

大数据

1

于艳华/YU Yanhua, 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期将介绍大数据获取、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)01-0060-04

1 大数据概念的提出

高度数字化使得数据充斥着整个地球,大数据也成为一种新的自然资源^[1],并成为当前所有行业最热门的话题之一。大数据概念的提出可以追溯到《自然》杂志2008年9月专刊中发表的文章:《Big Data: Science in the Petabyte Era》^[2],此后大数据这个概念被广泛应用和传播。2011年,麦肯锡公司发布了关于大数据的调研报告《大数据:下一个前沿,竞争力、创新力和生产力》,指出了大数据研究的地位以及将给社会带来的价值。2012年3月,美国奥巴马政府宣布投资2亿美元启动“大数据研发计划”,旨在提高和改进从海量和复杂数据中获取知识的能力,加速美国在科学和工程领域发明的步伐,增强国家安全。这是继1993年美国宣布“信息高速公路”计划后的又一次重大科技发展部署,由美国国家科学基金会、能源部等6个联邦部门共同投资。中国科学院计算技术研究所李国杰院士指出^[3]:“美国政府认为大数据是未来的新石油,将大数据研究上升为国家意志,对未来的科技与经

济发展必将带来深远影响。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分和企业间新的争夺焦点。”

维基百科对大数据的定义是:“大数据是由于规模、复杂性、实时而导致的使之无法在一定时间内用常规软件工具对其进行获取、存贮、搜索、分享、分析、可视化的数据集”。互联网数据中心将大数据定义为:为更经济地从高频率的、大容量的、不同结构和类型的数据中获取价值而设计的新一代架构和技术。

大数据和以往的海量数据、超大规模数据有什么区别呢?显然这些术语都表示系统需要管理的数据规模很大。相对于当时的CPU和存储技术水平而言,这些规模过大的数据在处理时需要特别对待。从历史发展来看,超大规模在提出时表示的是GB级别的数据,海量数据提出时表示的是TB级别数据,而大数据则是指PB(1 015)及以上级别的数据。

PB甚至更高级别的大数据的出现是近年来移动通信、互联网、传感器、物联网等技术发展和应用的结果。据IDC公司统计,2011年全球被

创建和被复制的数据总量为1.8 ZB(1 021),其中75%来自于个人(主要是图片、视频和音乐),远远超过人类有史以来所有印刷材料的数据总量(200 PB)。谷歌公司通过大规模集群和MapReduce软件,每个月处理的数据量超过400 PB;百度每天大约要处理几十PB数据;Facebook注册用户超过10亿,每月上传的照片超过10亿张,每天生成300 TB以上的日志数据;淘宝网会员超过3.7亿,在线商品超过8.8亿,每天交易数千万笔,产生约10—20 TB数据;雅虎的总存储容量超过100 PB^[3]。图灵奖获得者吉姆·格雷和IDC公司预测,全球数据量每18个月翻一翻,未来10年全球大数据将增加50倍左右。

大数据对于企业来说意味着巨大的经济效益。2009年,谷歌公司通过大数据业务对美国经济的贡献为540亿美元;eBay通过数据挖掘精确计算出广告中的每一个关键字,2007年以来eBay产品销售的广告费降低了99%,而顶级卖家占总销售额的百分比却上升至32%。

另一方面,大数据对IT业也意味着对海量、分散、变化、异构特性数据

进行分析和管理的技術挑战。IBM、Oracle、微软、谷歌、亚马逊、Facebook等都是大数据处理技术的主要推动者。大数据带来的技术挑战涉及数据的收集、存储、检索、共享、分析以及可视化等各个方面。首先,存储能力的增长已经远远赶不上数据的增长,设计更合理、高可扩展性的分层存储架构是数据管理系统的首要任务。数据移动已是数据管理系统最大开销,数据管理系统需要从数据围着处理器转改为处理能力围着数据转。除了数据的采集、数据存储外,新的数据表示方法、非结构化数据的存储和分析、数据的去冗余和高效存储、海量动态数据的实时数据挖掘甚至大数据管理带来的能源消耗都将成为大数据时代的亟待解决的技术挑战。

2 大数据的特点

和很多新出现的概念或技术一样,关于大数据的特点也有很多种不同说法。百度百科给出的大数据的特点是4个“V”,分别代表:数量巨大(Volume),类型繁多(Variety),价值高(Value),处理速度快(Velocity)。但作者更倾向于Forrester分析师布赖恩·霍普金斯和鲍里斯·埃韦尔松在《首席信息官,请用大数据扩展数字视野》报告中给出的大数据的4个特点,分别是:海量(Volume)、多样性(Variety)、高速(Velocity)和易变性(Variability)。

(1)海量。IDC给出了一个估算:2011年全球数据总量大约为1.8 ZB,如果用9 GB的DVD盘来保存,那么叠加起来这些DVD的高度超过260 000 km,大约是地球到月球距离的2/3;如果用1 TB的2.5寸硬盘保存,那么叠加起来的高度将会超过17 000 km,接近地球周长的一半。据IDC最近的报告预测,到2020年,全球数据量将扩大50倍。大数据的规模尚是一个不断变化的指标,单一数据集的规模范围从几十TB到数PB

不等。此外,各种意想不到的来源都能产生数据。例如,从巴塞罗那至沙特首府利雅得的单程航行中,一架商用喷气飞机上收集的传感器数据量将超过1 PB,当用一次飞行的数据量乘以每天所有飞行的航班数,数据总量也将非常惊人。

(2)多样性。数据多样性的增加主要是由于新型多结构数据。以及包括网络日志、社交媒体、互联网搜索、手机通话记录及传感器网络等数据类型造成。其中,部分传感器安装在火车、汽车和飞机上,每个传感器都增加了数据的多样性。

(3)高速。高速描述的是数据分析和处理的速度。在网络时代,通过基于实现软件性能优化的高速电脑处理器和服务器的,创建实时数据流已成为流行趋势。企业不仅需要了解企业如何快速创建数据,还必须知道如何快速处理、分析并返回给用户,以满足他们的实时需求。根据IMS Research研究机构关于数据创建速度的调查,通过跟踪物联网设备的激活量,发现联网设备增长的第二波浪潮正在加速到来。本轮增长后,将涌现更多新型可联网设备增长的浪潮。据预测,到2020年全球将拥有220亿部互联网连接设备。

(4)易变性。大数据具有多层结构,这意味着大数据会呈现出多变的形式和类型。相较传统的业务数据,大数据存在不规则和模糊不清的特性,造成很难甚至无法使用传统的应用软件进行分析。传统业务数据随时间演变已拥有标准的格式,能够被标准的商务智能软件识别。目前,企业面临的挑战是处理并从各种形式呈现的复杂数据中挖掘价值。

3 大数据的应用

2012年被称为大数据元年,因为在这一年大数据这个概念引起了人们的空前关注。首先是美国政府公布“大数据研发计划”,紧接着世界各国以及各大商业公司也对大数据给

予了极大的关注。中国的计算机学会、电子学会等学术机构以及淘宝、中兴通讯等企业也给予了积极响应。其实,对大数据相关的技术研究和应用一直在进行,2012年突然迸发,只是一个量变到质变的结果。下面将简单地介绍全球重要的企业、机构有关大数据的研究、开发、应用的一些情况。

3.1 Google

众所周知,Google所提出的GFS、BigTable、MapReduce技术奠定了云计算研究和应用的基础。正如很多成功的技术一样,Google提出这些技术是为了解决其业务提供中遇到的现实问题。这个问题用今天的眼光来看就是大数据问题。Google作为搜索领域先进技术的实践者,其面对的现实一方面是海量的网页数据,另一方面是海量的网页数据分布在全世界200多个地方,总计超过100万台服务器上,而且这些数据和服务器的数量还在快速增长。GFS是Google提出的分布式文件系统,可以支持对分布在大量廉价硬件的数据进行有效可靠的访问。BigTable是Google构建在GFS之上的一种压缩高效的专属数据库系统。MapReduce是支持在大规模集群上的大数据进行并行计算的软件框架。基于这3项针对大数据存储访问和计算的关键技术,Google可以进行海量数据的搜索和分析挖掘,保证了其在搜索领域的主导地位。

Google在公布了GFS、BigTable和MapReduce技术后,Apache软件基金会以其为基础用Java开发了开源软件框架Hadoop,该框架现在是云计算相关研究和应用的基础。因为Hadoop是用来进行“批处理”的平台,一个任务一般需要几分钟来完成。针对Hadoop时间延迟的问题,Google提出了可以实现在海量网页文档集或者数字图书馆进行快速查询的Dremel技术^[4],该技术既有传统结构

化查询语言(SQL)的易用风格,又可以极快地处理比如查询PB级别的数据。基于Dremel技术,Google从2011年底开始,向公众发布了它的大数据服务“BigQuery”,其目的是为了销售云端的数据存储,以及分析软件;BigQuery使用了UI和REST界面,该业务的应用意味着数据分析门槛的降低。

3.2 IBM

IBM在数据分析与挖掘领域的聚焦由来已久。从2005年开始,IBM投资160亿美元进行了30次与大数据有关的收购,包括2005年收购拥有ETL数据集成工具DataStage的Ascential软件公司,2007年收购商业智能(BI)领域重量级公司Cognos,2009年收购美国三大统计分析软件之一的统计产品与服务解决方案(SPSS),2010年收购大规模并行处理数据仓库厂商Netezza等。IBM现在是全球数学博士的最大雇主,数学家正在通过IBM的数据分析产品研发和解决方案将其数学分析的才能应用于石油勘探、医疗健康等各领域,提升各行业的智能水平和决策的科学性。

近来IBM在数据分析领域中的一个为人所称道的成功案例,是其拥有学习能力的超级电脑沃森,它在2011年美国的智力竞猜节目《危险边缘》中击败人类。在IBM技术创新全球副总裁Bernard S. Meyerson看来,沃森代表着一个大数据时代的一种全新的计算模式。他说:“未来沃森可以自主学习,如果答案是错的,它会改变思维方式,下次给出正确答案”,这显然与我们现在应用的给定相同输入就一定给出相同输出的电脑程序是不同的。

IBM大数据平台的几个核心能力包括基于Hadoop的云存储、流计算、数据仓库等。IBM在Hadoop系统领域的代表产品是InfoSphere BigInsights,IBM将其在数据管理上的

丰富经验与Hadoop开源平台高效整合,成为最主要的静态大数据分析工具和平台。IBM在流计算领域的代表产品是InfoSphere Streams,它是流数据处理技术产品,不仅能够对诸如气象信息、通信信息、金融交易数据的管理中动态捕捉信息并进行实时分析,还能够对静态数据的处理提供有效补充。IBM在数据仓库领域的代表产品是在线交易型数据仓库InfoSphere Warehouse和分析型数据仓库Netezza。Netezza克服了传统数据仓库在面临大数据挑战时的瓶颈,充分发挥了对称多处理能力,可以将大量数据整合到统一的平台上,计算能力高达TB级。通过与Cognos和SPSS等业务分析工具相结合,IBM数据仓库产品还能够实现定制化的分析挖掘功能。

3.3 Microsoft

微软在数据管理、商务智能、数据挖掘的研发和解决方案是以其结构化查询语言(SQL) Server平台为基础的,对大数据的布局也是以SQL Server平台为主,并集成Windows Azure公有云与Hadoop系统,形成覆盖整个产业链的完整解决方案。微软已发布了SQL Server平台的2012版本,其中加入了大数据处理和分析挖掘的功能。这些特性包括:能够处理结构化数据以及非结构化数据;提出了数据商店的概念;将SQL Server的活动目录与Hadoop集成。目前微软已有的大数据实施成功的案例,包括目前正在成都投入运作的云计算中心。该中心利用大数据平台、虚拟化、BI商业智能分析等一系列技术手段,实现了对肉类产品从喂养到售卖的实时监控。

3.4 淘宝

随着电子商务的迅速发展,淘宝所积累的庞大数据、所面对的大量复杂用户需求,客观要求采用大数据技术进行分析和处理,这主要包括在线

分析和离线分析两种。在线分析对相应时间的要求比较高(通常不超过若干秒),通过构建在云计算平台上的NoSQL系统(例如Hadoop上的HBase),实现了更好地开源、降低成本、易于扩展等效果,而且能够实时处理数千万甚至数亿条请求记录。离线数据分析基于开源的Hadoop的HDFS文件系统和MapReduce运算框架,用于较复杂和耗时的数据分析和处理。

采用传统市场调查方式(电话、邮件、信函等)抽样调查耗时耗力,且调查结果与客观情况误差较大,淘宝通过对实际访问、交易的真实数据分析可以发现一些有趣结果,利用它们可以帮助商家调整营销战略,提升竞争力。让我们来分析一个商品之间常常存在的内在关联实例^[5],比如买了奶粉的客户,很可能会买奶嘴等婴儿用品。过去人们更多依靠逻辑分析和抽样统计来发现这些关联关系,现在凭借大数据及其分析处理系统,可以更加清晰和准确地获取商品之间的内在关联。比如,购买了女装的客户,买女士内衣、箱包皮具和食品的比例最大;其次是买彩妆和女鞋;再次是服饰配件和饰品等,这是非常典型的女性消费者购物模式。这些信息可以有多种用途,例如商家在决定扩大或缩小经营范围时,可以借此来选择扩大或缩小商品的类别;搞促销活动时,商城运营人员可以借此选择促销的范围乃至不同商品的促销力度等。

4 大数据与云计算的关系

大数据和云计算是关系紧密的两个概念。大数据技术广义来讲涵盖了从数据的海量存储、处理到应用多方面的技术,包括海量分布式文件系统、并行计算框架、NoSQL数据库、实时流数据处理以及智能分析技术如模式识别、自然语言理解、应用知识库等。狭义来讲则主要指从大量、多样、分散和异构的数据集中提取有

用信息的核心技术,包括实时流数据处理以及智能分析技术如模式识别、自然语言理解、应用知识库等。

云计算之所以一经提出就得到广泛关注,是因为它使得人类“将计算能力作为公共事业设施来提供”的梦想变为现实,而使得“梦想照进现实”的关键技术是 GFS、BigTable 和 MapReduce。这 3 项技术是 Google 为了巩固其搜索领域的核心地位而提出的。Google 提出将文件和数据分割成块,以便支持分布式存储和并行处理,实现海量数据存储并提升大数据量下的快速数据处理^[6]。因此,云计算的核心是业务模式,本质是数据处理技术。

可以看出,云计算技术是广义大数据技术的一部分,也是狭义大数据技术的基础。可以说,大数据是资产,云为数据资产提供了保管、访问

的场所和渠道。如何盘活数据资产,使其为国家治理、企业决策乃至个人生活服务,是大数据研究的核心问题。一方面,大数据离不开云计算,正因为有了云计算的超强计算能力,大数据才显示出了堪比黄金钻石的价值。另一方面,大数据处理的兴起也将改变云计算的发展方向,云计算正在进入以分析即服务(AaaS)为主要标志的 Cloud 2.0 时代。(待续)

参考文献

- [1] Big data: The new natural resource[EB/OL]. <http://www.ibmbigdatahub.com/infographic/big-data-new-natural-resource>
- [2] Big data: Science in the petabyte era[J]. Nature, 2008, 455: 1-136.
- [3] 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.
- [4] MELNIK S, GUBAREV A, LONG J J, et al. Dremel: Interactive Analysis of Web-Scale Datasets[C]// Proceedings of the 36th International Conference on Very Large Data Bases (VLDB'10), Sep 13-17, 2010, Singapore. 2010: 330-339.

- [5] 阳振坤, 张清, 王勇, 等. 大数据的魔力[J]. 中国计算机学会通讯, 2012, 8(6): 17-21.
- [6] 王柏, 徐六通. 云计算[J]. 中兴通讯技术, 2010, 16(1): 57-60.

收稿日期: 2012-12-03

作者简介



于艳华, 北京邮电大学计算机学院副教授; 主要研究方向为网络管理与优化、数据挖掘等; 已发表论文 10 余篇, 申请专利 10 余项。



宋美娜, 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。

上接第 56 页

来。这样 Reducer 里就可以利用这些样本重新估计出 k 个聚类中心, 如(2)所示:

$$O_i = \frac{\sum_{j=1}^n O_{ij}}{n} \quad (2)$$

这样, 在一轮 MapReduce 完成后, 新的聚类中心也已经计算出来。通过比较本轮聚类中心与上一轮聚类中心差异度, 可确定算法是否收敛。

4 结束语

文章通过对数据挖掘和云计算技术的发展分析, 提出了基于云计算的数据挖掘平台架构以及数据挖掘服务化的思路。本平台不仅仅是基于云计算实现了一个数据挖掘平台, 同时也对数据挖掘平台进行了 SaaS 化。本平台可以为运营商、企业提供效益增值的数据挖掘应用, 同时也减少了运营商、企业在数据挖掘技术上的投入。运营商、企业即可以创建自己内部的数据挖掘私有云, 为内部产品提供数据挖掘服务, 也可以提供数据挖掘公用云, 为不同的企业提供数

据挖掘服务。

参考文献

- [1] 云时代企业数据挖掘面临的挑战(1)[EB/OL]. <http://cloud.watchstor.com/storage-134538-1.htm>
- [2] 陈康, 郑邦民. 云计算: 系统实例与研究现状[J]. 软件学报, 2009, 20(5): 1337-1348.
- [3] 纪俊. 一种基于云计算的数据挖掘平台架构设计与实现[D]. 青岛: 青岛大学, 2009.
- [4] J Han, M Kamber. Data mining concepts and techniques[M]. Third Edition. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2012.
- [5] 邵峰晶, 于忠清. 数据挖掘原理与算法[M]. 北京: 科学出版社, 2009.
- [6] 商琳, 骆斌. 一种基于数据仓库的数据挖掘系统的结构框架[J]. 计算机应用研究, 2000, 17(9): 63-65.
- [7] 杨帆友, 唐彦. 云计算总体架构及其应用与商业模式探讨[J]. 数字通信, 2012, (3): 3-6.
- [8] 何清. 基于云计算的海量数据挖掘[IC/OL]//第二届中国云计算大会, 2010年5月21-22日, 北京. http://blog.sina.com.cn/s/blog_66248a9e0100z38d.html
- [9] 杨勇, 董振江, 陆平. 具备云计算特性的业务交付平台及其关键技术研究[J]. 中兴通讯技术, 2011, 17(5): 55-57.
- [10] 吴朱华. 云计算核心技术剖析[M]. 北京: 人民邮电出版社, 2011.
- [11] 刘鹏. 云计算[M]. 北京: 电子工业出版社, 2011.
- [12] 夏英, 杨选伦. 云环境中基于金字塔模型的影像数据存储方法[J]. 重庆邮电大学学报(自然科学版), 2012, 24(6): 669-674.
- [13] 余永红, 向晓军, 高阳等. 面向服务的云数据挖掘引擎的研究[J]. 计算机科学与探索, 2012, 6(1), 46-57.
- [14] 李智龙, 宿绍莹, 唐鹏飞, 陈曾平. 基于数字信道化的正弦信号快速测频方法[J]. 雷达科学与

技术, 2011, 9(5): 55-58.

收稿日期: 2012-10-28

作者简介



丁岩, 南京邮电大学硕士学位毕业; 中兴通讯业务研究院业务平台系统部部长; 先后从事 BOSS 系统、核心网网管、SDP、Appstore、MISP 等产品和平台的总体架构设计和研发工作, 研究方向为 SDP、移动互联网、云计算等; 曾获深圳市科技创新奖, 并申请多个专利。



杨庆平, 哈尔滨工业大学硕士毕业; 中兴通讯业务研究院系统工程师; 研究方向为人工智能和数据挖掘, 主要从事数据挖掘技术规划、架构设计、需求分析等



钱煜明, 中兴通讯业务研究院总工、中科院客座研究员、江苏省双创人才; 对移动互联网、云计算及服务计算、大数据分析处理方面有较深入研究。

大数据

作者: [于艳华](#), [宋美娜](#)
作者单位: [北京邮电大学计算机学院, 北京, 100876](#)
刊名: [中兴通讯技术](#) 
英文刊名: [ZTE Communications](#)
年, 卷(期): 2013, 19(1)
被引用次数: 1次

引用本文格式: [于艳华](#). [宋美娜](#) [大数据](#) [期刊论文] - [中兴通讯技术](#) 2013(1)