文章编号:1674-7070(2014)05-0405-15

方巍[1,2,3]　郑玉[1,2]　徐江[1,2]

大数据:概念、技术及应用研究综述

摘要

随着物联网、云计算、移动互联网的迅猛发展,大数据(Big Data)吸引了越来越多的关注,正成为信息社会的重要财富,同时也给数据的处理与管理带来了巨大挑战.首先从大数据概念入手,阐述了大数据的来源、主要挑战、关键技术、大数据处理工具和应用实例等,并对比了大数据与云计算、物联网、移动互联网等技术之间关系,然后剖析了大数据核心技术、大数据企业解决方案,讨论了目前大数据应用实例,最后归纳总结了大数据发展趋势.旨在为了解大数据当前发展状况、关键技术以及科学地进行大数据分析与处理提供参考.

关键词

大数据;云计算;大数据处理;分布式系统;NoSQL

中图分类号 TP393.4

文献标志码 A

收稿日期 2014-09-01

资助项目 计算机软件新技术国家重点实验室(南京大学)开放课题(KFKT2014B21);2014年全国及江苏省大学生实践创新训练计划(201410300026);江苏高校优势学科建设工程资助项目

作者简介

方巍,男,博士,副教授、高级工程师,主要研究方向为智能信息处理、云计算和大数据分析.hsfangwei@sina.com

1 南京信息工程大学 江苏省网络监控中心,南京,210044

2 南京信息工程大学 计算机与软件学院,南京,210044

3 南京大学 计算机软件新技术国家重点实验室,南京,210046

0 引言

近来,当人们对“物联网”、“云计算”、“移动互联网”等热词还感觉模糊时,“大数据”(Big Data)又横空出世且其发展成燎原之势.2014年巴西世界杯与往届世界杯最大不同的是,其融入了诸多的科技元素如“云计算”、“大数据”等.IBM研究表明,在整个人类文明所获得的全部数据中,有90%是过去2年内产生的,到2020年,全世界所产生的数据规模将达到2009年的44倍.根据国际数据公司IDC监测,人类产生的数据量正在呈指数级增长,大约每2年翻一番,2020年全球数量将达到35 ZB.据统计,平均每一秒都有200万用户在使用Google搜索,Facebook注册用户超过10亿,每天生成300 TB以上的日志数据.同时,传感网、物联网、社交网络等技术迅猛发展,引发数据规模爆炸式增长,各种视频监控、监测、感应设备也源源不断地产生巨量流媒体数据,能源、交通、医疗卫生、金融、零售业等各行业也有大量数据不断产生,积累了TB级、PB级的大数据.上述情况表明,现在已进入大数据时代,大数据已经开始造福于人类,成为信息社会的宝贵财富.

大数据泛指大规模、超大规模的数据集,因可从中挖掘出有价值的信息而倍受关注,但传统方法无法进行有效分析和处理.《华尔街日报》将大数据时代、智能化生产和无线网络革命称为引领未来繁荣的3大技术变革.“世界经济论坛”报告指出大数据为新财富,价值堪比石油.因此,目前世界各国纷纷将开发利用大数据作为夺取新一轮竞争制高点的重要举措.

当前大数据分析者面临的主要问题有:数据日趋庞大,无论是入库和查询,都出现性能瓶颈;用户的应用和分析结果呈整合趋势,对实时性和响应时间要求越来越高;使用的模型越来越复杂,计算量指数级上升;传统技能和处理方法无法应对大数据挑战.

可喜的是,学术界、工业界甚至于政府机构都已经开始密切关注大数据问题,并对其产生浓厚的兴趣.就学术界而言,《Nature》和《Science》等国际顶级学术期刊相继出版专刊专门探讨大数据问题.2008年《Nature》出版了“Big Data”专刊[1],从互联网技术、网络经济学、超级计算、环境科学、生物医学等多个科技方面介绍大数据带来的挑战.《Science》也在2011年推出数据处理“Dealing with Data”专刊[2],讨论大数据所带来的挑战和大数据科学研究的重要性.IT产业

界如 IBM、Google、亚马逊、Facebook 等国际知名企业都是大数据的主要推动者,相继推出了各自的大数据产品.国内的大数据企业代表有百度、阿里巴巴、腾讯等.可以说,大数据兴起另一重要原因是经济利益驱动.大数据是一个具有国家战略意义的新兴产业,作为国家和社会的主要管理者,各国政府机构也是大数据技术的主要推动者.2012 年 3 月 29 日,美国政府宣布投资 2 亿美元启动"大数据研究和开发计划[3](Big Data Research and Development Initiative)",该计划旨在提高和改进人们从海量和复杂的数据中获取知识的能力,加快科学、工程领域的创新步伐,增强国家安全,把大数据看作"未来的新石油",并将对大数据的研究上升为国家意志,其 6 大机构合力研发核心技术,支持协同创新.英国、澳大利亚等国政府也开始大数据研究进程.我国对大数据研究也已提出指导性方针,《国家中长期科技发展规划纲要 2006—2020》、《"十二五"国家战略性新兴产业发展规划》中都提出支持海量数据存储、处理技术的研发和产业化.2013 年 2 月 1 日,科技部公布了国家重点基础研究发展计划(973 计划)2014 年度重要支持方向,其中,大数据计算的基础研究为重要支持方向之一.计世资讯认为,2011 年是中国大数据市场元年,一些大数据产品已经推出,2012—2016 年,将迎来大数据市场飞速发展,2016 年,整个国内大数据规模逼近百亿元.大数据研究是社会发展和技术进步的迫切需要.由上可见,大数据已引起了产业界、科技界和政府部门的高度关注.

大数据已在网络通信、医疗卫生、农业研究、金融市场、气象预报、交通管理、新闻报道等方面广泛应用.大数据背后隐藏着大量的经济与政治利益,尤其是通过数据整合、分析与挖掘,其所表现出的数据整合与控制力量已经远超以往.仅 2009 年,Google 公司通过大数据业务对美国经济的贡献就达 540 亿美元,而这只是大数据蕴含的巨大经济效益的冰山一角[4].可以说,现在大数据研究已经是社会发展和技术进步的迫切需要[5].

本文就目前大数据热点问题,就其概念、来源、存在问题、机遇和挑战、关键技术及应用实例等问题进行归纳和总结.首先分析大数据时代背景、研究现状及意义,阐述了大数据概念和来源、数据处理面临的问题及挑战,对大数据与云计算、物联网之间的关系也作了对比分析,着重分析了大数据技术框架、常用处理工具和大数据企业解决方案,还讨论了大数据行业应用实例,最后是全文总结并对大数据研究进行展望.

1 大数据概念

最早提出"大数据"时代到来的是全球知名咨询公司麦肯锡,该公司在《大数据:创新、竞争和生产力的下一个前沿领域》报告中称:"数据,已经渗透到当今每一个行业和业务职能领域,成为重要的生产因素.人们对于海量数据的挖掘和运用,预示着新一波生产率增长和消费者盈余浪潮的到来."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.同时强调,并不是说一定要超过特定 TB 级的数据集才能算是大数据[6].大数据是云计算、物联网之后 IT 行业又一大颠覆性的技术革命.

1.1 大数据定义

那么,大数据的定义是什么呢? 一般而言,大家比较认可关于大数据从早期的 3V、4V 说法到现在的 5V(新增 Value).大数据的 5 个"V",业界将其归纳为 Volume, Velocity, Variety, Veracity, Value,如图 1 所示.实际上也就是大数据包含的 5 个特征,包含 5 个层面意义:第一,数据体量(Volume) 巨大.指收集和分析的数据量非常大,从 TB 级别,跃升到 PB 级别,但在实际应用中,很多企业用户把多个数据集放在一起,已经形成了 PB 级的数据量.第二,处理速度(Velocity)快,需要对数据进行近实时的分析.以视频为例,连续不间断监控过程中,可能有用的数据仅仅有一两秒.这一点和传统的数据挖掘技术有着本质的不同.第三,数据类别(Variety)大,大数据来自多种数据源,数据种类和格式日渐丰富,包含结构化、半结构化和非结构化等多种数据形式,如网络日志、视频、图片、地理位置信息等.第四,数据真实性(Veracity).大数据中的内容是与真实世界中的发生息息相关的,研究大数据就是从庞大的网络数据中提取出能够解释和预测现实事件的过程.第五,价值密度低,商业价值(Value)高.通过分析数据可以得出如何抓住机遇及收获价值.

维基百科(http://en.wikipedia.org/wiki/Big_data)对大数据的定义则简单明了:大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集.也就是说大数据是一个体量特别大,数据类别特别大的数据集,并且这样的数据集无法用传统数据库工具对其内容进行抓取、管理

和处理.

图 1　大数据 5V 特征

Fig. 1　5Vs feature of Big Data

1.2　大数据的来源

大数据集通常是 PB 或 EB 的大小.这些数据集有各种各样的来源:传感器,气候信息,公开的信息,如杂志、报纸、文章,还包括购买交易记录,网络日志,病历,军事监控,视频和图像档案,及大型电子商务等.当前,根据来源不同,大数据大致分为如下几种类型[5,7]:

1) 来自人类活动:人们通过社会网络、互联网、健康、金融、经济、交通等活动过程所产生的各类数据,包括微博、病人医疗记录、文字、图形、视频等信息.

2) 来自计算机:各类计算机信息系统产生的数据,以文件、数据库、多媒体等形式存在,也包括审计、日志等自动生成的信息.

3) 来自物理世界:各类数字设备、科学实验与观察所采集的数据.如摄像头所不断产生的数字信号,医疗物联网不断产生的人的各项特征值,气象业务系统采集设备所收集的海量数据等.

1.3　大数据处理的挑战

大数据时代数据存在多源异构、分布广泛、动态增长、先有数据后有模式等诸多特点.正是这些与传统数据处理不同的特点,使得大数据时代数据管理面临新的挑战.目前大数据处理和分析工具却相当落后,问题很严重:在大数据背景下,传统的数据分析软件都是失效的.通过目前主流软件工具,无法在合理时间内撷取数据、管理数据、处理数据,并整理成为帮助企业经营或主管部门决策提供支持的数据.

在应对处理大数据时代的各种技术挑战过程中,以下几个方面问题需高度关注[8-11].

1.3.1　数据的异构性和不完备性

大数据的广泛存在和来源的多样性使数据越来越多分散在不同的数据管理系统中,目前采集到的85%以上是非结构化和半结构化的数据,不能用已有的简单数据结构来描述它们.而传统关系数据库无法高效处理复杂的数据结构表示的数据,但处理同质的数据则非常有效.因此,如何将数据组织成合理的结构,进行数据的集成是大数据处理的一个重要挑战问题.

数据的不完备性是指在大数据条件下所获取的数据常常包含一些不完整的信息和错误的数据[11].在进行大数据分析处理之前必须对数据的不完备性进行有效处理才能分析出有价值的信息,通常在数据采集与预处理阶段完成.例如,某医疗过程数据一致且准确,但遗失某些患者既往病史,从而存不完备性,可能导致不正确的诊断甚至严重医疗事故.由于大数据的 5 类特征存在,对不完备性处理方法是一项挑战,为表述信息,文献[12]提出了另一种关系数据库的扩展模型,给出了封闭世界假设和开放世界假设的概念.文献[13]提出“open null”的概念,提出了在封闭式假设下数据库缺失属性值的表示方法.另外,文献[14-16]在概率数据管理方面的一些研究成果会为未来的不确定、不完备的数据管理提供新的方法.工业界在多种数据清洗和质量控制方面开发出多种工具,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 等.

可以说,大数据异构性和不完备性处理即数据集成问题将是面临的首要挑战问题.

1.3.2　数据处理的时效性

传统的数据分析主要针对结构化数据,利用数据库技术来存储结构化数据,并在此基础上构建数据仓库进行联机分析处理(Online Analytical Processing,LAP).现有方法对处理相对较少的结构化数据时极为高效,但对于大数据而言,半结构化和非结构化数据量的迅猛增长,给传统数据分析处理带来巨大冲击和挑战.

随着时间的流逝,数据中所蕴含的知识价值也随之衰减,因此,大数据处理的速度非常重要.一般来讲,数据规模越大,分析处理时间就会越长,而在许多情况下,用户要求立即得到数据的分析结果.大数据则要求为复杂结构的数据建立合适的索引结构,这就要求索引结构的设计简单、高效,能够在数据模式发生变化时很快进行调整适应[11,17].在数据模式变更的假设前提下设计新的索引方案将是大数据处理的主要挑战之一[8].大数据存储系统的要求是:高可用、低成本、高性能、低开销.

1.3.3 数据的安全与隐私保护

隐私问题由来已久.互联网技术的发展使数据的传输、共享更加便利,而数据隐私问题则越来越严重,如最近爆发的"棱镜门"事件."棱镜"项目是一项由美国国家安全局(NSA)自2007年起开始实施的绝密电子监听计划,年耗资近2 000亿美元,用于监听全美电话通话记录,据称还可以使情报人员通过"后门"进入9家主要科技公司的服务器,包括微软、雅虎、Google、Facebook、PalTalk、美国在线、Skype、YouTube、苹果.该事件加剧了人们对大数据安全与隐私的担忧.人们在互联网上的一言一行都掌握在互联网商家手中,例如淘宝知道用户的购物习惯、腾讯知道用户的好友联络情况、百度知道用户的检索习惯等.大数据的隐私保护与安全是大数据分析和处理的一个重要方面.

大数据的隐私保护既是技术问题也是社会学问题,需要学术界、商业界和政府部门共同参与[10].目前中国并没有专门的法律法规来界定用户隐私,处理相关问题时多采用其他相关法规条例来解释.但随着民众隐私意识的日益增强,合法合规地获取数据、分析数据和应用数据,是进行大数据分析时必须遵循的原则.

大数据时代的安全与传统安全相比,变得更加复杂,面临更多挑战.如何在大数据环境下确保信息共享的安全性和如何为用户提供更为精细的数据共享安全控制策略等问题值得深入研究.

1.3.4 大数据能耗问题

随着大数据规模的不断扩张,而能源价格持续上涨,同时数据中心存储规模不断扩大,高能耗已逐渐成为制约大数据快速发展的一个主要瓶颈[7].要达到低成本、低能耗、高可靠性目标,通常要用到冗余配置、分布式和云计算技术,在存储时要按照一定规则对数据进行分类,通过过滤和去重,减少存储量,同时进行索引便于查询操作.大数据管理系统中,能耗主要由2大部分组成:硬件能耗和软件能耗,二者之中又以硬件能耗为主.据《纽约时报》2012年调查显示[18],Google数据中心年电功率约为3亿W,而Facebook则达6 000万W左右,最令人惊讶的是这些巨大能耗中,实际只有6%至12%的能量是真正用于响应用户查询请求,绝大部分电能是用来确保系统服务器处于正常待机状态,以应对突如其来的用户查询的网络流量高峰.从已有的一些研究成果来看,可以考虑以下2个方面来改善大数据能耗问题:采用新型低功耗硬件,建立计算核心与二级缓存的直通通道,从应用、编译器、体系结构等多方面协同优化;引入可再生的新能源.

1.3.5 大数据管理易用性问题

大数据时代,数据的数量和复杂度的提高对数据的处理、分析、理解和呈现带来极大挑战.从开始的数据集成到数据分析,到最后的数据解释过程,易用性贯穿于整个大数据处理的流程.易用性的挑战突出体现在2个方面[8]:首先大数据的数据量大,分析更复杂,得到的结果更加多样化,其复杂程度已远超传统的关系数据库;其次大数据已广泛渗透到人们生活的方方面面,复杂的分析过程和难以理解的分析结果制约了各行各业从大数据中获取知识的能力,大数据分析结果的可视化呈现将是大数据管理易用性的又一挑战问题.

1.4 传统数据库和大数据的比较

现有数据处理技术大多采用数据库管理技术,从数据库到大数据,看似一个简单的技术升级,但仔细考察不难发现两者存在一些本质上区别.

传统数据库时代的数据管理可以看作"池塘捕鱼",而大数据时代数据管理类似"大海捕鱼","鱼"表示待处理的数据."捕鱼"环境条件的变化导致"捕鱼"方式的根本性差异[8].具体差异归纳如表1所示.

表1 传统数据库和大数据比较

Table 1 Comparison between traditional database and Big Data

项目	传统数据库	大数据
数据规模	以MB为基本单位	常以GB,甚至是TB、PB为基本处理单位
数据类型	数据种类单一,往往仅有一种或少数几种,且以结构化数据为主	种类繁多,数以千计,包括结构化、半结构化和非结构化数据
产生模式	先有模式,才会产生数据	难以预先确定模式,模式只有在数据出现之后才能确定,且模式随着数据量的增长不断演化
处理对象	数据仅作为处理对象	数据作为一种资源来辅助解决其他诸多领域问题
处理工具	一种或少数几种就可以应对	不可能存在一种工具处理大数据,需要多种不同处理工具应对

1.5 大数据与云计算、物联网之间的关系

大数据产生有其必然性,主要归结于互联网、移动设备、物联网和云计算等快速崛起,全球数据量大

幅提升.可以说,移动互联网、物联网以及云计算等热点崛起在很大程度上是大数据产生的原因.要了解大数据的概念,就必须了解大数据与云计算、物联网、移动互联网之间的关系.

《互联网进化论》一书中提出"互联网的未来功能和结构将与人类大脑高度相似,也将具备互联网虚拟感觉、虚拟运动、虚拟中枢、虚拟记忆神经系统",并绘制了一幅互联网虚拟大脑结构图,形象生动地描绘了大数据、物联网、云计算等之间的关系,如图 2 所示[19].

从图 2 可以看出:物联网对应了互联网的感觉和运动神经系统,是数据的采集端;云计算是互联网的核心硬件层和核心软件层的集合,也是互联网中枢神经系统萌芽,是数据的处理中心;大数据代表了互联网的信息层(数据海洋),是互联网智慧和意识产生的基础.物联网,传统互联网和移动互联网在源源不断地向互联网大数据层汇聚数据和接受数据.其中,大数据与云计算(Cloud Computing)的异同总结如表 2 所示.

因此,不难发现大数据与云计算两者是相辅相成的.云计算和大数据实际上是工具与用途的关系,即云计算为大数据提供了有力的工具和途径,大数据为云计算提供了很有价值的用武之地.

而物联网(The Internet of Things)作为新一代信息技术的重要组成部分,是互联网的应用拓展,广泛应用于智能交通、环境保护、政府工作、公共安全、平安家居、智能消防、气象灾害预报、工业监测、个人健康、照明管控、情报收集等诸多领域.物联网、移动互联网和传统互联网,每天都产生海量数据,为大数据提供数据来源,而大数据则通过云计算的形式,将这些数据分析处理,提取有用的信息,即大数据分析.

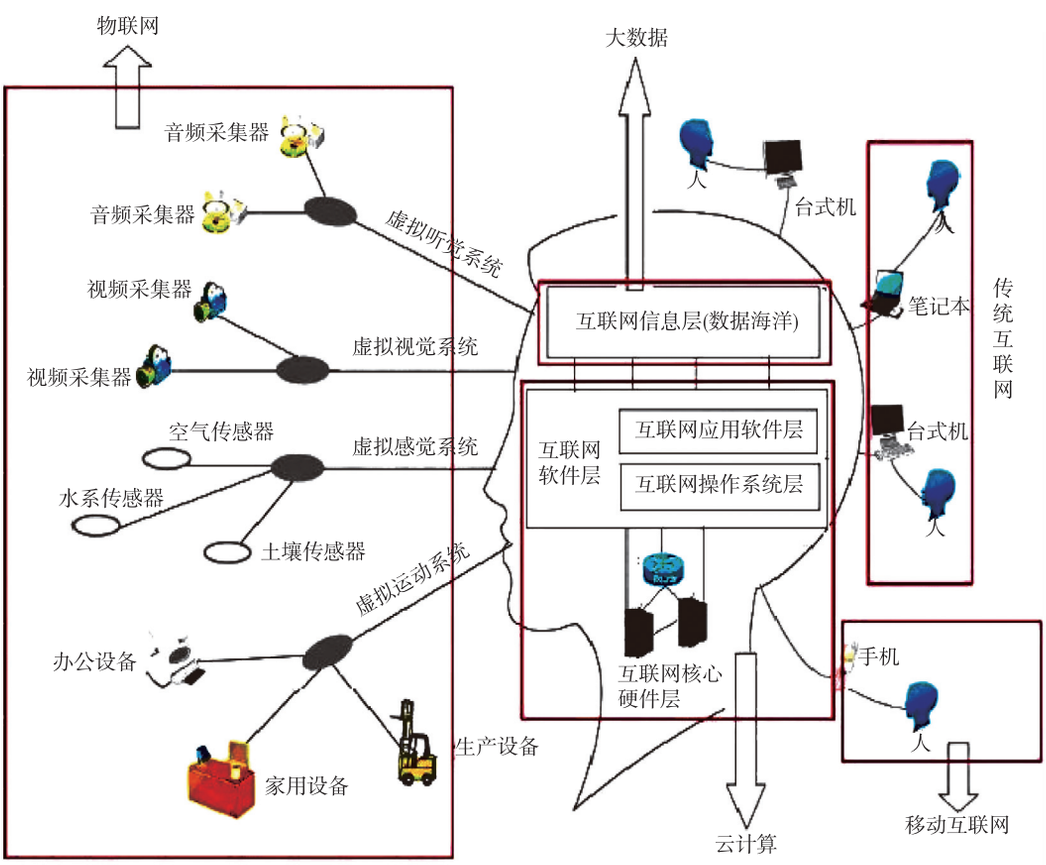

图 2 大数据、云计算、物联网和移动互联网之间关系

Fig. 2 Relationship among Big Data, cloud computing and mobile internet

表 2　大数据与云计算关系

Table 2　Relationship between Big Data and cloud computing

		大数据	云计算
总体关系		大数据着眼于“数据”,关注实际业务,云计算着眼于“计算”,关注 IT 解决方案,提供 IT 基础架构,看重数据处理能力. 云计算为大数据提供有力的工具和途径,大数据为云计算提供用武之地.	
相同点		1)目的相同:都是为数据存储和处理服务,需占用大量的存储和计算资源. 2)技术相似:大数据根植于云计算,云计算关键技术中的海量数据存储技术、海量数据管理技术、MapReduce 编程模型,都是大数据技术的基础.	
不同点	背景	不能胜任社交网络和物联网产生的大量异构但有价值数据	基于互联网服务日益丰富和频繁
	目标	充分挖掘海量数据中的信息	扩展和管理计算机软硬件资源和能力
	对象	各种数据	IT 资源、能力和应用
	推动力量	从事数据存储与处理的软件厂商和拥有大量数据的企业	存储及计算设备的生产厂商和拥有计算及存储资源的企业
	价值	发现数据中的价值	节省 IT 资源部署成本

2　大数据技术

大数据处理技术正在改变当前计算机的运行模式,正在改变着这个世界.它能处理几乎各种类型的海量数据,无论是微博、文章、电子邮件、文档、音频、视频,还是其他形态的数据.它实时、高效、可视化呈现结果.它依托云计算将计算任务分布在大量计算机构成的廉价的资源池上,使用户能够按需获取计算资源、存储资源、网络资源和信息服务.云计算技术的应用使得大数据处理和利用成为可能.

大数据作为信息金矿,对其采集、传输、处理和应用的相关技术就是大数据处理技术,是一系列使用非传统的工具来对大量的结构化、半结构化和非结构化数据进行处理,从而获得分析和预测结果的一系列数据处理技术,或简称大数据技术.

2.1　大数据技术框架

根据大数据处理的生命周期,大数据的技术体系涉及大数据的采集与预处理、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、大数据可视化分析及大数据隐私与安全等几个方面[7-8,10].图 3 是大数据技术主要架构示意.

2.1.1　大数据采集与预处理

大数据的一个重要特点就是数据源多样化,包括数据库、文本、图片、视频、网页等各类结构化、非结构化及半结构化数据.因此,大数据处理的第一步

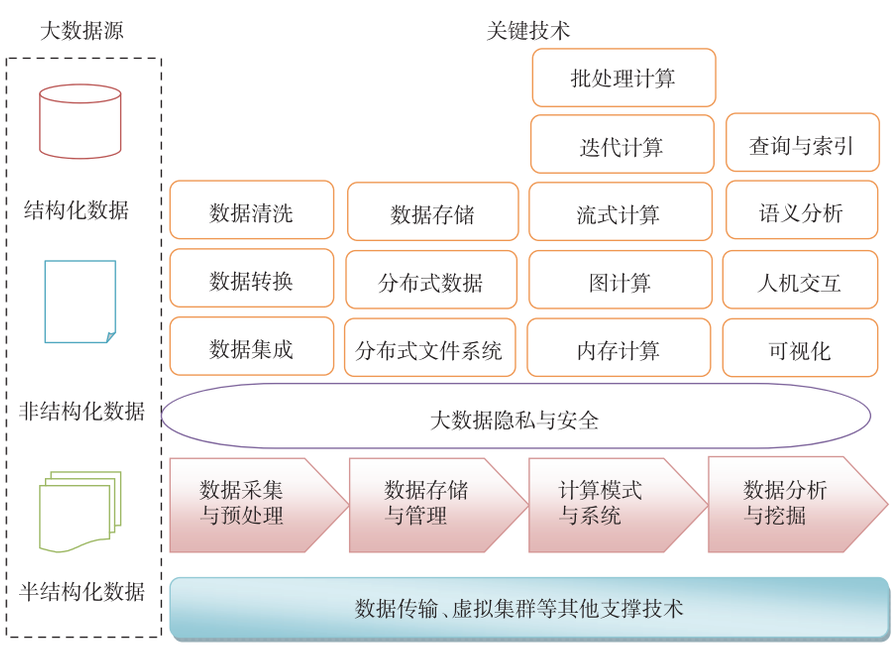

图 3　大数据技术架构

Fig. 3　Technology architecture of Big Data

是从数据源采集数据并进行预处理和集成操作,为后继流程提供统一的高质量的数据集.

现有数据抽取与集成方式可分为以下 4 种类型[21]:基于物化或 ETL 引擎方法、基于联邦数据库引擎或中间件方法、基于数据流引擎方法和基于搜索引擎方法.

常用 ETL 工具负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成,最后加载到数据仓库或数据集市中,成为联机分析处理(OLAP)、数据挖掘的基础.

由于大数据的来源不一,异构数据源的集成过程中需要对数据进行清洗,以消除相似、重复或不一致数据.文献[21-22]中数据清洗和集成技术针对大数据特点,提出非结构化或半结构化数据的清洗以及超大规模数据的集成方案.

2.1.2 *大数据存储与管理*

数据存储与大数据应用密切相关.大数据给存储系统带来 3 个方面挑战:1)存储规模大,通常达到 PB 甚至 EB 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[10].

大数据存储与管理,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,并且对数据处理的实时性、有效性提出更高要求,传统常规技术手段根本无法应付.某些实时性要求较高的应用,如状态监控,更适合采用流处理模式,直接在清洗和集成后的数据源上进行分析.而大多数其他应用需要存储,以支持后续更深度数据分析流程.根据为上层应用访问接口和功能侧重不同,存储和管理软件主要包括文件系统和数据库.大数据环境下,目前最适用的技术是分布式文件系统、分布式数据库以及访问接口和查询语言[10].

目前,一批新技术提出来应对大数据存储与管理的挑战,这方面代表性的研究包括分布式缓存(包括 CARP、mem-cached)、基于 MPP 的分布式数据库、分布式文件系统(GFS[23]、HDFS[24]),各种 NoSQL 分布式存储方案(http://nosqldatabase.org/)(包括 Mongodb、CouchDB、HBase、Redis、Neo4j 等).各大数据库厂商如 Oracle、IBM、Greenplum 都已经推出支持分布式索引和查询产品.

2.1.3 *大数据计算模式与系统*

大数据计算模式指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或模型,它的出现有力推动了大数据技术和应用的发展.例如,MapReduce 是一个并行计算编程模型[25],Bekerley 大学著名的 Spark 系统中"分布内存抽象 RDD(RDD, a distributed memory abstraction[26])"、CMU 著名的图计算系统 GraphLab 的"图并行抽象"(Graph Parallel Abstraction[27])等.

大数据处理的主要数据特征和计算特征维度有:数据结构特征、数据获取方式、数据处理类型、实时性或响应性能、迭代计算、数据关联性和并行计算体系结构特征.根据大数据处理多样性需求和上述特征维度,目前已有多种典型、重要的大数据计算模式和相应大数据计算系统和工具[28].典型大数据计算模式与系统如表 3 所示[10].

表 3 典型大数据计算模式与系统

Table 3 Typical Big Data computing models and systems

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 Streaming 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

其中,大数据查询分析计算模式是为应对数据体量极大时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.大数据查询分析计算的典型系统包括 Hadoop(http://hadoop.apache.org)下的 HBase 和 Hive,Facebook 开发的 Cassandra,Google 公司的交互式数据分析系统 Dremel[29],Cloudera 公司的实时查询引擎 Impala.

最适合完成大数据批处理计算模式是 Google 公司的 MapReduce.一个完整的 MapReduce 过程如图 4 所示[25].

流式计算是一种实时性的计算模式,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,避免数据堆积和丢失,尽可能快地对最新数据做出分析并给出结果是流式计算的目标,其执行流程如图 5 所示[30].采用流式计算的大数据应用场景有网页点击数实时统计、传感器网络、电力、金融交易、道路监控等,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

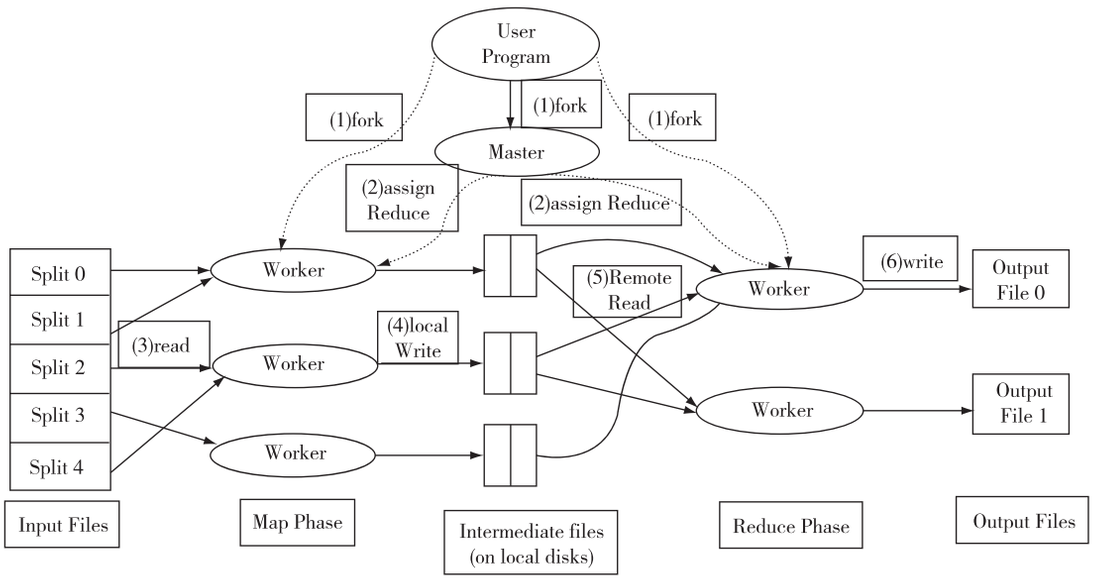

图 4　MapReduce 执行流程

Fig. 4　Execution flow chart of MapReduce

需要具备高实时性的流式计算能力. Facebook 的 Scribe 和 Apache 的 Flume 都提供相应机制构建日志数据处理流图. 比较有代表性的流式计算开源系统如 Twitter 的 Storm(https://github.com/nathanmarz/storm)、Yahoo 的 S4[31]、Linkedin 的 Kafka[32] 以及 Berkeyly AMPLab 的 Spark Steaming[33].

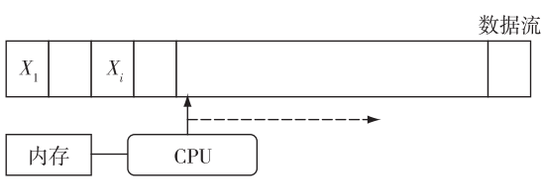

图 5　基本流式计算模型

Fig. 5　Basic flow computing model

原始的 MapReduce 编程模型并不能很好支持迭代计算,计算代价很大,而迭代计算是图计算、数据挖掘等领域常见运算模式. 为了克服 MapReduce 难以支持迭代计算的缺陷,HaLoop 将迭代控制放到 MapReduce 作业执行的框架内部,通过各个 task tracker 对数据进行缓存和创建索引,以减少迭代间的数据传输的 I/O 磁盘开销[34]. Twister 系统[35] 则将 Hadoop 全部数据存放在内存中,引入可缓存的 Map 和 Reduce 对象. iMapReduce 在此基础上保持 Map 和 Reduce 任务的持久性,规避启动和调度开销[36]. iHadoop 实现了 MapReduce 的异步迭代,但在 task 之间复用上无多大改进[37]. 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[26],其将中间结果放在内存中实现快速的迭代计算,但适用异步细粒度更新状态应用.

图一般用来表示真实社会广泛存在事物之间联系的一种有效手段,在社交网络、Web 链接关系、各种社会关系等方面存在大量图数据,这些图数据规模通常达到数十亿个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,使得如何对它们进行高效处理成为一个巨大的挑战. 目前出现的分布式图计算系统,主要有 Google 公司的 Pregel[38], Hadoop 的开发商 Yahoo 对 Pregel 的开源实现 Giraph,微软基于内存的分布式图数据库系统 Trinity[39], Berkeley AMPLab 的 GraphX[40], Infinite Graph[41] 以及 CMU 的 GraphLab 以及目前性能最快的图数据系统 PowerGraph[27] 等.

内存计算是指 CPU 直接从内存,而不是硬盘上读取数据,并进行计算、分析,是对传统数据处理方式的一种加速. 内存计算非常适合处理海量数据,以及需要实时获得结果的数据. 随着信息技术的高速发展,计算机硬件价格持续下降,其中内存价格非常低廉使服务器可配置的高内存容量成为可能,用内存计算完成实时的大数据处理已成为大数据计算的一个重要发展趋势. 分布内存计算的典型开源系统是 Spark,SAP 公司的 Hana 则是一个全内存式的基于开放式架构设计的内存计算系统,也是一个高性能大数据管理平台,其他如 Oracle 的 TimesTen,IBM 的 SolidDB.

2.1.4 大数据分析与挖掘

由于大数据环境下数据呈现多样化、动态异构,而且比小样本数据更有价值等特点,需要通过大数据分析与挖掘技术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.

针对大数据环境非结构化或半结构化的数据挖掘问题,文献[42]提出针对图片文件的挖掘技术,文献[43]提出一种大规模文本文件的检索与挖掘技术.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 公司对 R 和 Hadoop 进行集成[44].R 是开源的统计分析软件,通过 R 和 Hadoop 深度集成,可进行数据挖掘和并行处理,使 Hadoop 获得了强大的深度分析能力.另有研究者实现了 Weka(一种类似 R 的开源数据挖掘工具软件)和 MapReduce 的集成,可实现大数据的分析与挖掘.

2.1.5 大数据可视化分析

从上分析可知,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、分析和理解的巨大挑战.数据分析是大数据处理的核心,但是用户往往更关心结果的展示.如果分析的结果正确但是没有采用适当的解释方法,则所得到的结果很可能让用户难以理解,极端情况下甚至会误导用户.由于大数据分析结果具有海量、关联关系极其复杂等特点,采用传统的解释方法基本不可行.目前常用的方法是可视化技术和人机交互技术.

可视化技术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,有助于用户更快更好地从复杂数据中得到新的发现.用形象的图形方式向用户展示结果,已作为最佳结果展示方式之一率先被科学与工程计算领域采用.常见的可视化技术有原位分析(In Situ Analysis)、标签云(Tag Cloud)、历史流(history flow)、空间信息流(Spatial information flow)、不确定性分析等.可以根据具体的应用需要选择合适的可视化技术.现有研究如文献[45-46],通过数据投影、维度降解和电视墙等方法来解决大数据显示问题.

另外,以人为中心的人机交互技术也是解决大数据分析结果的一种重要技术,让用户能够在一定程度上了解和参与具体的分析过程.这个既可以采用人机交互技术,利用交互式的数据分析过程来引导用户逐步进行分析,使得用户在得到结果的同时更好理解分析结果的由来,也可以采用数据起源技术,通过该技术可以帮助追溯整个数据分析的过程,有助于用户理解结果.

2.1.6 大数据隐私与安全

当前大数据的发展仍然面临着许多问题,安全和隐私问题是人们公认的关键问题之一[7].其中,隐私问题由来已久,计算机的出现使得越来越多的数据以数字化的形式存储在电脑中,互联网的发展则使数据更加容易产生和传播,数据隐私问题越来越严重.

大数据在存储、处理、传输等过程中面临安全风险,具有数据安全和隐私保护需求.而实现大数据安全与隐私保护,较其他安全问题(如云安全中数据安全等)更为棘手.呈现出的安全隐私问题主要有[10,47]:

1)大数据时代的安全与传统安全相比,变得更加复杂;

2)使用过程中的安全问题;

3)对大数据分析较高的企业和团体,面临更多的安全挑战;

4)基于位置的隐私数据暴露严重;

5)缺乏相关的法律法规保证;

6)大数据的共享问题;

7)数据动态性;

8)多元数据的融合挑战.

目前针对上述问题,主要研究解决方法有[7,10]:文件访问控制技术、基础设备加密、匿名化保护技术、加密保护技术、数据水印技术、数据溯源技术、基于数据失真的技术、基于可逆的置换算法.

2.2 大数据处理工具

Hadoop 是目前最为流行的大数据处理平台.Hadoop 由 Apache 公司为实现 Google 的 MapReduce 编程模型的一个云计算开源平台,Hadoop 是可伸缩、高效的,能够处理 PB 级数据.Hadoop 平台包括最底部的文件系统(HDFS)、数据库(HBase、Cassandra)、数据处理(MapReduce)、数据仓库(Hive)、大数据分析语言接口(Pig)等功能模块在内的完整生态系统(Ecosystem).某种程度上可以说 Hadoop 已经成为大数据处理工具事实上的标准[8].

现有的大数据处理工具大多是对开源的 Hadoop 平台进行改进并将其应用于各种场景的.Hadoop 完整生态系统中各子系统都有相对应大数据处理的改进产品.目前大数据分析处理工具中常用的有:Hadoop、HPCC、Storm、Apache Drill、RapidMiner、Pentaho BI 等.下面针对前述大数据处理技术生命周期的各个阶段,

归纳总结了现今各阶段一些主流的大数据处理平台和工具,如表 4 所示,这些平台和工具或已经投入商业使用,或是开源软件.在已经投入商业使用的产品中,绝大部分也是在开源 Hadoop 平台基础上进行功能扩展,或者提供与 Hadoop 的数据接口[8].

表 4　常用大数据处理工具

Table 4　Commonly used Big Data processing tools

种类		工具示例
平台	Local	Hadoop、MapR、Cloudera、Hortonworks、BigInsights、HPCC
	Cloud	AWS、Google Coumpute Engine、Azure
数据库	SQL	MySql(Oracle)、MariaDB、PostgreSQL、TokuDBGreenplum、Aster Data、Vertica
	NoSQL	HBase、Cassandra、MongoDB、Redis
	NewSQL	Spanner、Megastore、F1
数据仓库		Hive、HadoopDB、Hadapt
数据收集		scraperWIKI、needlebase、bazhuayu
数据清洗		DataWrangler、Google Refine、OpenRefine
数据处理	批处理	MapReduce、Dyrad
	流计算	Sorm、S4、Kafka
	内存计算	Drill、Dremel、Spark
查询语言		HiveQL、Pig Latin、DyradLINQ、MRQL、SCOPE
统计与机器学习		Mahout、Weka、R、RapidMiner
数据分析		Jaspersoft、Pentaho、Splunk、Loggly、Talend
可视化分析		Google Chart API、Flot、D3、Processing、FUSION TABLES、Gephi、SPSS、SAS、R、Modest Maps、OpenLayers

2.3　大数据技术发展趋势

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,大数据技术各个环节的技术发展呈现出新的发展趋势和挑战,具体归纳如表 5 所示[7-10].

表 5　大数据技术发展趋势

Table 5　Development trend of Big Data technology

主要技术	发展趋势
采集与预处理	1) 数据源的选择与高质量原始数据的采集方法; 2) 多源数据的实体识别和解析方法; 3) 数据清洗和自动修复方法; 4) 高质量数据的整合方法; 5) 数据演化的溯源管理.
存储与管理	1) 大数据索引和查询技术; 2) 实时/流式大数据存储与处理.
计算模式与系统	1) Hadoop 改进后与其他计算模式和平台共存; 2) 混合计算模式成为大数据处理有效手段.
数据分析与挖掘	1) 更复杂和大规模分析与挖掘; 2) 大数据实时分析与挖掘; 3) 大数据分析与挖掘的基准测试.
可视化分析	1)原位分析; 2)人机交互; 3)协同与众包可视分析; 4)可扩展性与多级层次问题; 5)不确定性分析和敏感分析; 6)可视化与自动数据计算挖掘结合; 7)面向领域和大众的可视化工具库.
数据隐私与安全	1) NoSQL 有待进一步完善; 2) APT 攻击研究; 3) 社交网络的隐私保护; 4) 数字水印技术; 5) 风险自适应访问控制; 6) 数据采集、存储、分析 3 个过程“三权分立”.
其他	1) 大数据高效传输架构和协议; 2) 大数据虚拟机集群优化研究.

2.4　大数据企业解决方案

为充分发挥大数据的业务价值,企业需要一个可扩展、灵活而可管理的数据基础架构.面对大数据的机遇与挑战,国内外各大公司都提出了相应大数据解决方案.表 6 归纳总结了国内外一些主要企业的大数据应用解决方案(http://wenku.it168.com/redian/data/;http://www.itongji.cn/article/062622c2013.html).

3　大数据应用实例

说到大数据应用实例,当属 Google 有一个名为“Google 流感趋势”的工具(http://www.google.org/flutrends/).Google 认为,某些搜索词有助于了解流感疫情.Google 流感趋势会根据汇总的 Google 搜索数据,近乎实时地对全球当前的流感疫情进行估测.它对于健康服务产业和流行病专家来说是非常有用的,因为它的时效性极强,能够很好地帮助到疾病暴发的跟踪和处理.

表 6　大数据企业解决方案

Table 6　Industrial solutions for Big Data

主要企业	技术产品	特点
Google	BigQuery (包含 MapReduce、BigTable、GFS、Chubby 等技术)	BigQuery 是真正为大数据而生的企业级云计算产品,可用于对 TB 级别的大数据进行实时的分析处理
IBM	InfoSphere 大数据分析平台(包括 BigInsights 和 Streams 产品)	提供了全面的大数据解决方案,可与 DB2、Netezza 等集成,这使大数据平台更适合企业级应用
Oracle	硬件:大数据机 软件:Oracle Linux、Oracle JDK、Cloudera Hadoop Distribution、Cloudera Manager、NoSQL DataBase 等	提供大数据软硬件一体化集成解决方案
Microsoft	Windows Azure 平台上提供基于云端的 Hadoop 服务 Windows Server 集成 Hadoop 版本 Microsoft Business Intelligence (BI) 数据分析平台	帮助企业在原有微软软件产品基础上快速部署其大数据解决方案
EMC	Greenplum 统一分析平台(UAP)结合 Greenplum DB 和 Greenplum Hadoop	为企业构建高效处理结构化、半结构化、非结构化数据的大数据分析平台.
Intel	Intel Hadoop Manager	对 Intel 平台上的 Hadoop 进行了优化
亚马逊	Amazon EC2、Amazon EC2、Karmasphere Analyst	可应对各种数据密集型任务
阿里云	ODPS (Open Data Processing Service)	通过 ODPS 在线服务,不用花大钱建数据中心,就能分析海量数据.

最近《羊城晚报》发布的 2014 中国 10 大堵城排行榜(http://legal.gmw.cn/2014-08/25/content),据高德地图介绍,是将 3 亿高德地图导航用户作为数据蓝本,基于大数据计算出来的.

大数据将给各行各业带来变革性机会,但真正的大数据应用仍处于发展初级阶段.下面就目前大数据在电子政务、网络通信行业、医疗行业、能源、气象等行业的应用进行简单介绍[10,48].

3.1　大数据在电子政务中的应用

大数据的发展,将极大改变政府现有管理模式和服务模式.具体而言,就是依托大数据的发展,节约政府投入、及时有效进行社会监管和治理,提升公共服务能力.以大数据应用支撑政务活动为例,美国积极运用大数据推动政府管理方式变革和管理能力提升,越来越多的政府部门依托数据及数据分析进行决策,将之用于公共政策、舆情监控、犯罪预测、反恐等活动.例如,作为大数据的强力倡导者,奥巴马及其团队创新性地将大数据应用到竞选活动中,通过对近 2 年搜集、存储的海量数据进行分析挖掘,寻找和锁定潜在的己方选民,运用数字化策略定位拉拢中间派选民及筹集选举资金,成为将大数据价值与魅力发挥到淋漓尽致的典型.借助大数据,还能逐步实现立体化、多层次、全方位的电子政务公共服务体系,推进信息公开,促进网上电子政务开展,创新社会管理和服务应用,增强政府和社会、百姓的双向交流、互动[49].

3.2　大数据在网络通信业的应用

大数据与云计算相结合所释放出的巨大能量,几乎波及到所有的行业,而信息、互联网和通信产业将首当其冲.特别是通信业,在传统话音业务低值化、增值业务互联网化的趋势中,大数据与云计算有望成为其加速转型的动力和途径.对于大数据而言,信息已经成为企业战略资产,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、支撑系统中产生海量有价值的数据,基于这些大数据的商业智能应用将为通信运营商带来巨大机遇和丰厚利润.

例如,电信业者可通过数以千万计的客户资料,分析出多种使用者行为和趋势,卖给需要的企业,这是全新的资料经济.中国移动通过大数据分析,对企业运营的全业务进行针对性的监控、预警、跟踪,系统在第一时间自动捕捉市场变化,再以最快捷的方式推送给指定负责人,使他在最短时间内获知市场行情.据计世资讯预测,到 2015 年,电信业大数据应用市场规模预计将达到 18.3 亿元.

3.3　大数据在医疗行业的应用

伴随医疗卫生行业信息化进程的发展,在医疗业务活动、健康体检、公共卫生、传染病监测、人类基因分析等医疗卫生服务过程中将产生海量高价值的数据.数据内容主要包括医院的 PACS 影像、B 超、病理分析、大量电子病历、区域卫生信息平台采集的居民健康档案、疾病监控系统实时采集的数据等[10].面

对大数据,医疗行业遇到前所未有的挑战和机遇.例如,Seton Healthcare 是采用 IBM 最新沃森技术医疗保健内容分析预测的首个客户.该技术允许企业找到大量病人相关的临床医疗信息,通过大数据处理,更好地分析病人的信息.在加拿大多伦多的一家医院,针对早产婴儿,每秒钟有超过 3 000 次的数据读取[48].通过这些数据分析,医院能够提前知道哪些早产儿出现问题并且有针对性地采取措施,避免早产婴儿夭折.大数据让更多的创业者更方便地开发产品,比如通过社交网络来收集数据的健康类 App.也许在数年后,它们搜集的数据能让医生给你的诊断变得更为精确,比方说不是通用的成人每日 3 次,1 次 1 片,而是检测到你的血液中药剂已经代谢完成会自动提醒你再次服药.社交网络为许多慢性病患者提供临床症状交流和诊治经验分享平台,医生借此可获得在医院通常得不到的临床效果统计数据.基于对人体基因的大数据分析,可以实现对症下药的个性化治疗.对于公共卫生部门,可以通过全国联网的患者电子病历库,快速检测传染病,进行全面疫情监测,并通过集成的疾病监测和响应程序,快速进行响应.

3.4 大数据在能源行业的应用

能源勘探开发数据的类型众多,不同类型数据包含的信息各具特点,只有综合各种数据所包含的信息才能得出真实的地质状况.能源行业企业对大数据产品和解决方案的需求集中体现在:可扩展性、高带宽、可处理不同格式数据的分析方案.智能电网现在欧洲已经做到了终端,也就是所谓的智能电表.在德国,为了鼓励利用太阳能,会在家庭安装太阳能,除了卖电给你,当你的太阳能有多余电的时候还可以买回来.通过电网收集每隔 5 min 或 10 min 收集一次数据,收集来的这些数据可以用来预测客户的用电习惯等,从而推断出在未来 2~3 个月时间里,整个电网大概需要多少电.预测后,就可以向发电或者供电企业购买一定数量的电.因为电有点像期货一样,如果提前买就会比较便宜,买现货就比较贵.通过预测可以降低采购成本[48].维斯塔斯风力系统,依靠的是 BigInsights 软件和 IBM 超级计算机,然后对气象数据进行分析,找出安装风力涡轮机和整个风电场最佳的地点.利用大数据,以往需要数周的分析工作,现在仅需要不足 1 h 便可完成.

3.5 大数据在零售行业的应用

从商业价值来看,大数据究竟能往哪些方面挖掘出巨大的商业价值呢? 根据 IDC 和麦肯锡的大数据研究结果的总结,大数据主要能在以下 4 个方面挖掘出巨大的商业价值:对顾客群体细分,然后对每个群体量体裁衣般地采取独特的行动;运用大数据模拟实境,发掘新的需求和提高投入的回报率;提高大数据成果在各相关部门的分享程度,提高整个管理链条和产业链条的投入回报率;进行商业模式、产品和服务的创新.

在商业领域,沃尔玛公司每天通过 6 000 多个商店,向全球客户销售超过 2.67 亿件商品,为了对这些数据进行分析,HP 公司为沃尔玛公司建造了大型数据仓库系统,数据规模达到 4 PB,并且仍在不断扩大[48].沃尔玛公司通过分析销售数据,了解顾客购物习惯,得出适合搭配在一起出售的商品,还可从中细分顾客群体,提供个性化服务.在金融领域,华尔街德温特资本市场公司通过分析 3.4 亿微博账户留言,判断民众情绪,依据人们高兴时买股票、焦虑时抛售股票的规律,决定公司股票的买入或卖出.

阿里巴巴公司根据在淘宝网上中小企业的交易状况筛选出财务健康和讲究诚信的企业,对他们发放无需担保的贷款.当我们去购物时,我们的数据会结合历史购买记录和社交媒体数据来为我们提供优惠券、折扣和个性化优惠.零售企业也监控客户的店内走动情况以及与商品的互动,它们将这些数据与交易记录相结合来展开分析,从而在销售哪些商品、如何摆放货品以及何时调整售价上给出意见,此类方法已经帮助某领先零售企业减少了 17% 的存货,同时在保持市场份额的前提下,增加了高利润率自有品牌商品的比例.

3.6 大数据在气象行业的应用

与世界大数据时代的进程相同,气象数据量不断翻番.目前,每年的气象数据已接近 PB 量级(1 024 GB = 1 TB, 1 024 TB = 1 PB)(中国气象报. http://www.cma.gov.cn/kppd/kppdsytj/201306/t20130627_217674.html).

以气象卫星数据为例:虽然气象卫星是用来获取与气象要素相关的各类信息的,然而在森林草场火灾、船舶航道浮冰分布等方面,气象卫星却同样也能发挥出跨行业的实时监测服务价值.气象卫星、天气雷达等非常规遥感遥测数据中包含的信息十分丰富,有可能挖掘出新的应用价值,从而拓展气象行业新的业务领域和服务范围.比如,可以利用气象大数据为农业生产服务.美国硅谷有家专门从事气候数

据分析处理的公司,从美国气象局等数据库中获得数十年来的天气数据,然后将各地降雨、气温、土壤状况与历年农作物产量的相关度做成精密图表,可预测各地农场来年产量和适宜种植品种,同时向农户出售个性化保险服务.气象大数据应用还可在林业、海洋、气象灾害等方面拓展新的业务领域.

除了上述行业应用外,大数据在教育科研、生产制造、金融保险、交通运输等行业也有密切应用.大数据在金融行业可用于客户洞察、运营洞察和市场洞察.大数据在智能交通、智慧城市建设方面也有出色表现.随着社会、经济的发展,各行业各类用户对于智能化的要求将越来越高,今后大数据技术会在越来越多领域得到广泛应用,通过大数据的采集、存储、挖掘与分析,大数据在营销、行业管理、数据标准化与情报分析和决策等领域将大有作为,将极大提升企事业单位的信息化服务水平.随着云计算、物联网、移动互联网等技术的快速发展,大数据未来发展空间将更加广阔.

4 小结

继实验科学、理论科学、计算机科学之后,以大数据为代表的数据密集型科学将成为又一次历史性技术变革,成为人类科学研究的第四大范式[50].大数据虽然表面上是个技术术语,但实际上涉及到社会生活、经济运行、国防军事、科学技术等方方面面.面对大数据的机遇与挑战,大数据时代呼唤创新型人才,给国内自主处理器芯片研发行业提供重大战略机遇,将会有更多应用大数据技术的新兴的公司和运营模式出现.美国 Gartner 咨询公司预测大数据将为全球带来 440 万个 IT 新岗位和成千上万个非 IT 岗位.

根据中国计算机学会大数据专委会发布的2014年大数据发展 10 大趋势预测[10]:大数据从“概念”走向“价值”、大数据架构的多样化模式并存、大数据安全与隐私、大数据分析与可视化、大数据产业成为战略性产业、数据商品化与数据共享联盟化、基于大数据的推荐与预测流行、深度学习与大数据智能成为支撑、数据科学的兴起和大数据生态环境逐步完善.

据 Gartner 公司最新发布的《2012—2013 年技术曲线成熟度报告》指出,大数据成为市场的主流产品需要 2~5 年,大数据核心处理技术尚未成熟.所以,也要清醒地意识到,我国发展大数据产业不能再重蹈光伏、风电、物联网等新兴产业盲目跟风、一哄而上的老路,而要注意科学规划,提出适合我国实际情况的大数据战略和发展路径,形成良好的大数据发展环境.

美国维克托在《大数据时代》一书中提到:“未来,数据将会像土地、石油和资本一样,成为经济运行中的根本性资源.”

总之,未来信息世界是:三分技术,七分数据,得数据者得天下.

参考文献

References

[1] Nature.Big Data[EB/OL].[2014-08-23].http://www.nature.com/news/specials/bigdata/index.htm

[2] Science.Special online collection:Dealing with data[EB/OL].(2011-02-11).[2014-08-23].http://www.sciencemag.org/site/special/data/

[3] Big Data across the federal government.[EB/OL].[2014-08-23].http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf

[4] Agrawal D,Bernstein P,Bertino E,et al.Challenges and opportunities with Big Data[R].Cyber Center Technical Reports,2012

[5] 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.Research status and scientific thinking of Big Data[J].Bulletin of Chinese Academy of Sciences,2012,27(6):647-657

[6] Manyika J,Chui M,Brown B,et al.Big Data:The next frontier for innovation,competition,and productivity[EB/OL].[2014-09-02].http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[7] 冯登国,张敏,李昊.大数据安全与隐私保护[J].计算机学报,2014,37(1):246-258
FENG Dengguo,ZHANG Min,LI Hao.Big Data security and privacy protection[J].Chinese Journal of Computers,2014,37(1):246-258

[8] 孟小峰,慈祥.大数据管理:概念、技术与挑战[J].计算机研究与发展,2013,50(1):146-169
MENG Xiaofeng,CI Xiang.Big Data management:Concepts,techniques and challenges[J].Journal of Computer Research and Development,2013,50(1):146-169

[9] 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 on Big Data research[J].Communications of China Computer Federation,2012,8(9):8-15

[10] 中国计算机学会大数据专家委员会.中国大数据技术与产业发展白皮书[R].2013
Big Data Expert Committee in China Computer Federation.White paper on China's Big Data technology

and industry development[R].2013
[11] 周晓方,陆嘉恒,李翠平,等.从数据管理视角看大数据挑战[J].中国计算机学会通讯,2012,8(9):16-20
ZHOU Xiaofang, LU Jiaheng, LI Cuiping, et al. Big Data challenges from the point view of data management[J]. Communications of China Computer Federation, 2012, 8(9):16-20
[12] Vardi M. On the integrity of databases with incomplete information[C]// Proceedings of the 5th ACM SIG ACT-SIGMOD Symposium on Principles of Database Systems, 1985:252-266
[13] Gottlob G, Zicari R. Closed world databases opened through null values[C]// Bancilhon F, deWitt D J. Proceedings of the 14th International Conference on Very Large Databases, 1988:50-61
[14] Dalvi N N, Suciu D. Management of probabilistic data: Foundations and challenges[C]// Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems, 2007: 1-12, doi:10.1145/1265530.1265531
[15] Kanagal B, Li J, Deshpande A. Sensitivity analysis and explanations for robust query evaluation in probabilistic databases[C]// SIGMOD, 2011:841-852
[16] Li J, Saha B, Deshpande A. A unified approach to ranking in probabilistic databases[J]. The VLDB Journal, 2011, 20(2):249-275
[17] Cooper B F, Sample N, Franklin M J, et al. A fast index for semistructured data[C]// Proceedings of the International Conference on VLDB, 2001:341-350
[18] Times N Y. Power, pollution and the internet[EB/OL]. [2014-08-25]. http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?pagewanted=all
[19] 刘锋.互联网进化论[M].北京:清华大学出版社,2012
LIU Feng. Internet evolution[M]. Beijing: Tsinghua University Press, 2012
[20] Haas L. Integrating extremely large data is extremely challenging[C]// Proceedings of XLDB Asia 2012. http://idke.ruc.edu.cn/xldb/www.xldb-asia.org/program.html
[21] Li X, Dong X L, Lyons K, et al. Truth finding on the deep web: Is the problem solved?[C]// Proceedings of the 39th International Conference on Very Large Data Bases (VLDB'2013), 2013:97-108
[22] Arasu A, Chaudhuri S, Chen Z, et al. Experiences with using data cleaning technology for bing services[J]. IEEE Data Engineering Bulletin, 2012, 35(2):14-23
[23] Ghemawat S, Gobioff H, Leung S-T. The Google file system[C]// Proceedings of the 19th ACM Symposium on Operating Systems Principles, 2003:29-43
[24] HDFS Architecture Guide[EB/OL]. [2014-08-25]. http://hadoop.apache.org/docs/stable/hdfs_design.htm, 20130512
[25] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1):107-113
[26] 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 Symposium on Networked Systems Design and Implementation, 2012:2-16
[27] Gonzalez J E, Low Y, Gu H, et al. PowerGraph: Distributed graph-parallel computation on natural graphs[C]// Proceeding of the 10th USENIX Symposium on Operating Systems Design and Implementation, 2012: 17-30
[28] 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. Big Data computing paradigm divergence and blending[J]. Programmer, 2013(9):104-108
[29] Melnik S, Gubarey A, Long J J, et al. Dremel: Interactive analysis of web-scale datasets[J]. Communications of the ACM, 2011, 54(6):114-123
[30] Kumar R. Two computational paradigm for Big Data[EB/OL]. [2014-08-25]. http://kdd2012.sigkdd.org/sites/images/summerschool/Ravi-Kumar.pdf
[31] Neumeyer L, Robbins B, Nair A, et al. S4: Distributed stream computing platform[C]// IEEE International Conference on Data Mining Workshops, 2010:170-177
[32] Goodhope K, Koshy J, Kreps J, et al. Building LinkedIn's real time activity data pipeline[J]. IEEE Data Engineering Bulletin, 2012, 35(2):33-45
[33] Zaharia M, Das T, Li H Y, et al. Discretized streams: An efficient and fault-tolerant model for stream processing on large cluster[C]// Proceedings of the 4th USENIX conference on Hot Topics in Cloud Computing, 2012:10-16
[34] Bu Y Y, Howe B, Balazinska M, et al. HaLoop: Efficient iterative data processing on large cluster[J]. Proc VLDB Endow, 2010, 3(1/2):285-296
[35] Ekanayake J, Li H, Zhang B J, et al. Twister: A runtime for iterative MapReduce[C]// Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, 2010: 810-818, doi: 10.1145/1851476.1851593
[36] Zhang Y F, Gao Q X, Gao L X, et al. iMapReduce: A distributed computing framework for iterative computation[J]. Journal of Grid Computing, 2012, 10(1):47-68
[37] Elnikety E, Elsayed E, Ramadan H E. iHadoop: Asynchronous iterations for mapreduce[C]// IEEE 3rd International Conference on Cloud Computing Technology and Science, 2011:81-90
[38] Malewicz G, Austern M, Bik A, et al. Pregel: A system for large-scale graph processing[C]// Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, 2010:135-146
[39] Shao B, Wang H X, Li Y T, et al. Trinity: A distributed graph engine on a memory cloud[C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013:1-12
[40] Xin R, Gonalez J, Franklin M. Graph X: A resilient distributed graph system on spark[C]// Proceedings of the First International Workshop on Graph Data Management Experience and System, 2013:12-18
[41] InfiniteGraph, the Distributed Graph Database[EB/OL]. [2014-08-25]. http://www.infinitegraph.com/. 2011: 7-29
[42] Kang U, Chau D H, Faloutsos C. PEGASUS: Mining

billion-scale graphs in the cloud [C] // IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012: 5341-5344, doi: 10.1109/ICASSP.2012.6289127

[43] Gubanov M, Pyayt A. MEDREADFAST: A structural information retrieval engine for big clinical text[C] // Proceedings of the 13th International Conference on Information Reuse and Integration (IRI), 2012: 371-376

[44] Das S, Sismanis Y, Beyer K S, et al. Ricardo: Integrating R and Hadoop [C] // Proceedings of the 2010 International Conference on Management of Data, 2010: 987-998

[45] Ahrens J, Brislawn K, Martin K, et al. Large-scale data visualization using parallel data streaming[J]. IEEE Computer Graphics and Applications, 2001, 21(4): 34-41

[46] Scheidegger L, Vo H T, Kruger J, et al. Parallel large data visualization with display walls[C] // Proceedings of the 2012 Conference on Visualization and Data Analysis (VDA), 2012: 1-8

[47] Schadt E E. The changing privacy landscape in the era of Big Data[J]. Molecular System Biology, 2012, 8(1): 612

[48] 大数据应用与案例分析[EB/OL].[2014-08-25].中国人民大学经济学论坛,http://bbs.pinggu.org/bigdata
Application and analysis of Big Data [EB/OL]. Economics Forum of Renmin University of China, http://bbs.pinggu.org/bigdata

[49] 刘琼.专家解读大数据时代的美国经验与启示[EB/OL].[2014-08-25].人民网(人民论坛),http://theory.people.com.cn/n/2013/0521/c112851-21551972.html
LIU Qiong. Expert interpretation: American experience and enlightenment for Big Data era[EB/OL].[2014-08-25].People's Daily Online: People's Tribune, http://theory.people.com.cn/n/2013/0521/c112851-21551972.html

[50] 工业和信息化部赛迪智库.大数据时代信息安全面临的挑战与机遇[EB/OL].[2014-08-25].科技日报,http://digitalpaper.stdaily.com/http_www.kjrb.com/kjrb/html/2013-06/24/content_209820.htm?div=-1
CCID think tank, Ministry of Industry and Information Technology of the PRC. Challenges and opportunities of information security in time of Big Data [EB/OL].[2014-08-25]. Science and Technology Daily, http://digitalpaper.stdaily.com/http_www.kjrb.com/kjrb/html/2013-06/24/content_209820.htm?div=-1

Big Data: Conceptions, key technologies and application

FANG Wei[1,2,3] ZHENG Yu[1,2] XIU Jiang[1,2]

1 Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044

2 School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044

3 State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing 210046

Abstract With the rapid development of internet of things, cloud computing, and mobile internet, the rise of Big Data has attracted more and more concern, which brings not only great benefits but also crucial challenges on how to manage and utilize Big Data better. This paper describes the main aspects of Big Data including definition, data sources, key technologies, data processing tools and applications, discusses the relationship between Big Data and cloud computing, internet of things and mobile internet technology. Furthermore, the paper analyzes the core technologies of Big Data, Big Data solutions from industrial circles, and discusses the application of Big Data. Finally, the general development trend on Big Data is summarized. The review on Big Data is helpful to understand the current development status of Big Data, and provides references to scientifically utilize key technologies of Big Data.

Key words Big Data; cloud computing; Big Data processing; distributed system; NoSQL