# 大数据

维基百科，自由的百科全书

**大数据**（英语：**Big data**[1][2][3]），又称为**巨量资料**，指的是传统数据处理应用软件不足以处理它们的大或复杂的数据集的术语[4][5]。大数据也可以定义为来自各种来源的大量非结构化和结构化数据。从学术角度而言，大数据的出现促成了广泛主题的新颖研究。这也导致了各种大数据统计方法的发展。大数据并没有抽样;它只是观察和追踪发生的事情。因此，大数据通常包含的数据大小超出了传统软件在可接受的时间内处理的能力。由于近期的技术进步，发布新数据的便捷性以及全球大多数政府对高透明度的要求，大数据分析在现代研究中越来越突出。[6] [7]

## 目录



## 概述

截至2012年，技术上可在合理时间内分析处理的数据集大小单位为艾字节（EB）[8]。在许多领域，由于数据集过度庞大，科学家经常在分析处理上遭遇限制和阻碍；这些领域包括气象学、基因组学[9]、神经网络体学、复杂的物理模拟[10]，以及生物和环境研究[11]。这样的限制也对网络搜索、金融与经济信息学造成影响。数据集大小增长的部分原因来自于信息持续从各种来源被广泛收集，这些来源包括搭载感测设备的移动设备、高空感测科技（遥感）、软件记录、相机、麦克风、无线射频辨识（RFID）和无线感测网络。自1980年代起，现代科技可存储数据的容量每40个月即增加一倍[12]；截至2012年，全世界每天产生2.5艾字节（$2.5\times10^{18}$字节）的数据[13]。

大数据几乎无法使用大多数的数据库管理系统处理，而必须使用“在数十、数百甚至数千台服务器上同时平行运行的软件”（计算机集群是其中一种常用方式）[14]。大数据的定义取决于持有数据组的机构之能力，以及其平常用来处理分析数据的软件之能力。“对某些组织来说，第一次面对数百GB的数据集可能让他们需要重新思考数据管理的选项。对于其他组织来说，数据集可能需要达到数十或数百TB才会对他们造成困扰。”[15]

随着大数据被越来越多的提及，有些人惊呼大数据时代已经到来了，2012年《纽约时报》的一篇专栏中写到，“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉。但是并不是所有人都对大数据感兴趣，有些人甚至认为这是商学院或咨询公司用来哗众取宠的buzzword，看起来很新颖，但只是把传统重新包装，之前在学术研究或者政策决策中也有海量数据的支撑，大数据并不是一件新兴事物。

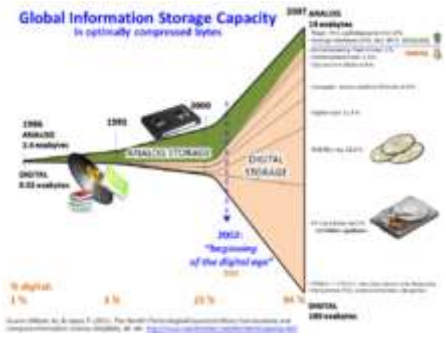
大数据时代的来临带来无数的机遇，但是与此同时个人或机构的隐私权也极有可能受到冲击，大数据包含各种个人信息数据，现有的隐私保护法律或政策无力解决这些新出现的问题。有人提出，大数据时代，个人是否拥有“被遗忘权”，被遗忘权即是否有权利要求数据商不保留自己的某些信息，大数据时代信息为某些互联网巨头所控制，但是数据商收集任何数据未必都获得用户的许可，其对数据的控制权不具有合法性。2014年5月13日欧盟法院就“被遗忘权”（right to be forgotten）一案作出裁定，判决谷歌应根据用户请求删除不完整的、无关紧要的、不相关的数据以保证数据不出现在搜索结果中。这说明在大数据时代，加强对用户个人权利的尊重才是时势所趋的潮流。

## 定义

IBM对维基百科的编辑纪录数据进行可视化的呈现。维基百科上总计数兆字节的文字和图片正是大数据的例子之一

大数据由巨型数据集组成，这些数据集大小常超出人类在可接受时间下的收集、废用、管理和处理能力[16]。大数据的大小经常改变，截至2012年，单一数据集的大小从数太字节（TB）至数十兆亿字节（PB）不等。

在一份2001年的研究与相关的演讲中[17]，麦塔集团（META Group，现为高德纳）分析员道格·莱尼（Doug Laney）指出数据增长的挑战和机遇有三个方向：量（Volume，数据大小）、速（Velocity，数据输入输出的速度）与多变（Variety，多样性），合称“3V”或“3Vs”。高德纳与现在大部分大数据产业中的公司，都继续使用3V来描述大数据[18]。高德纳于2012年修改对大数据的定义：“大数据是大量、高速、及/或多变的信息资产，它需要新型的处理方式去促成更强的决策能力、洞察力与最优化处理[原文 1][19]。”另外，有机构在3V之外定义第4个V：真实性（Veracity）为第四特点[20]。




全球信息存储容量成长图

大数据必须借由计算机对数据进行统计、比对、解析方能得出客观结果。美国在2012年就开始着手大数据，奥巴马更在同年投入2亿美金在大数据的开发中，更强调大数据会是之后的未来石油。

数据挖掘（data mining）则是在探讨用以解析大数据的方法。

## 应用示例

大数据的应用示例包括大科学、RFID、感测设备网络、天文学、大气学、交通运输、基因组学、生物学、大社会数据分析[21]、互联网文件处理、制作互联网搜索引擎索引、通信记录明细、军事侦查、金融大数据，医疗大数据，社交网络、通勤时间预测、医疗记录、照片图像和视频封存、大规模的电子商务等[22]。

### 巨大科学

应用于运动界

大型强子对撞机中有1亿5000万个感测器，每秒发送4000万次的数据。实验中每秒产生将近6亿次的对撞，在过滤去除99.999%的撞击数据后，得到约100次的有用撞击数据[23][24][25]。

将撞击结果数据过滤处理后仅记录0.001%的有用数据，全部四个对撞机的数据量复制前每年产生25拍字节（PB），复制后为200拍字节。

如果将所有实验中的数据在不过滤的情况下全部记录，数据量将会变得过度庞大且极难处理。每年数据量在复制前将会达到1.5亿拍字节，等于每天有近500艾字节（EB）的数据量。这个数字代表每天实验将产生相当于500垓

（$5\times10^{20}$）字节的数据，是全世界所有数据源总和的200倍。

### 科学研究

### 卫生学

国际卫生学教授汉斯·罗斯林使用“Trendalyzer”工具软件呈现两百多年以来全球人类的人口统计数据，跟其他数据交叉比对，例如收入、宗教、能源使用量等。

### 公共部门

目前，发达国家的政府部门开始推广大数据的应用。2012年奥巴马政府投资近两亿美元开始推行《大数据的研究与发展计划》，本计划涉及美国国防部、美国卫生与公共服务部门等多个联邦部门和机构，意在通过提高从大型复杂的的数据中提取知识的能力，进而加快科学和工程的开发，保障国家安全。

#### 信息审查

中国政府计划创建全面的个人信用评分体系，其包含不少对个人行为的评定，有关指标会影响到个人贷款、工作、签证等生活活动。高科技公司在被政治介入和指挥下为其目的服务，个人大部分行为和社交关系受掌控，几乎无人可免于监控[26]。除获取网络数据外，中国政府还希望从科技公司获得分类和分析信息的云计算能力，通过闭路电视、智能手机、政府数据库等搜集数据，以建造所谓的智能城市和安全城市。人权观察驻香港研究员王松莲指出，整个安全城市构想无非是一个庞大的监视项目[27]。

### 民间部门

- 亚马逊公司，在2005年的时点，这间公司是世界上最大的以LINUX为基础的三大数据库之一[28]。
- 沃尔玛可以在1小时内处理百万以上顾客的消费处理。相当于美国议会图书馆所藏的书籍之167倍的情报量[29]。
- Facebook，处理500亿枚的用户照片[30]。
- 全世界商业数据的数量，统计全部的企业全体、推计每1.2年会倍增[31]。
- 西雅图文德米尔不动产分析约1亿匿名GPS信号，提供购入新房子的客户从该地点使用交通工具(汽车、脚踏车等)至公司等地的通勤时间估计值[32]。
- 软银，每个月约处理10亿件（2014年3月现在）的手机LOG情报，并用其改善手机信号的信号强度[33]。
- 大企业对大数据技能需求量大，吸引了许多大学诸如伯克利大学开专门提供受过大数据训练的毕业者的大学部门。硅谷纽约为主《The Data Incubator》公司,2012年成立，焦点是数据科学与大数据企业培训，提供国际大数据培训服务。

### 社会学

大数据产生的背景离不开Facebook等社交网络的兴起，人们每天通过这种自媒体传播信息或者沟通交流，由此产生的信息被网络记录下来，社会学家可以在这些数据的基础上分析人类的行为模式、交往方式等。美国的涂尔干计划就是依据个人在社交网络上的数据分析其自杀倾向，该计划从美军退役士兵中拣选受试者，透过Facebook的行动app收集资料，并将用户的活动数据传送到一个医疗资料库。收集完成的数据会接受人工智能系统分析，接着利用预测程序来即时监视受测者是否出现一般认为具伤害性的行为。

## 市场

大数据的出现提升了对信息管理专家的需求，Software AG、甲骨文、IBM、微软、SAP、易安信、惠普和戴尔已在多间数据管理分析专门公司上花费超过150亿美元。在2010年，数据管理分析产业市值超过1,000亿美元，并以每年将近10%的速度成长，是整个软件产业成长速度的两倍[29]。

经济的开发成长促进了密集数据科技的使用。全世界共有约46亿的移动电话用户，并有10至20亿人链接互联网[29]。自1990年起至2005年间，全世界有超过10亿人进入中产阶级，收入的增加造成了识字率的提升，更进而带动信息量的成长。全世界通过电信网络交换信息的容量在1986年为281兆亿字节（PB），1993年为471兆亿字节，2000年时增长为2.2艾字节（EB），在2007年则为65艾字节[12]。根据预测，在2013年互联网每年的信息流量将会达到667艾字节[29]。

# 相关条目

- 数据挖掘
- 数据库
- 对象数据库
- 关系数据库
- 统计学
- 商务智能
- 分布式计算、分布式数据库、分布式文件系统、分布式运算环境
- 超级计算机
- 运筹学
- MapReduce
- 合成作战中心
- 工业大数据

# 注释

1. 原文：Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

# 参考文献


1. White, Tom. Hadoop: The Definitive Guide. O'Reilly Media. 2012-05-10: 3. ISBN 978-1-4493-3877-0.
2. MIKE2.0, Big Data Definition.
3. 巨量资料与进阶分析解决方案. 已忽略文本“ Microsoft Azure ” (帮助)
4. Kusnetzky, Dan. What is "Big Data?" . ZDNet. （原始内容存档于2010-02-21）.
5. Vance, Ashley. Start-Up Goes After Big Data With Hadoop Helper. New York Times Blog. 2010-04-22.
6. Li, Rita Yi Man. Have Housing Prices Gone with the Smelly Wind? Big Data Analysis on Landfill in Hong Kong, Sustainability 2018, 10(2), 341; doi:10.3390/su10020341. MDPI.
7. MIKE2.0, Big Data Definition.
8. Francis, Matthew. Future telescope array drives development of exabyte processing. 2012-04-02 [2012-10-24].
9. Community cleverness required. Nature. 4 September 2008, **455** (7209): 1. doi:10.1038/455001a.
10. Sandia sees data management challenges spiral. HPC Projects. 2009-08-04. （原始内容存档于2011-05-11）.
11. Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. Challenges and Opportunities of Open Data in Ecology. Science. 2011, **331** (6018): 703–5. doi:10.1126/science.1197962.
12. Hilbert & López 2011
13. IBM What is big data? — Bringing big data to the enterprise. www.ibm.com. [2013-08-26].
14. Jacobs, A. The Pathologies of Big Data. ACMQueue. 6 July 2009.
15. Magoulas, Roger; Lorica, Ben. Introduction to Big Data. Release 2.0 (Sebastopol CA: O'Reilly Media). 2009-02, (11).
16. Snijders, C., Matzat, U., & Reips, U.-D. (2012). ‘Big Data’ : Big gaps of knowledge in the field of Internet science. *International Journal of Internet Science, 7*, 1-5. http://www.ijis.net/ijis7_1/ijis7_1_editorial.html
17. Douglas, Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety (PDF). Gartner. [2001-02-06].
18. Beyer, Mark. Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. Gartner. [2011-07-13]. （原始内容存档于2011-07-10）.
19. Douglas, Laney. The Importance of 'Big Data': A Definition. Gartner. [21 June 2012].
20. What is Big Data?. Villanova University.
21. Erik Cambria; Dheeraj Rajagopal, Daniel Olsher, and Dipankar Das. 13. Big social data analysis. Taylor & Francis. 2013.
22. Hogan, M. What is Big Data. 2013-06-20 [2018-02-18]. （原始内容存档于2017-07-22）.

23. LHC Brochure, English version. A presentation of the largest and the most powerful particle accelerator in the world, the Large Hadron Collider (LHC), which started up in 2008. Its role, characteristics, technologies, etc. are explained for the general public.. CERN-Brochure-2010-006-Eng. LHC Brochure, English version. CERN. [20 January 2013].
24. LHC Guide, English version. A collection of facts and figures about the Large Hadron Collider (LHC) in the form of questions and answers.. CERN-Brochure-2008-001-Eng. LHC Guide, English version. CERN. [20 January 2013].
25. Brumfiel, Geoff. High-energy physics: Down the petabyte highway. Nature **469**. 19 January 2011: 282–83. doi:10.1038/469282a.
26. 陈迎竹. 慎防大数据助长独裁. 2017-10-15.
27. 华尔街日报：阿里、腾讯成为政府监视国民的耳目. 立场新闻. 2017-12-01.
28. Layton, Julia. Amazon Technology. Money.howstuffworks.com. [2013-03-05].
29. Data, data everywhere. The Economist. 2010-02-25 [2012-12-09].
30. Scaling Facebook to 500 Million Users and Beyond. Facebook.com. [2013-07-21].
31. eBay Study: How to Build Trust and Improve the Shopping Experience. Knowwpcarey.com. 2012-05-08 [2013-03-05]. （原始内容存档于2012-06-19）.
32. Wingfield, Nick. Predicting Commutes More Accurately for Would-Be Home Buyers - NYTimes.com. Bits.blogs.nytimes.com. 2013-03-12 [2013-07-21].
33. 柴山和久. ビッグデータを利益に変える方法. 幻冬舎. 2014. ISBN 978-4344952393 **（日语）**.

# 延伸阅读

- Big Data for Good (PDF). ODBMS.org. 2012-06-05 [2013-11-12].
- Hilbert, Martin; López, Priscila. The World's Technological Capacity to Store, Communicate, and Compute Information. Science. 2011, **332** (6025): 60–65. PMID 21310967. doi:10.1126/science.1200970.
- The Rise of Industrial Big Data. GE Intelligent Platforms. [2013-11-12].
- ISBN 978-986-320-191-5 《大數據》
- ISBN 978-986-241-673-0 《云时代的杀手级应用：Big Data大数据分析》
- IEEE Big Data Service. ODBMS.org. 2014-09-07 [2014-09-07].

# 外部链接

- 大数据的相关报导文章 (https://web.archive.org/web/20140227020937/http://www.wired.tw/events/big_data_magzine) （《Wired》中文网站）
- 处理大数据的挑战 (http://web.mit.edu/professional/onlinex-programs/courses/tackling_the_challenges_of_big_data.html)（美国麻省理工学院在线课程）

---