

大数据

2

韩晶/HAN Jing, 宋美娜/SONG Meina

(北京邮电大学计算机学院, 北京 100876)

[编者按] 数据是与自然资源一样重要的战略资源,大数据技术就是从数量巨大、结构复杂、类型众多的数据中,快速获得有价值信息的能力,它已成为学术界、企业界甚至各国政府关注的热点。本讲座将分3期对大数据进行讨论:第1期介绍了大数据的提出、含义、特点,大数据和云计算的关系以及大数据典型应用;第2期介绍大数据获取、存贮、搜索、分享、分析、可视化等方面的关键技术,并对当前热点技术——可视化进行重点分析;第3期将探讨数据流挖掘等实时数据分析技术,介绍大数据中非结构化数据处理和挖掘技术,并给出大数据发展面临的挑战与应用前景。

中图分类号: TN91 文献标志码: A 文章编号: 1009-6868 (2013) 02-0058-05

5 大数据生态系统

5.1 大数据生态系统

2011年6月,IBM架构师Stephen Watt在《Deriving new business insights with Big Data》文中对大数据生态系统进行了简单描述,提出大数据生态系统实际上就是数据的生命周期,即数据采集、存储、查找、分析和可视化的过程^[1],见图1。

在这样的生态系统中,每个环节都存在着不同的商业需求,而需求的出现必然会导致创新的产生。所以,在每一个环节都有不少企业在深耕自己所在的领域,试图通过新技术和新方法来实现新的商业模式。

5.2 大数据生态图谱

随着大数据生态系统的逐步形成,很多人在尝试绘制和更新大数据生态系统图谱,希望通过对大数据领域的公司、技术、产品进行细分,及时了解到大数据生态系统全貌。在众多图谱当中,比较有代表性的是美国On Grid Ventures公司Matt Turck等人

于2012年10月绘制更新的大数据生态图谱V2.0^[2],如图2所示。

尽管各个图谱的分类方法、全面性、时效性、权威性各不相同,但我们仍可以观察到:

(1)大数据领域的企业主要集中在数据集市、数据存储(基础设施)、数据分析、数据应用4个层面,其中数据应用层面又包含数据服务、数据检索、商务智能、可视分析等。这正符合数据科学中对数据全生命周期管理的描述。此外,很多企业业务覆盖大数据多个层面,有的企业甚至已经建立了完整的大数据栈,成为大

数据应用服务提供商。

(2)在大数据领域,活跃着的除了IBM、ORACLE等众多知名公司外,像Splunk、Tableau等专业大数据公司也及时跟上了大数据的浪潮,成功地获得了投资者和业界的关注。

(3)开源软件与大数据的结合迸发出惊人的颠覆性力量,更多厂商开始使用开源大数据工具,以支持其大数据业务。

大数据生态系统中覆盖大量的技术和产品,其中一些在大数据技术发展道路中起到了巨大的推动作用。

(1)Hadoop

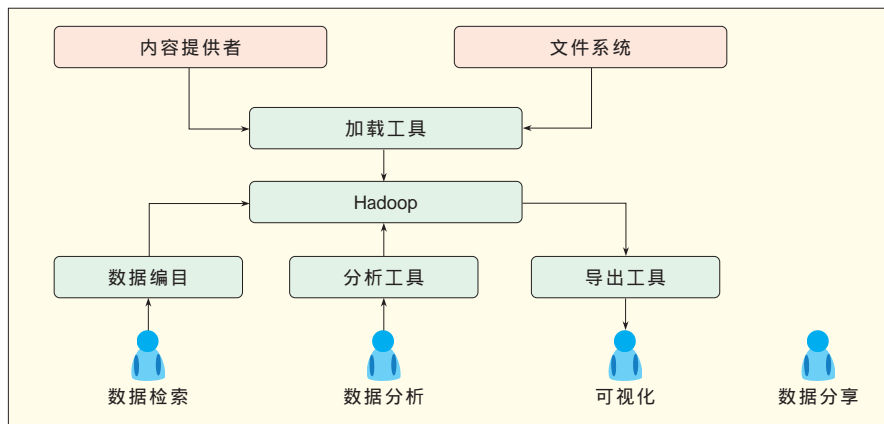


图1 大数据生态系统

收稿日期: 2013-01-05

网络出版时间: 2013-03-01



▲图2 大数据生态系统

在大数据时代, Hadoop 可以说是最耀眼的明星。凭借其开源和易用的特性, Hadoop 不仅是大数据时代数据处理的首选, 也是拥有海量数据处理需求的公司的标准配置。此外, 许多商业创新也都围绕 Hadoop 展开的, 并在大数据时代占据一席之地, 如 Cloudera 推出的软件发布包可以帮助企业更方便地搭建以 Hadoop 为中心的数据管理平台; MapR 则将 Hadoop 的速度改造为原来的 3 倍; 海量数据管理软件商 Platfora 旨在提供一个更为友好且更具操作性的用户界面, 它可以兼容包括 Cloudera 和 MapR 等多种 Hadoop 版本, 能够大大降低使用 Hadoop 的门槛; 而 AsterData (已被 TeraData 收购) 的核心技术 SQL-to-MapReduce 可将海量非结构化数据的处理技术和结构化数据的数据仓库技术结合, 以弥补传统数据仓库的公司所欠缺的高速处理海量非结构化数据的能力。

(2) NoSQL

与 Hadoop 密切相关的 NoSQL 也一直是大数据领域的热点。NoSQL 凭借其高性能和可扩展性等优势, 成

为关系数据库的强劲对手, 在大数据时代占据一席之地。根据存储模型和特征, NoSQL 大致可分为列存储、文档存储、key-value 存储、图存储、对象存储、XML 数据库等类型, 虽然也存在个别数据库可被归为多种类别的现象, 其中, HBase、MongoDB、Cassandra、CouchDB、Neo4j、HyperTable 等 NoSQL 已被相当多的企业和开发人员所熟知。

(3) NewSQL

无论 NoSQL 是被解释为 NoSQL, 还是后来的 Not Only SQL, 其不支持结构化查询语言 (SQL) 语言的特性为开发人员带来诸多不便。因此, 为了同时满足高性能和支持 SQL 两个方面, NewSQL 被设计出来。NewSQL 作为全新的关系数据库产品, 或将关系模型的优势发挥到分布式体系结构中, 或将关系数据库的性能提升到不必进行横向扩展的程度, 这使得 NoSQL 面临前所未有的挑战。典型的 NewSQL 有 VoltDB、Marklogic、Xeround、NuoDB 等。

(4) Data Marketplace

除了解决大数据处理、存储问题

之外, 开放数据资源也在相当程度上加速了大数据技术的发展。目前大部分的企业所面对的数据都是由内部系统或者交易记录日志之类的东西所产生的, 然而如果能够获得企业自己无法获得, 或者已经被处理过的外部数据, 那么内外数据融合分析后产生的价值将不可估量。因此, 能够下载或者访问数据集, 自然而然也就成为了商业需求, 甚至美国政府都推出了官方的数据集网站。

2009 年 5 月, 美国联邦政府正式启用了官方公共数据资源分享网站 Data.gov, 其数据内容涵盖了所有美国联邦政府行政部门在运营管理过程中采集、生产或转换而来的、有潜在价值的、可供再次开发利用的数据集。Data.gov 鼓励个人开发者使用政府发布的数据集, 开发出新颖的应用。值得一提的, 该网站于近期正式对外发布了全新的 开源政府平台 (OGPL), 该平台的代码将会对全球的开发者开放。

在中国, 数据堂 (datatang.com) 是目前最为专业的科研数据共享服务平台, 该平台致力于为全球科研机构、企业及个人提供科研数据支持, 其数据内容主要是科研数据集, 同时也提供浮动车历史数据、路况历史数据和车牌数据等, 用户也可以上传发布自己的数据。通过该平台不仅使得中国的科研机构、企业、高校和个人之间可以充分共享数据, 也促进各类科研数据价值的最大化。

在全球范围的大数据热潮中, 对于大多数企业来说, 大数据与自己有什么关系? 如何快速直观地理解和发现大数据中的价值? 没有足够大数据的情况下如何才能在大数据时代获益? 虽然这些问题还没有完美的答案, 但许多企业已经进行了积极的尝试, 通过数据可视化尝到了大数据的甜头。

6 可视化和可视分析

在众多描述大数据的词语中,

金矿、油田等的描述最为常见,这意味着人们开始意识到大数据中蕴含着丰富的价值。然而,巨大的数量、数据的固有复杂性及未知的分析目标都放大了任务的难度。如果能够有一种简单的方式对数据规律进行直观展现,必将使大数据中的价值得到快速理解和发现,可视化就是这样的方式。

6.1 数据可视化、信息可视化和可视分析概述

可视化由来已久,1861年法国工程师 Charles Joseph Minard 绘制了《拿破仑征俄战役图》可以看作可视化领域的经典案例。到了18世纪后期数据图形学诞生,抽象信息的视觉表达手段一直被人们用来揭示数据及其他隐匿模式的奥秘。随着20世纪50年代计算机图形学的出现,信息技术加速了可视化的演变。时至今日,可视化已经发展为数据可视化、科学可视化、信息可视化、可视分析这几大方向。

数据可视化起源于20世纪50年代,其基本思想是将数据库中每个数据项作为可视化图形中单个元素,同时将数据的各个属性值以多维数据的形式表示,通过从不同维度观察数据而达到对数据深入洞察和分析的目的。

科学可视化是一个典型的交叉学科,源于1987年布鲁斯·麦考梅克等人编写的网络文件系统(NFS)报告《Visualization in Scientific Computing》(意为科学计算之中的可视化)。科学可视化主要是将具有几何结构的三维数据转换为图像,应用领域涵盖科学和工程的多个方面。

信息可视化也是一个跨学科领域,出现于20世纪90年代,旨在为许多应用领域之中大规模非数值型信息资源的视觉呈现提供支持,这些资源可能是软件系统之中众多的文件、大规模并行程序的日志痕迹信息、网站内容等。与科学可视化相

比,信息可视化侧重于异质数据集,如非结构化文本当中的点。

可视分析则起源于2005年,它是一门通过交互可视界面来分析、推理和决策的科学,通过将可视化和数据处理分析方法结合,提高可视化质量的同时也为用户提供更完整的大规模数据解决方案^[3]。如今,针对可视分析的研究和应用逐步发展,已经覆盖科学数据、社交网络数据、电力等多个行业。

虽然在这几大方向之间的边界还未完全清晰,不过,其相互关系和区别可以总结如下:数据可视化外延不断扩大,可以认为数据可视化包含科学可视化、信息可视化和可视分析;科学可视化处理的是那些具有天然几何结构的数据;信息可视化处理的是异质的抽象的数据结构;可视分析则主要通过意会、推理、互动融合的方式来挖掘数据中的问题和原因。

可视化融合了问题的求解和艺术表现方式两个方面,允许我们同时通过理性和感官方式来感受数据,那么怎样才是成功的可视化? Noah Iliinsky 在《数据可视化之美》一书中提到^[4],一个称得上美的可视化,必须具备新颖、充实、高效和美观4个关键要素。新颖性体现在必须从崭新的视角观察数据,传统可视化展现方式(如柱形图)虽易理解,但不够新奇有趣,是不足以激发读者新的理解的;充实性体现在可视化一定要为读者提供获取信息的途径,从而向读者传递信息甚至知识;高效性指成功的可视化须尽可能直截了当,而不允许展示太多与目标和主题无关的信息;美观是指合理的图形构建(坐标轴、布局、色彩、线条等)是实现可视化之美的必要因素。这四要素必须同时具备,否则不能对数据进行有意义地呈现。

6.2 可视化之美

美丽的可视化作品不同于传统的可视化,它们能够通过创造不同于

惯例的图形构建方式,揭示数据显性和隐性的特征,使读者在对可视化效果感到惊喜的同时收获启示。通过以下的一些案例我们可以充分体会到这一点。

(1) 电信数据可视化 《都市移动族》

当今城市被通讯数据所充斥,每个打电话发短信的人都生成特定时间地点的数据包,然而这些数据中有什么规律? 2008年,法国 faberNovel 公司对巴黎国际音乐节和新年夜产生的手机数据进行监测和可视化,帮助法国电信运营商 Orange 建立《都市移动族 Urban Mobs》(图3)^[5]。它不仅让我们发现城市活动中丰富的一面,同时也使电信运营商在流量分析、业务推荐等方面获得启示。

(2) 电信数据可视化 《活力日内瓦》

手机可以看作是实时记录并上传用户地理位置信息的移动传感器,2011年,瑞士日内瓦市政府与 Interactive Things 公司合作,将市民每天在日内瓦市的行动轨迹的手机GPS数据进行记录,并制作城市生活(Ville Vivante)^[6]动态显示瑞士电信每时每刻的数据流向。图4展示的是晚上六点到午夜之间人们移动的轨迹。这种融合基于位置的服务(LBS)和电信数据的可视化方式不仅使政府和公众对城市生活有了重新认识,同时也产生不可估量的经济政治效益。

(3) 智慧城市 《实时新加坡》

现代城市中每天都在产生海量的数据,如何才能让政府和市民更快了解城市每时每刻的变化,帮助政府提高管理效能,为市民提供生活便利? 2011年,美国麻省理工大学可感知实验室为新加坡建立了 LIVESingapore 实时新加坡平台^[7](图5),该平台能够为公众提供实时的城市活动及环境信息。其中,实时通讯显示新加坡语音通讯、短信及网络使用情况,等时地图实时呈



图3
都市移动族 Urban Mobs



图4
活力热内瓦

现新加坡居民交通耗时情况；雨天打车 结合降雨监测和出租车数据进行可视化，从而在雨天智能调配出租车；城市热岛 将新加坡区域温度与能源消耗的关系进行可视化。通过对城市生活、环境数据的可视化，可助力提高城市公共服务质量，改善市民生活，真正意义上实现智慧城市。

(4) 北京大学 PKUVIS 微博可视分析工具

结合社会网络分析中的概念和可视化的呈现方法，佐以统计和智能数据挖掘的方法，可以为海量复杂社会网络提供快速、直观和智能的分析和呈现方法^[8]。2012年北京大学可视化与可视分析小组开发了支持可视化浏览和分析微博热点事件的 PKUVIS 微博可视分析工具（图6）^[9]。该工具将一条条独立的微博连接起来，通过直观的视图清晰地呈现出一个事件中微博转发的过程，从而让用户能够迅速地发现事件中的关键人

物、关键微博、重要观点，同时通过可视化的方式可以更好地分析新浪微博传播脉络以及事件的发生与发展的过程。

(5) 电力大数据可视化

美国 Space-Time 是一家提供新一代地理空间和可视化解决方案的创业公司，2011年，Space-Time 为美

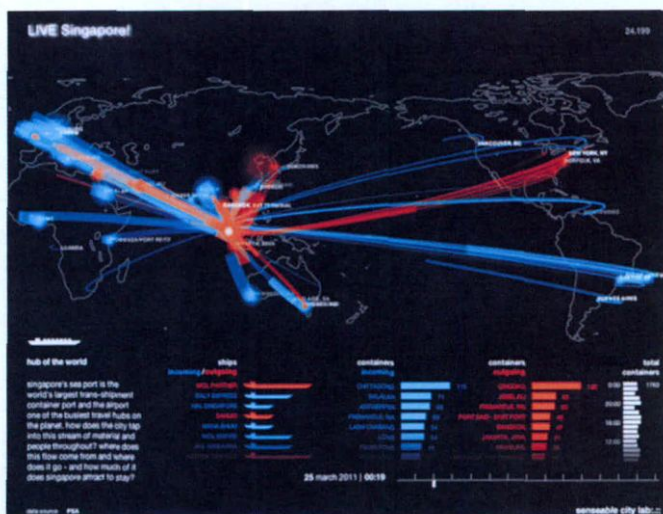
国加州独立系统运营商设计了一套可以实时监控电力传输系统能源基础设施的可视化软件 Space-Time Insight（图7）^[10]，该可视化系统通过控制室中的一个80英寸的显示屏，在地图上实时展示长达25 000 km的输电线路状况，工作人员一旦发现一个地区出现了问题，就可以根据该地区问题的严重性和临近地区的反应来做决策。不仅简化了日常运营复杂度，还在尽可能降低影响的情况下解决问题。这种大数据可视化实践对中国的电力大数据分析展示乃至整个能源相关行业都具有巨大的参考价值。

6.3 开源可视化工具

如果读者已经被以上可视化案例所吸引，并且愿意尝试将企业数据进行可视化，那么开源的数据可视化编程语言和环境将会是不错的选择。可视化领域中重要而常用的可视化编程语言和环境有 Processing、Processing.js、R、D3、Impure、ParaView、Circos 等，它们具备的一个共同特点就是为用户提供了常见的专业可视化模版或图形库，用户可以通过简单调用即可很快实现可视化效果，此外，由于软件的开源优势，专业用户可以根据其需求，对图形源代码进行定制修改。

在可见的未来，大数据可视化机

图5
实时新加坡



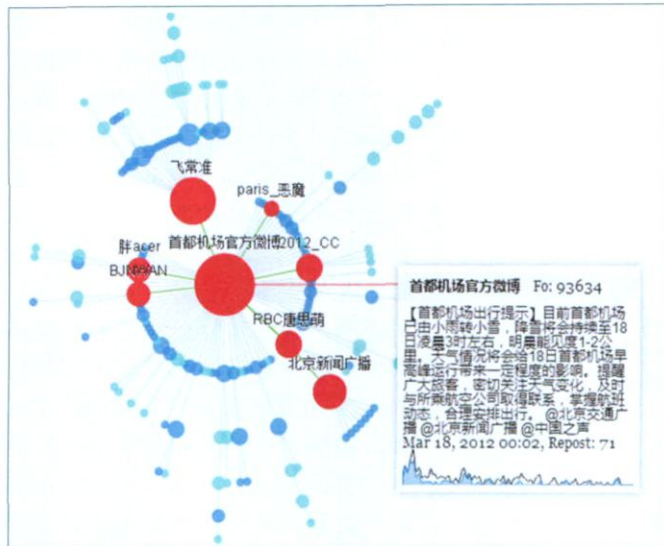


图6 北京大学 PKU VIS 微博可视分析工具微博可视化



图7 Space-Time Insight 电力大数据可视化

遇挑战并存^[11],大数据可视化将越来越广泛地为各领域所使用,也将引发新一轮的投资热潮,而构建面向电子政务、电信、电力等特定行业大数据的可视分析工具是一个可以深入探

索的重要发展方向。 (待续)

参考文献

- [1] Deriving new business insights with big data [EB/OL]. <http://www.ibm.com/developerworks/library/os-bigdata>

- [2] Big data landscape v2.0[EB/OL]. <http://www.ongridventures.com/2012/10/23/the-big-data-landscape/>
- [3] 俞宏峰. 大规模科学可视化[J]. 中国计算机学会通讯, 2012, 8(9): 29-37.
- [4] STEELE J, ILIINSKY N. Beautiful visualization [M]. Sebastopol, CA, USA: O'Reilly Media, 2010.
- [5] Urban Mobs[EB/OL]. <http://www.urbanmobs.fr/en/>
- [6] Ville Vivante[EB/OL]. <http://www.villevivante.ch/>
- [7] LIVE Singapore[EB/OL]. <http://www.live-singapore.com.sg/>
- [8] 袁晓如, 张昕, 肖何等. 可视化研究前沿及展望[J]. 科研信息化技术与应用, 2011, 2(4): 3-13.
- [9] PKU VIS 微博可视分析工具[EB/OL]. <http://vis.pku.edu.cn/weibova/weiboevents/>
- [10] Space-time insight[EB/OL]. <http://www.spacetimeinsight.com/>
- [11] 黄伯仲, 沈汉威, 克里斯托弗·约翰逊等. 超大规模数据可视分析十大挑战[J]. 中国计算机学会通讯, 2012, 8(9): 38-43.

作者简介



韩晶, 北京邮电大学计算机学院在读博士; 主要研究方向为大数据管理; 已发表论文10余篇, 申请专利和软著9项。



宋娜娜, 北京邮电大学计算机学院教授; 主要研究方向为分布式系统、服务计算、数据工程等; 已发表论文50余篇, 申请专利20余项。

综合信息

中兴通讯 iCity 智慧城市解决方案进入 GSMA 全球移动大奖 短名单

【本刊讯】2013年2月21日消息, GSMA 公布了第十八届 全球移动大奖 的入围短名单, 中兴通讯的 iCity 智慧城市解决方案在 600 多个提案中脱颖而出, 进入无线智慧城市最佳创新 奖项的短名单, 也是进入该短名单唯一的智慧城市综合解决方案。

GSMA 全球移动大奖 是移动行业领域的最高奖项, 由 160 多名评委共同打分评定, 包括行业分析师、媒体、学者以及 14 家运营商的技术执行官。这是中兴通

讯面向政府和企业的解决方案第一次进入该类奖项短名单。

中兴通讯 iCity 智慧城市解决方案涵盖 维稳定、促增长、保民生 三大领域, 包含电子政务、智慧交通、平安城市、智慧教育等 12 项重点应用, 首次在业界提出包含 信息 (Information)、智能 (Intelligent)、创新 (Innovation) 以及 我和城市 (I with City) 的 4I 智慧城市理念, 并引入云计算平台。中兴通讯此次能够进入短名单, 是中兴通讯产品实力的最佳体现。中兴通讯希望能够创建更加智慧的城市, 助力经济发展。