空间大数据信息基础设施

Big Spatial Data Infrastructure

吴朝晖 陈华钧 杨建华 著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

空间大数据信息基础设施

吴朝晖　陈华钧　杨建华　著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目(CIP)数据

空间大数据信息基础设施 / 吴朝晖，陈华钧，杨建华著. — 杭州：浙江大学出版社，2013.1
ISBN 978-7-308-11016-7

Ⅰ. ①空… Ⅱ. ①吴… ②陈… ③杨… Ⅲ. ①空间信息技术—基础设施—研究 Ⅳ. ①P208

中国版本图书馆 CIP 数据核字(2013)第 006655 号

空间大数据信息基础设施
吴朝晖　陈华钧　杨建华　著

责任编辑　陈静毅　黄娟琴
封面设计　续设计
出版发行　浙江大学出版社
（杭州天目山路 148 号　邮政编码 310007）
（网址：http://www.zjupress.com）
排　　版　杭州星云光电图文制作工作室
印　　刷　浙江印刷集团有限公司
开　　本　787mm×1092mm　1/16
印　　张　9
字　　数　162 千
版 印 次　2013 年 1 月第 1 版　2013 年 1 月第 1 次印刷
书　　号　ISBN 978-7-308-11016-7
定　　价　42.00 元

前　言

大数据计算是指规模在PB级(10^{15})-EB级(10^{18})-ZB级(10^{21})的极大规模数据处理。从相对概念上来说,大数据计算又指传统文件系统、关系数据库、并行处理等技术无法有效处理的极大规模数据计算或极限级计算(Extreme-scale Computing)。它也指相对于设备的计算能力而言足够大数据量的计算,如移动设备上的TB级数据处理、内存的TB级数据处理等。早在2003年,Google研究人员Ghemawat S,Gobioff H和Leung S T在国际操作系统设计与实现会议上发表论文,提出谷歌文件系统(Google File System, GFS),这其实是信息产业界最早对大数据的研究和应用。2011年2月,《科学》杂志围绕大数据刊登了专题,指出大数据的收集、维护和使用已经成为科学研究的重要工作。2011年9月,高德纳咨询公司将其列为2012年全世界十大技术趋势之一。2012年2月,大数据计算被《华尔街日报》认为是改变21世纪的三个重大突破之一,就像改变20世纪的电话和电力。

空间信息基础设施是指为获取和存储空间信息,对空间信息进行处理与分析并将结果分发给不同用户,集成融合不同空间信息系统及实现互操作与共享,对空间信息进行管理维护与组织协调,最终将结果显示给用户的计算机基础设施体系。空间数据处理是典型的大数据计算应用。空间大数据包括各种遥感图像、传感器观测数据、元数据以及原始数据的合成数据等。作为典型的特定领域内的大数据,它具有大数据的主要特点:数据规模极大,数据间关联性复杂,类型多样化,时效性高,分析全面深入。同时,空间大数据还具有空间信息领域内的特点,比如在深度数据分析上,需要同时在空间和时间两个维度发掘数据之间的关联,进行时间轴和空间轴上的数据预测等。

随着大数据时代的到来,传统的空间信息基础设施已经逐渐暴

露弊端，越来越不能满足当前海量数据处理规模的需求。所以，探索一种能够承载海量空间大数据处理业务的新架构成为空间信息大数据时代的必由之路。另外，在大数据的时代背景下，传统的空间信息基础平台需要逐步向存储海量化、处理规模化、功能开放化、管理集中化和客户端轻量化等方向转化。为了实现这些转化，传统的空间信息基础设施架构必须进行相应的改变。

本书结合空间信息的典型特征，探讨发展新一代面向大数据的空间信息基础设施体系架构的必要性，并从空间数据表达模型、空间数据存储模型、空间数据计算模型、空间数据服务模型、空间信息应用开发模型和空间信息应用服务模型六个维度阐述新一代基础设施体系架构应该具有的特性和价值。

本书的组织结构如下：第 1 章介绍大数据计算的基本概念和主要发展现状；第 2 章结合大数据处理的最新进展，介绍大数据处理的主要关键技术；第 3 章介绍传统空间信息基础设施的不同发展阶段和主要发展趋势；第 4 章结合大数据的特征，探讨面向空间大数据的新一代空间信息基础设施体系架构；第 5 章介绍空间大数据处理的若干典型技术；第 6 章结合综合防灾减灾介绍空间大数据处理的主要应用前景。

本书是浙江大学高分辨率对地观测工程中心的研究人员多年的共同努力成果，以下人员为本书的撰写和审校作出过贡献，在此一并表示感谢，他们是：陈矫彦、张宁豫、曾明宇、陶金火、胡磊、黄梅龙、高啸、陈曦、陈云路等。

作者

2012 年 12 月于浙江大学求是园

目　录

第 1 章

大数据计算概述

本章重点对大数据(Big Data)计算进行概述,主要从三个方面展开。首先,概述大数据计算的发展历史、内涵定义和主要特征。大数据的五大特征分别是:极大数据规模、复杂数据关联、整体深度数据分析、类型多样化和高度时效性。其次,介绍大数据计算的发展现状,分析它的三个重要发展方向:分布式数据集群、高性能并行计算机和一体化大数据分析平台。最后,分析大数据计算在地球观测、强子对撞机数据处理、生物信息、医学影像、网络日志和商业智能等多个领域的重要应用。

1.1 大数据计算简介

1.1.1 发展历史

在过去十几年,随着大规模数据指数级增长,特别是在互联网和传感器网络领域,现有的数据处理技术已经严重限制信息的最大化利用。所以,关于大数据处理的学术讨论和研究得到日益重视。同时,无论是在工业界,还是在学术界,对大数据的研究和应用都取得了不少重要的成果。下面分阶段介绍大数据在信息产业技术和学术研究上的重要发展历程和成果。

总体而言,大数据计算的发展历史可以分为以下三个阶段。

1. 以计算为中心的大数据处理时代

早期的数据处理主要集中在大规模的数据计算,包括多核高性能计算

机、计算网格等，而数据的存储主要采用大规模的冗余磁盘阵列、存储网格等。这个时期互联网还没有兴起，海量数据并未在各个商业领域出现，主要集中在军事、科学计算、科学实验、特定制造等领域。但是这个时候的并行计算、网格计算等也在一定程度上影响了后来分布式集群和大数据计算的理论技术。

20 世纪 60 年代，超级计算机首先被提出。最早的超级计算机是 CDE 的 Cray S 设计的。到 20 世纪 70 年代，超级计算机使用为数不多的处理器。而从 20 世纪 90 年代到 20 世纪末，成千上万的处理器开始在超级计算机中应用，海量并行的超级计算机也开始出现。下面列出一些超级计算机发展史上的重要事件。

1966 年，世界上第一台超级计算机 CDC6600 诞生。

1976 年，Cray1 诞生，它后来成为世界上最成功的超级计算机之一。

1985 年，Cray2 诞生，它具有 8 个处理器，每秒能进行 1.9 亿次计算，一直到 20 世纪 90 年代，它都是世界上最快的计算机。

1994 年，被称为 Numerical Wind Tunnel 的超级计算机诞生，它包含 166 个矢量处理器，而每一个这样的处理器每秒可进行 1.7 亿次运算。

1996 年，具有 2048 个处理器的 Hitachi SR2201 超级计算机诞生，它处于巅峰时每秒可执行 600 亿次计算。

2003 年，蓝色基因的原型机诞生。IBM 蓝色基因超级计算系统对世界超级计算领域产生了深远的影响。它凭借空前的可持续计算性能，以每秒钟 280.6 万亿次浮点运算速度夺得 TOP 500 超级计算排名冠军。

2004 年，美国能源部在 ASCI 计划中，使用超级计算机真实地模拟了核爆炸，这也是超级计算在应用上的一个重要里程碑。

2010 年 10 月，我国天河-1A 安装完毕，以每秒 2.5 千万亿次运算速度夺得全球第一，比第二名的美国国家实验室的计算机快 30%。

2011 年 6 月，日本超级计算机“京”(K Computer)以每秒 8162 万亿次运算速度成为全球最快的超级计算机。

在超级计算机快速发展的同时，网格计算也得到了长足的进步，下面列出网格计算发展史上的重大时间点。

(1)1990 年至 1995 年，出现了以高性能网络把超级计算机节点连接起来，为高性能应用提供计算机资源的想法和需求。这种需求逐渐孕育出网格计算的概念。FAFNER[1] 及其后继 SETI[2] 和 I-WAY[3] 项目是当时典型的网格计算项目代表。

(2)1995 年至 2000 年，Globus[4] 等网格平台中间件被提出来。这些工具

集为网格计算的发展提供了强大的推动力，而 Globus 甚至成为网格计算的事实标准。

(3)2000 年至今，开放的网格服务体系结构(OGSA)[5]被提出。OGSA 是网格计算走向成熟的重要标志，网格计算的研究、开发和应用也大量出现。并且随着互联网的出现，网格计算和互联网逐渐走向融合。

2. 以存储为中心的大数据处理时代

这个时代的特点是利用 Hadoop 分布式集群来处理半结构化数据和非结构化数据。它更贴近互联网应用，突出商业价值，强调利用廉价的集群来处理海量的数据。这个阶段并没有真正提出大数据的概念，但基本奠定了当前大数据处理主流方法的理论和平台。

早在 2003 年，Google 研究人员 Ghemawat S，Gobioff H 和 Leung S T 在国际操作系统设计与实现会议上发表论文，提出谷歌文件系统(Google File System，GFS)[6]，这其实是信息产业界最早对半结构化和非结构化大数据的研究和应用。

2004 年，Google 发表论文《MapReduce：简化大规模集群上的数据处理》[7]，向全世界介绍他们的分布式计算框架，它针对半结构化和非结构化大数据的特点进行设计。需要指出的是，文中结构化数据是指存入数据库的有属性和类型的数据值；半结构化数据指有一些位置规律，但是没有类型没有数据关系的数据，比如服务器日志；而非结构化数据是没有位置规律、没有数据关联、没有数据类型的数据，比如计算机无法识别的人类语言、用户上传的图片视频等。事实上，后来的 Hadoop 就是使用 MapReduce 作为其并行计算框架，并且根据 GFS 设计了分布式文件系统(Hadoop Distributed File System，HDFS)。所以，Google 是这个大数据处理时代的开创者之一。

伴随着 GFS 和 MapReduce 的提出，信息界日渐形成了以这种分布式廉价集群为基础的大数据处理热潮。相对于传统的超级计算机、网格计算，以 GFS 和 MapReduce 为核心的集群在半结构化和非结构化大数据处理上具有非常明显的优势，并且迅速在产业界得到广泛应用。

2005 年年初，著名的开源项目 Apache Nutch 完成了 Hadoop 分布式集群框架的早期工作。开发人员在吸收了 GFS 的理念后，转向 NDFS 的研发，使得著名的开源搜索引擎 Nutch 的主要算法完成向 MapReduce 的移植。

2006 年开发人员将 NDFS 和 MapReduce 移出 Nutch，形成 Lucene 的一个子项目，也就是后来著名的 Hadoop 项目。与此同时，Hadoop 的创始人 Cutting D加入雅虎，在雅虎的支持下，Hadoop 逐渐成长为一个能够处理 Web

数据的系统。

到2008年，Hadoop已经在雅虎等单位完成上千个节点规模的集群构建，成功地进行了TB级数据的处理任务。2008年4月，Hadoop打破世界纪录，成为世界上最快的TB级数据排序系统。

随着以Hadoop为核心的开源项目的深入发展，Hadoop的大数据处理生态圈也逐渐形成。Hadoop所使用的分布式文件系统HDFS已经日渐成熟。2007年，分布式按列存储数据库HBase原型完成，2008年成为Apache下面的顶层项目，并且发布第一个稳定版本。2008年9月，运行在MapReduce和HDFS集群之上的数据流语言和运行环境"Pig"发布，用于检索非常大的数据结合。除此以外，分布式按列存储数据仓库Hive，分布式高可用协调服务ZooKeeper，数据库和HDFS之间的高效数据传输工具Sqoop等工具也陆续发布稳定版本。

在以Hadoop为核心的海量半结构化非结构化数据处理技术不断发展的同时，大数据的概念也逐渐被提出来，并且得到广泛的认可。世界权威的学术期刊*Nature*在2008年发表多篇论述大数据的文章。文章[8]根据当前的数据来源，在一定程度上预示了大数据时代的到来是必然的。文章[9]提出生物学领域大数据的概念，分析了生物学领域应对大数据挑战的方法。文章[10]表述了大数据时代给数据挖掘带来的挑战。

3. 以系统为中心的大数据分析机时代

Hadoop这样的分布式集群在互联网领域得到广泛应用，但也存在一些问题。作为一个开源软件，它需要一个技术团队维护集群，并且无法提供专业的服务。同时，当它要扩展到其他非互联网领域时，比如用户商业智能分析，也需要进行特定的开发，需要解决不少问题。因此，各大数据服务商开始以整个系统为中心，针对特定的领域提供Hadoop和硬件一体化设计的大数据分析机。同时，大数据计算的概念也开始被提出，并且得到多种解释和应用。

2010年，EMC收购了Greenplum——一家为企业级数据云(EDC)和商务智能(BI)提供解决方案和咨询服务的公司。2011年4月，EMC发布了EMC Greenplum大数据计算装置。2011年12月，EMC发布了EMC Greenplum统一分析平台(Unified Analytics Platform，UAP)。UAP集成了关系数据库、Hadoop等一系列技术，提供通用集成化的大数据分析平台。

2010年，甲骨文公司推出Oracle Exadata数据库云服务器，是一个由数据库软件、硬件服务器和存储设备组成的软件和硬件集成式系统，真正实现软硬件一体化。2011年10月，甲骨文公司在甲骨文全球大会上宣布推出Oracle大

数据机(Big Data Aplliance),并同步发布了 Oracle NoSQL 数据库。2012 年 2 月,甲骨文宣布 Exalytics In-Memory Machine 出货,Exalytics 基于内存数据库技术,具有实时在线高效大数据挖掘分析能力。

2011 年,微软发布 SQL Server R2 平行数据仓库(Parallel Data Warehouse, PDW)。PDW 采用非关系数据管理模型,与基于 Windows 的 Hadoop 云存储实现 SQL Azure Hadoop 服务,重点支持大数据计算,并支持在 Hadoop 上用 C# 编程。

2011 年 6 月,惠普发布了 Vertica 分析平台 5.0 新版本,让 IT 系统能够实时地分析超过 1PB 的数据。

2012 年 1 月,IBM 宣布推出一款全新的一体机分析工具——IBM Netezza 客户智能处理装置。这款分析利器集成 Hadoop 云存储、大数据内存计算等技术,可以帮助行业客户在数秒之间运行复杂的实时分析。

除了这些大数据一体化解决方案上的突破,大数据计算的概念也得到相应的发展。除了概念、定义等的明确提出,这些成果也向人们明确指出了大数据的方向和重要性。

2011 年 2 月出版的《科学》杂志刊登的专题[11],围绕目前各类数据的爆发式增长展开了讨论,并且得出结果:海量数据的收集、维护和使用已经成为科学研究的重要工作。

2011 年 5 月,麦肯锡咨询公司认为,大数据是未来创新力、竞争力和生产力的最前沿[12]。同时,著名的 Forrester 研究公司指出,大数据存在巨大的机遇[13]。

2011 年 9 月,高德纳咨询公司将其列为 2012 年全世界十大技术趋势之一[14]。

2012 年 2 月,大数据计算被《华尔街日报》认为是改变 21 世纪的三个重大突破之一,就像改变 20 世纪的电话和电力。

1.1.2　内涵定义

大数据计算的概念自从被提出以来,就有多种不同版本的定义。这些定义从不同的领域、不同的计算特点出发,对大数据进行了定性的描述。

无论是哪种大数据定义方式,一个业界公认的大数据描述就是数据规模的相对性。从绝对概念上来说,大数据计算是指规模在 PB 级(10^{15})- EB 级(10^{18})- ZB 级(10^{21})的极大规模数据处理。从相对概念上来说,大数据计算又指传统文件系统、关系数据库、并行处理等技术无法有效处理的极大规模数据

计算或极限级计算(Extreme-scale Computing)。它也指相对于设备的计算能力而言足够大数据量的计算,如移动设备上的TB级数据处理、内存的TB级数据处理等。

全球最大的战略咨询公司麦肯锡咨询公司对大数据的定义作了三个方面的概括:第一,强调是传统数据库技术和工具无法处理的;第二,强调"大"是一个相对概念;第三,数据规模"大"对于不同领域或工具软件来说可能差异很大。

IBM的大数据定义包括三部分内容:数据量大、数据类型多样化和时效性高。数据量大一方面指企业级数据规模达到TB级和PB级,另一方面也就是所谓的相对性概念。数据类型多样化指的是数据的复杂度越来越高,非结构化和半结构化数据的增长已经远远超出传统的结构化数据。为了从这些数据中挖掘出巨大的信息价值,需要不断丰富数据的类型以提高计算机对各种数据的处理能力,这也是大数据计算的一个必然趋势。数据时效性高包括两个方面:一方面,信息的价值在于最终服务于决策,而数据的时效性直接影响了信息价值的最大化;另一方面,传感器网络、在线服务得到空前的发展,也产生了对海量流数据处理的需求,而海量流数据处理事实上也是大数据处理高度时效性的一个重要体现。

维基百科对信息科学中大数据的定义是"一个巨大并且复杂的数据集合"[15]。它是一个相对概念,指相对现有的系统而言,它给数据的获取、建模、存储、搜索、共享、分析和虚拟化带来巨大的挑战。

《著云台》的分析师团队对大数据的定义更多停留在企业级数据智能。大数据通常用来形容一个公司创造的大量非结构化和半结构化数据,这些数据在传统的关系型数据库中分析处理时,会花费过多时间和金钱。而大数据技术就是从各种各样类型的数据中,快速获得有价值信息的能力。

相比于简单对大数据本身进行规模或者复杂度等的定义,英特尔在名为《技术商业:大数据,下一个黄金增长点》的文章中,从技术、商业的角度对整个大数据的生态系统进行了描述。文章指出,大数据包含三个内涵:第一,日渐膨胀的复杂数据;第二,一个正在快速发展的生态系统;第三,新技术、新技能、新实践和新的商业模式[16]。

综合分析各个企业、学术研究机构对大数据的定义和大数据本身的特点、应用情况,本书给出对大数据内涵的定义,也就是大数据的3-H定义。3-H分别指的是极大规模数据(Huge Data)、复杂关联数据(Hyper Data)、整体深度数据分析(Holistic Data)。

(1)大数据指的是极大规模数据。和一般大数据计算定义类似的是数据规模极大,并且这个数据规模是一个相对"大"的概念。而和一般大数据计算定义

不同的是，极大规模的概念还强调了对不同规模的大数据的处理是有可伸缩性的，能够实现跨多种规模（GB、TB、PB、EB、ZB级）的弹性数据处理。

（2）**大数据是复杂关联的数据**。数据之间不是孤立的，而是存在具有实际意义的语义关联。数据本身并非无意义的字节存储，而是机器能够识别，能够快速检索、展现、分析的语义化数据，它是作为一个知识体为人们提供服务。大数据计算需要实现跨领域、跨行业、跨组织的大数据语义互联，建立复杂的超数据关联。

（3）**大数据计算是整体深度的数据分析**。大数据计算需要实现跨时间、跨空间的大数据时空管理，在时空两个维度上建立大数据的整体系统性分析方法。同时，在时空每一维上，数据的挖掘需要综合考虑多个因素，进行深度的分析。

1.1.3　主要特征

通过多种大数据计算的定义，下面用五大特征来描述大数据计算。

1. 极大数据规模

进入21世纪以来，人们正面临的一个严峻考验是：在信息相关产业的任何一个领域，数据都正以指数级的速度增长。麦肯锡咨询公司首先提出了“大数据”的概念以及大数据时代的到来，并认为从海量数据中能挖掘具有巨大经济价值的信息[17,18]。国内互联网周刊，更直接撰文风趣地指出“从2012年开始，我们将从大陆时代，移民进入大数据时代”[19]。的确，在各个领域，数据都呈现出爆发式的增长趋势。文章选择了三个具有代表性的领域来展现数据增长的迅猛势头：互联网、企业级数据分析和传感器网络。

全球互联网自20世纪90年代进入商用以来迅速拓展，目前已经成为当今世界推动经济发展和社会进步的重要信息基础设施。经过短短十几年的发展，截至2007年1月，全球互联网已经覆盖五大洲的233个国家和地区，网民达到10.93亿人，用户普及率为16.6%。而根据2012年1月16日中国互联网络信息中心（CNNIC）在北京发布的《第29次中国互联网络发展状况统计报告》显示，截至2011年12月底，中国网民规模突破5亿人，达到5.13亿人，全年新增网民5580万人[20]。可以这样说，互联网直接连接了世界上多数的数据系统、个人数据中心，甚至部分传感器数据，并且它的规模还在成倍地增加。

目前，在互联网领域，21亿网民80亿互联网设备产生了大量的数据。图1-1用具体的数字展示了当前的数据规模。就在普通的一天里，Google产生了

20PB的数据流量，200万篇博客在互联网上发布，2940亿封电子邮件发出，1亿8700万小时的音乐在Pandora上播放，2200万小时的电影和电视节目在Netfix上播放，1288个新应用程序可供下载，超过3500万个应用程序被下载，5亿3200万条状态在社交网站上更新，2亿5000万张照片上传至Facebook。这些一天中产生的数据，足够说明互联网数据增长之快。从互联网企业来看，雅虎是一个典型的例子。雅虎的云计算高级副总裁Shugar S指出，雅虎每天为1000亿事件产生120TB数据输入，目前储存了70PB，而其最高存储容量是170PB。雅虎每天处理3PB数据，每个月在38000台服务器上运行超过百万个任务。

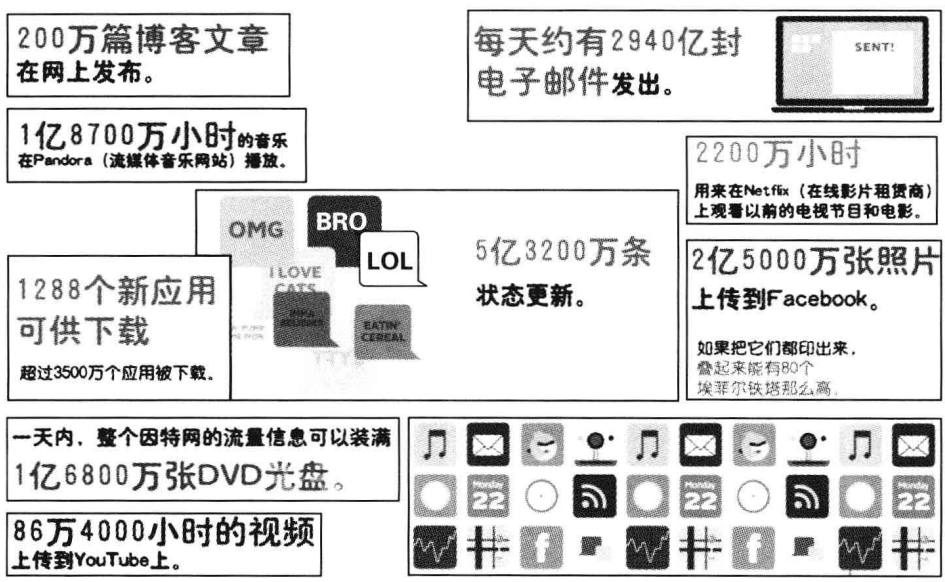

图1-1　典型互联网应用的数据规模

大数据分析和决策信息的获取已经成为企业提升经济价值的关键途径。麦肯锡在《你准备好迎接大数据时代了吗》中指出：在过去的几年，企业信息呈现爆发式增长。在美国17个经济部门中的15个部门，员工超过1000人的企业平均存储了235TB的数据，超出了美国国会图书馆的藏书。虽然大量信息来源于金融交易和客户互动，但从新设备和价值链各环节中产生的信息增长速度惊人。文章[21]进行了企业级用户调查，已有超过57.1%的企业数据量突破TB级别，其中18.1%的企业数据量已经超过10TB，企业数据管理系统已经逐渐告别GB，迈入TB时代。数据规模的概念是相对的，单个企业TB级数据