述问题,研究人员提出了参数服务器(Parameter Server)的概念[26],如图5所示。在参数服务器系统中,大规模的模型参数被集中存储在一个分布式的服务器集群中,大规模的训练数据则分布在不同的工作节点(worker)上,这样每个工作节点只需要保存它计算时所依赖的少部分参数即可,从而有效解决了超大规模大数据分析模型的训练问题。目前参数服务器的实现主要有卡内基梅隆大学的Petuum[27]、PSLit[28]等。

在大数据分析的应用过程中,可视化通过交互式视觉表现的方式来帮助人们探索和理解复杂的数据。可视化与可视分析能够迅速和有效地简化与提炼数据流,帮助用户交互筛选大量的数据,有助于使用者更快更好地从复杂数据中得到新的发现,成为用户了解复杂数据、开展深入分析不可或缺的手段。大规模数据的可视化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。通常情况下,大规模数据可视化的技术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、任务并行化、管道并行化和数据并行化4种基本技术[29]。微软公司在其云计算平台Azure上开发了大规模机器学习可视化平台(Azure Machine Learning),将大数据分析任务形式为有向无环图并以数据流图的方式向用户展示,取得了比较好的效果。在国内,阿里巴巴旗下的大数据分析平台御膳房也采用了类似的方式,为业务人员提供的互动式大数据分析平台。

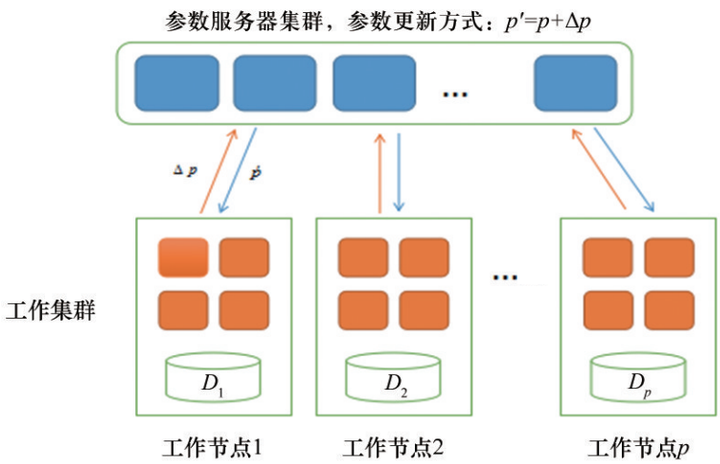

图5 参数服务器工作原理

3 大数据技术发展趋势

随着对大数据技术的不断发展和研究,其各个环节的技术发展呈现出新的发展趋势和挑战。2015年12月,中国计算机学会(CCF)大数据专家委员会发布了中国大数据技术与产业发展报告[30],并对中国大数据发展趋势进行了展望,主要包含以下6个方面。

3.1 可视化推动大数据平民化

近几年大数据概念迅速深入人心,大众直接看到的大数据更多是以可视化的方式体现。可视化是通过把复杂的数据转化为可以交互的图形,帮助用户更好地理解分析数据对象,发现、洞察其内在规律。可视化实际上已经极大拉近了大数据和普通民众的距离,即使对IT技术不了解的普通民众和非技术专业的常规决策者也能够更好地理解大数据及其分析的效果和价值,从而可以从国计、民生两方面都充分发挥大数据的价值。建议在大数据相关的研究、开发和应用中,保持相应的比例用于可视化和可视分析。

3.2 多学科融合与数据科学的兴起

大数据技术是多学科多技术领域的融合,数学和统计学、计算机类技术、管理类等都有涉及,大数据应用更是与多领域产生交叉。这种多学科之间的交叉融合,呼唤并催生了专门的基础性学科——数据学科。基础性学科的夯实,将让学科的交叉融合更趋完美。在大数据领域,许多相关学科从表面上看,研究的方向大不相同,但是从数据的视角看,其实是相通的。随着社会的数字化程度逐步加深,越来越多的学科在数据层面趋于一致,可以采用相似的思想进行统一研究。从事大数据研究的人不仅包括计算机领域的科学家,也包括数学等方面的科学家。希望业界对于大数据的边界采取一个更宽泛、更包容的姿态,包容所谓的“小数据”,甚至将领域的边界泛化到“数据科学”所对应的整个数据领域和数据产业。建议共同支持“数据科学”的基础研究,并努力将基础研究的成果导入技术研究和应用的范畴中。

3.3 大数据安全与隐私令人忧虑

大数据带来的安全与隐私问题主要包括以下3个方面:第一,大数据所受到的威胁也就是常说的安全问题,当大数据技术、系统和应用聚集了大量价值时,必然成为被攻击的目标;第二,大数据的过度滥用所带来的问题和副作用,比较典型的就是个人隐私泄露,还包括大数据分析能力带来的商业秘密泄露和国家机密泄露;第三,心智和意识上的安全问题。对大数据的威胁、大数据的副作用、对大数据的极端心智都会阻碍和破坏大数据的发展。建议在大数据相关的研究和开发中,保持一个基础的比例用于相对应的安全研究,而让安全方面产生实质性进步的驱动力可能是对于大数据的攻击和滥用的负面研究。

3.4 新热点融入大数据多样化处理模式

大数据的处理模式更加多样化,Hadoop不再成为构建大数据平台的必然选择。在应用模式上,大数据处理模式持续丰富,批量处理、流式计算、交互式计算等技术面向不同的需求场景,将持续丰富和发展;在实现技术上,内存计算将继续成为提高大数据处理性能的主要手段,相对传统的硬盘处理方式,在性能上有了显著提升。特别是开源项目Spark,目前已经被大规模应用于实际业务环境中,并发展成为大数据领域最大的开源社区。Spark拥有流计算、交互查询、机器学习、图计算等多种计算框架,支持Java、Scala、Python、R等语言接口,使得数据使用效率大大提高,吸引了众多开发者和应用厂商的关注。值得说明的是,Spark系统可以基于Hadoop平台构建,也可以不依赖Hadoop平台独立运行。

很多新的技术热点持续地融入大数据的多样化模式中,形成一个更加多样、平衡的发展路径,也满足大数据的多样化需求。建议将大数据研究和开发有意识地链接和融入大数据技术生态中,或者利用技术生态的成果,或者回馈技术生态。

3.5 深度分析推动大数据智能应用

在学术技术方面,深度分析会继续成为一个代表,推动整个大数据智能的应用。这里谈到的智能,尤其强调是涉及人的相关能力延伸,比如决策预测、精准推荐等。这些涉及人的思维、影响、理解的延展,都将成为大数据深度分析的关键应用方向。

相比于传统机器学习算法,深度学习提出了一种让计算机自动学习产生特征的方法,并将特征学习融入建立模型的过程中,从而减少了人为设计特征引发的不完备。深度学习借助深层次神经网络模型,能够更加智能地提取数据不同层次的特征,对数据进行更加准确、有效的表达。而且训练样本数量越大,深度学习算法相对传统机器学习算法就越有优势。

目前,深度学习已经在容易积累训练样本数据的领域,如图像分类、语音识别、问答系统等应用中获得了重大突破,并取得了成功的商业应用。预测随着越来越多的行业和领域逐步完善数据的采集和存储,深度学习的应用会更加广泛。由于大数据应用的复杂性,多种方法的融合将是一个持续的常态。建议保持对于智能技术发展的持续关注。在各自的分析领域(如在策划阶段、技术层面、实践环节等)尝试深度学习。

3.6 开源、测评、大赛催生良性人才与技术生态

大数据是应用驱动,技术发力,技术与应用一样至关重要。决定技术的是人才及其技术生产方式。开源系统将成为大数据领域的主流技术和系统选择。以Hadoop为代表的开源技术拉开了大数据技术的序幕,大数据应用的发展又促进了开源技术的进一步发展。开源技术的发展降低了数据处理的成本,引领了大数据生态系统的蓬勃发展,同时也给传统数据库厂商带来了挑战。新的替代性技术,都是新技术生态对于旧技术生态的侵蚀、拓展和进化。

对数据处理的能力、性能等进行测试、评估、标杆比对的第三方形态出现,并逐步成为热点。相对公正的技术评价有利于优秀技术占领市场,驱动优秀技术的研发生态。各类创业创新大赛纷纷举办,为人才的培养和选拔提供了新模式。大数据技术生态是一个复杂环境。2016年,“开源”会一如既往占据主流,而测评和大赛将形成突破性发展。建议不要闭门搞大数据技术和系统,要开门融入世界性的技术生态中。

4 结论

大数据技术的兴起正完成对各传统领域的颠覆。全球范围内,运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势。各国已相继制定实施大数据战略性文件,大力推动大数据发展和应用。从全球大数据发展的趋势来看,大数据产业推动社会生产要素的网络化共享、集约化整合、协作开发和高效利用,改变了传统的生产方式和经济运行机制,可显著提升经济运行水平和效率。中国是数据生产大国。目前,中国互联网、移动互联网用户规模居全球第一,拥有丰富的数据资源和应用市场优势。如果能在大数据管理和分析技术的研发与应用方面取得突破,可持续推动互联网创新企业和创新应用的高速成长。

参考文献(References)

[1] 李国杰. 大数据研究的科学价值[J]. 中国计算机学会通讯, 2012, 8(9): 8-15.
Li Guojie. Scientific value of big data research[J].China Computer Society Newsletter, 2012, 8(9): 8-15.

[2] 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[J]. 中国科学院院刊, 2012, 27(6): 647-657.
Li Guojie, Cheng Xueqi. Big data research: The major strategic areas of the development of the future science and economic society[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6):647-657.

[3] 王元卓, 靳小龙, 程学旗. 网络大数据:现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138.
Wang Yuanzhuo, Jin Xiaolong, Cheng Xueqi. Network big data: Present and future[J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138.

[4] 李国杰, 程学旗, 赵国栋, 等. 2014中国大数据技术与产业发展报告[M]. 北京: 机械工业出版社, 2013: 6-11.
Li Guojie, Cheng Xueqi, Zhao Guodong, et al. 2014 China big data technology and industry development report[M]. Beijing: China Mechine Press, 2013: 6-11.

[5] 周慧. 国家发改委: 资金支持大数据重大建设项目[EB/OL]. 2016-01-20 [2016-04-08]. http://news.hexun.com/2016-01-20/181906965.html.
Zhou Hui. National development and reform commission: funding support major construction projects of big data[EB/OL]. 2016-01-20 [2016-04-08]. http://news.hexun.com/2016-01-20/181906965.html.

[6] Apache H. What is apache hadoop?[EB/OL]. 2013-08-26[2016-04-13]. http://hadoop.apache.org.

[7] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[8] Zaharia M, Chowdhury M, Das T, et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing[C]//Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2012: 141-146.

[9] Lublinsky B, Smith K T, Yakubovich A. Professional hadoop solutions[M]. Birmingham: Wrox Press, 2013.

[10] Gartner Research Report. Magic quadrant for data quality tools [EB/OL]. [2016-04-12]. http://useready.com/wp-content/uploads/2013/07/Gartner-Data-Quality-2012.pdf

[11] Gonzalez J E, Low Y, Gu H, et al. Powergraph: Distributed graph-parallel computation on natural graphs[C]//Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation. Berkeley, CA: USENIX Association, 2012: 17-30.

[12] 吴甘沙. 大数据计算范式的分野与交融[J]. 程序员, 2013(9): 104-108.
Wu Gansha. The difference and blending of big data computing paradigm[J]. Programmer, 2013(9): 104-108.

[13] Engle C, Lupher A, Xin R, et al. Shark: Fast data analysis using coarse-grained distributed memory[C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management. New York: ACM, 2012.

[14] Neumeyer L, Robbins B, Nair A, et al. S4: Distributed stream computing platform[C]//Proceedings of the 10th International Conference on Data Mining Workshops. Washington, DC: IEEE, 2010: 170-177.

[15] Zaharia M, Das T, Li H, et al. Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters[C]//Proceedings of the 4th USENIX conference on Hot Topics in Cloud computing. Berkeley CA: USENIX Association, 2012: 10-16.

[16] Bu Y, Howe B, Balazinska M, et al. HaLoop: Efficient iterative data processing on large clusters[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 285-296.

[17] Zhang Y, Gao Q, Gao L, et al. iMapReduce: A distributed computing framework for iterative computation[C]//Proceedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum. Washington, DC: IEEE, 2011: 1112-1121.

[18] Ekanayake J, Li H, Zhang B, et al. Twister: A runtime for iterative mapreduce[C]//Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. New York: ACM, 2010: 810-818.

[19] Malewicz G, Austern M, Bik A, et al. Pregel: A system for large-scale graph processing[C]//Proceedings of the 2010 International Conference on Management of Data. New York: ACM, 2010: 135-146.

[20] Shao B, Wang H, Li Y, et al. Trinity: A distributed graph engine on a memory cloud[C]//Proceedings of the 2013 ACM SIGMOD International Conference on Management. New York: ACM, 2013: 1-12.

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

[22] Xin R, Gonalez J, Franklin M. GraphX: A resilient distributed graph system on spark[C]//Proceedings of the First International Workshop on Graph Data Management Experience and System. New York: ACM, 2013: 12-18.

[23] 程学旗, 王元卓. 大数据计算的技术体系与引擎系统[J]. 高科技与产业化, 2013, 9(5): 62-65.
Cheng Xueqi, Wang Yuanzhuo. Technology architecture and engine system for big data computing[J]. High-Technology & Industrialization, 2013, 9(5): 62-65.

[24] Xing E P, Qirong H Xie P T, et al. Strategies and principles of distributed machine learning on big data[J]. ArXiv Preprint ArXiv:1512.09295, 2015.

[25] Gropp W, Lusk E, Thakur R. Using MPI-2: Advanced features of the message-passing interface[M]. Cambridge MA: MIT Press, 1999.

[26] Smola A, Narayanamurthy S. An architecture for parallel topic models[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 703-710.

[27] Xing E P, Ho Q, Dai W, et al. Petuum: A new platform for distributed machine learning on big data[J]. IEEE Transactions on Big Data, 2015, 1(2): 49-67.

[28] Li M, Andersen D G, Park J W, et al. Scaling distributed machine learning with the parameter server[C]//11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14). Berkeley, CA: USENIX Association, 2014: 583-598.

[29] Bethel E W, Childs H, Hansen C. High performance visualization: Enabling extreme-scale scientific insight[M]. Boca Raton, FL: CRC Press, 2012.

[30] 潘柱廷, 程学旗, 袁晓如. CCF大专委2016年大数据发展趋势预测——解读和行动建议[J]. 大数据, 2016, 2(1): 2016012.
Pan Zhuting, Cheng Xueqi, Yuan Xiaoru. Developing trend forecasting of big data in 2016 from CCF TFBD: Interpretation and proposals[J]. Big Data Research, 2016, 2(1): 2016012.

Technological progress and trends of big data

CHENG Xueqi, JIN Xiaolong, YANG Jing, XU Jun

CAS Key Laboratory of Network Data Science & Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract With the rapid development of IT technology, worldwide businesses are experiencing informatization changes, and almost every industry is trying to exploit and utilize the value of big data. In order to take full advantage of the opportunity while effectively dealing with the challenges brought by big data, academia, industry and government actively promote the layout adjustment and establish the holistic strategic plan. This paper first introduces the background and dynamics of big data, and expounds the national policy environment and industry development status. Then it describes the technical progress on big data, with a comprehensive overview about the technology architecture and innovative features. Finally, the development trend is forecasted and some suggestions relating to big data visualization, multidisciplinary integration, security and privacy, depth analysis etc. are proposed.

Keywords big data technology; information technology; big data ecosystem

（编辑　韩丹岫）