高等学校应用型本科创新人才培养计划指定教材
高等学校云计算与大数据专业“十三五”课改规划教材

云计算与大数据概论

青岛英谷教育科技股份有限公司　编著

西安电子科技大学出版社
http://www.xduph.com

高等学校应用型本科创新人才培养计划指定教材

高等学校云计算与大数据专业“十三五”课改规划教材

云计算与大数据概论

青岛英谷教育科技股份有限公司　编著

西安电子科技大学出版社

内 容 简 介

本书从云平台和大数据的概念出发，以当前国内外云平台和大数据的发展现状为背景，详细介绍了云计算技术、云计算平台、大数据技术、Hadoop 开发平台、MapReduce 应用、Pig 简介、HBase 简介、云计算与大数据安全等知识。

本书重点突出，偏重在宏观上讲解云平台和大数据的概念、用途、设计思想、应用原理等，为云平台和大数据专业的读者提供了理论基础。

本书适用面广，可作为本科大数据、计算机科学与技术、软件工程、计算机软件、计算机信息管理、统计学、数学等专业的教材，也可作为云计算和大数据从业者及爱好者的参考用书。

图书在版编目(CIP)数据

云计算与大数据概论 / 青岛英谷教育科技股份有限公司编著.
—西安：西安电子科技大学出版社，2017.8(2018.1 重印)
高等学校云计算与大数据专业“十三五”课改规划教材
ISBN 978-7-5606-4609-1

Ⅰ. ① 云… Ⅱ. ① 青… Ⅲ. ① 云计算 ② 数据处理 Ⅳ. ① TP393.027 ② TP274

中国版本图书馆 CIP 数据核字(2017)第 170508 号

策　　划　毛红兵
责任编辑　毛红兵
出版发行　西安电子科技大学出版社(西安市太白南路 2 号)
电　　话　(029)88242885　88201467　　邮　　编　710071
网　　址　www.xduph.com　　电子邮箱　xdupfxb001@163.com
经　　销　新华书店
印刷单位　陕西天意印务有限责任公司
版　　次　2017 年 8 月第 1 版　2018 年 1 月第 2 次印刷
开　　本　787 毫米×1092 毫米　1/16　印　张　14.5
字　　数　336 千字
印　　数　301～3300 册
定　　价　37.00 元

ISBN 978-7-5606-4609-1/TP

XDUP 4901001-2

高等学校云计算与大数据专业
“十三五”课改规划教材编委会

❖❖❖ 前　言 ❖❖❖

随着互联网、物联网、云计算等技术的快速发展，以及智能终端、网络社会、数字地球的普及和建设，全球数据量出现爆炸式增长，据 IDC 预计，到 2020 年全球数据量将增加 50 倍。毋庸置疑，云计算和大数据时代已经到来。

云计算和大数据的发展是相辅相成的。一方面，云计算为大数据提供存储和运算平台，并运用人工智能技术从海量的、多样化的数据中发现知识、规律和趋势，为决策提供信息参考；另一方面，大数据利用云计算的强大计算能力，可以提高数据分析的效率，从而更迅速地从海量数据中挖掘出有价值的信息，其不断增加的业务需求也拓展了云计算的应用领域。然而，云计算和大数据的发展也进一步增加了信息的开放程度，隐私数据及敏感信息的泄露事件亦随之时有发生。面对云计算与大数据产业的新特点和新挑战，如何保障数据安全也逐渐成为业界内外十分关注的重点课题。

目前，云计算已成为 IT 建设不可缺少的基础技术。国际领先的信息技术企业(如 Google、Amazon、微软、IBM 等)都构建了自己的云平台生态系统。中国的云计算市场经过技术和商业模式的积累，也已进入稳定发展阶段，云计算产业链、行业生态环境基本稳定，其中，阿里巴巴、百度、腾讯借助自身互联网业务的优势，逐步奠定了国内云计算市场的霸权地位；用友、金蝶等企业的软件云服务也逐渐成型；在国家政策的引导下，政府和企事业单位的公共云服务布局也在加快建设当中。

伴随着云计算技术的成熟，大数据也得到了日益广泛的应用。国际上，IBM、微软、Google、甲骨文、Amazon 等主要大数据服务提供商的大数据业务均呈稳步增长态势；在国内，阿里巴巴、腾讯、百度等企业的大数据业务也已经全面展开，为政府决策、日常出行、就餐购物、休闲娱乐等提供了诸多便利。

作为引领实体经济转型升级的重要突破口，云计算和大数据产业的发展已受到国家决策层的高度关注，国务院先后出台多个文件，支持和规范相关行业发展。2014 年 3 月，“大数据”首先进入《政府工作报告》；2015 年 1 月，国务院印发了《关于促进云计算创新发展培育信息产业新业态的意见》，以促进云计算创新发展，积极培育信息产业新业态；2015 年 6 月，国务院常务会议决定将在重点领域引入大数据监管系统；2015 年 7 月，国务院办公厅印发《关于运用大数据加强对市场主体服务和监管的若干意见》，进一步拓展大数据的应用领域；截至 2017 年 3 月，大数据已第四次进入《政府工作报告》，云计算和大数据产业的发展已然被提升到国家战略的高度，是具有广阔发展前景和旺盛人才需求的朝阳产业。

本书是面向高等院校云计算和大数据专业方向的标准化教材，兼顾完善的理论性和较强的实用性。前三章主要介绍了云计算技术，包括云计算的概念、云计算的核心技术、当前主流的云计算平台等内容；第 4 章至第 8 章主要介绍了大数据技术，包括大数据的概念、大数据的关键技术、当前的主流大数据服务和开源开发平台等内容；最后一章重点介

绍了云计算与大数据当前面临的主要安全问题及解决方案。希望读者能够通过对本书的学习，了解云计算和大数据的发展概况，掌握云计算技术及其体系架构，了解 Hadoop 等主流云计算平台，掌握大数据开发技术以及 MapReduce、Pig 和 HBase 等大数据分析工具的使用，并对云计算与大数据安全的标准和规范有一个基本的了解。

本书由青岛英谷教育科技股份有限公司编写，参与本书编写的人员有张伟洋、张杰、王万琦、马秀娟、侯方超、孟洁等。本书在编写期间得到了各合作院校的专家及一线教师的大力支持和协作，在此要特别感谢给予我们开发团队大力支持和帮助的领导及同事，感谢合作院校的师生给予我们的支持和鼓励，更要感谢开发团队每一位成员所付出的艰辛劳动。

书中难免有错误或不当之处，读者在阅读过程中如有发现，可以通过邮箱(yinggu@121ugrow.com)与我们联系，以期进一步完善。

编　者

2017 年 6 月

❖❖❖ 目　录 ❖❖❖

第1章　云计算与大数据概述

本章目标

- 掌握云计算的基本原理
- 掌握云计算的特点与优势
- 掌握大数据的特点
- 了解全球和国内云计算和大数据的发展状况
- 掌握云计算的分类
- 了解主流云服务供应商提供的云服务
- 掌握云计算和大数据的关系

目前，云计算和大数据时代已经到来，云计算已经普及并成为 IT 行业的主流技术。云计算的实质是由越来越大的计算量以及越来越多、越来越动态、越来越实时的数据需求催生出来的一种基础架构和商业模式。云计算时代，个人用户可以将文档、照片、视频、游戏存档记录上传至“云”中永久保存，企业客户根据自身需求，也可以搭建自己的“私有云”，或者托管、租用“公有云”上的 IT 资源与服务。

“大数据”在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业的存在已有时日，近年来，互联网和信息行业的发展令其越发引起人们的关注。最早提出“大数据”时代已经到来的是全球知名咨询公司麦肯锡，麦肯锡称：“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。”

本章旨在让读者更好地认识和了解云计算和大数据的基本知识，包括云计算和大数据的概念及发展状况、云计算的分类、主流云计算和大数据服务供应商以及云计算与大数据的联系等。

1.1 云计算和大数据的概念

云计算是什么？它是一种开创性的新计算机技术，还是一种新的信息化应用模式？什么是大数据？它有什么特点？对企业的发展又有什么影响？本节将回答这些问题，介绍云计算和大数据的概念、特点与优势。

1.1.1 云计算概述

本节主要从两个方面来介绍云计算：一是云计算思想如何产生，二是什么是云计算。

1. 云计算思想的产生

传统模式下，企业建立和开发一套系统，或者个人使用计算机软件，都需要花费较高的成本。对企业来说，企业如果需要建立一套软件系统，不仅需要购买硬件等基础设施，还需要购买软件的许可证，并需要专门的人员维护，随着企业规模的扩大，更是需要升级各种软、硬件设施以满足需要，但计算机的硬件和软件本身并非企业真正需要的，它们仅仅是完成工作、获取盈利、提高效率的工具而已；对个人来说，使用电脑需要安装许多软件，而有些软件是收费的，对不经常使用该软件的用户来说，购买软件非常不划算。

那么，服务提供商可不可以提供某种服务，将软件以租赁的方式提供给用户？这样，用户只需要在使用时交纳少量租金，即可租用这些软件服务，从而节省许多购买软、硬件的资金。

这种服务模式其实在日常生活中已经存在：我们每天都用电，但不是每家自备发电机，电是由电厂集中提供的；我们每天都用自来水，但不是每家都有井，水是由自来水厂集中提供的。这种模式极大地节约了资源，方便了我们的生活。将这种服务模式在计算机应用中推广的想法最终导致了云计算的产生。

云计算模式即为电厂集中供电模式在计算机行业的应用：在云计算模式下，用户的计

算机可以变得十分简单，不大的内存就可以满足需求，甚至可能不需要硬盘和各种应用软件，因为用户的计算机只需要通过浏览器给“云”发送请求和接收数据，就可以很方便地使用云服务提供商提供的服务，比如计算资源、存储空间和各种应用软件等，这就像连接显示器和主机的电线无限长，从而可以把显示器放在使用者的面前，而把主机放在很远乃至计算机使用者本人也不知道的地方。云计算把连接显示器和主机的电线变成了网络，把主机变成了云服务提供商的服务器集群。

在云计算环境下，用户的使用观念也会发生彻底的转变：从“购买产品”转向“购买服务”，因为他们直接面对的将不再是复杂的硬件和软件，而是最终的服务。用户不需要拥有看得见、摸得着的硬件设施，也不需要为机房支付设备供电、空调制冷、专人维护等费用，更不需要等待漫长的供货周期或冗长的项目实施时间，只需要把钱汇给云计算服务提供商，就能马上得到需要的服务。

云计算的最终目标：将计算、服务和应用作为一种公共设施提供给公众，使人们能够像使用水、电、煤气和电话那样使用计算机资源。

2．云计算的概念

云计算(Cloud Computing)是由分布式计算(Distributed Computing)、并行处理(Parallel Computing)和网格计算(Grid Computing)发展而来的，是一种新兴的商业计算模式。云计算与网络密不可分，云计算的原始含义即是通过互联网提供计算能力。云计算一词的起源与 Amazon 和 Google 两家公司有十分密切的关系，它们最早使用了“Cloud Computing”的表述方式。随着技术的发展，对云计算的认识也在不断地发展变化，目前云计算仍没有形成普遍一致的定义。

狭义的云计算指的是厂商通过分布式计算和虚拟化技术搭建数据中心或超级计算机，以免费或按需租用的方式向技术开发者或者企业客户提供数据存储、分析以及科学计算等服务，比如 Amazon 数据仓库出租服务、阿里服务器出租服务等。

广义的云计算指厂商通过建立网络服务器集群，向各种不同类型的客户提供在线软件使用、硬件租借、数据存储、计算分析等不同类型的服务。广义的云计算包括了更多的厂商和服务类型，例如国内用友、金蝶等管理软件厂商推出的在线财务软件，Google 发布的 Google 应用程序套装等。

通俗的理解是，云计算的“云”就是存在于互联网上的服务器集群上的资源，它包括硬件资源(服务器、存储器、CPU 等)和软件资源(如应用软件、集成开发环境等)，本地计算机只要通过互联网发送一个需求信息，远端就会有成千上万的计算机提供所需资源，并将结果返回到本地计算机，本地计算机几乎不需要做什么，所有的处理都可以由云计算提供商所提供的计算机群完成。

1.1.2　云计算的特点和优势

云计算是信息行业的一项技术变革，下面简单介绍云计算的特点和优势。

1．云计算的特点

云计算将计算分布在大量的分布式计算机上，而非本地计算机或远程服务器中。打个

比方，这种新型计算方式相当于使企业从古老的单台发电机模式转向了电厂集中供电的模式，意味着计算和存储能力也可以作为一种服务形式提供给用户，而用户则可以通过购买获取云端提供的产品和服务。

目前，被大众普遍接受的云计算特点如下：

(1) 超大规模。

组成“云”的集群一般由较多台机器构成。例如，Google 云系统已拥有一百多万台服务器，Amazon、IBM、微软、Yahoo 等的“云”均拥有几十万台服务器，企业私有云也一般拥有数百上千台服务器，这些机器可以一起提供庞大的计算能力。

(2) 虚拟化。

云计算支持用户在任意位置使用各种终端获取应用服务，所请求的资源来自“云”，而不是固定的有形实体。应用在“云”中某处运行，但用户无需了解，也不用关心应用运行的具体位置，只需要一台笔记本或者一个手机就可以通过网络获取所需的一切服务，甚至包括超级计算这样的任务。

(3) 高可靠性。

“云”使用了数据多副本容错、计算节点同构可互换等措施来保障服务的高可靠性，使用云计算比使用本地计算机可靠。

(4) 通用性。

云计算不专属于特定的应用，在“云”的支持下可以构造出千变万化的应用，同一个“云”可以同时支持不同的应用运行。

(5) 高可扩展性。

云计算的规模可以动态伸缩，满足应用和用户规模增长的需要。

(6) 按需服务。

云计算有一个庞大的资源池，用户按需购买，可以像使用自来水、电、煤气一样计费。

(7) 极其廉价。

“云”的特殊容错措施使其可以用极其廉价的节点来构成；“云”的自动化集中式管理使大量企业无需负担日益高昂的数据中心管理成本；“云”的通用性使资源的利用率较之传统系统大幅提升。用户可以充分享受“云”的低成本优势，经常只要花费几百美元、几天时间就能完成以前需要数万美元、数月时间才能完成的任务。

2. 云计算的优势

云计算是一种新型的商业和服务模式，它的主要优势在于由技术特征和规模效应所带来的较高性价比，简单来说就是：通过廉价的普通机器即可建立集群，并能向使用者提供高性价比的计算和存储等服务。

全球企业的 IT 开销大致分为三部分：系统建设、能耗和管理成本。根据 IDC(Internet Data Center，互联网数据中心，为企业、政府提供服务器托管、租用以及相关增值等服务的公司)在 2007 年做过的一个调查和预测，从 1996 年到 2010 年，全球企业 IT 开销中的硬件开销是基本持平的，但能耗和管理的成本上升非常迅速，以至于到 2010 年，管理成本占了 IT 开销的大部分，而能耗成本则越来越接近硬件开销，如图 1-1 所示。

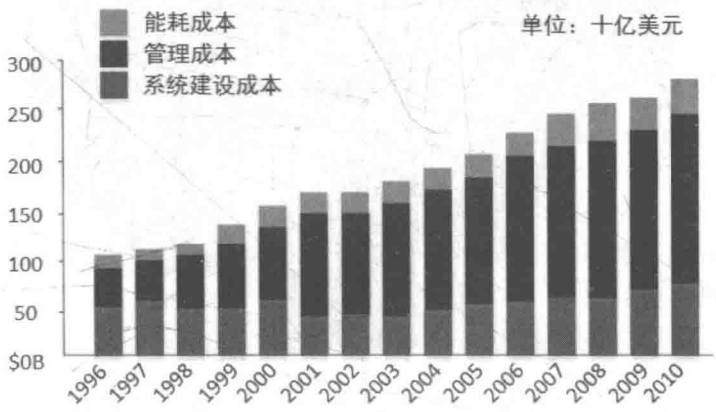

图 1-1　全球企业 IT 开销发展趋势

使用云计算的话，在系统建设和管理成本方面与传统数据中心会有很大区别，如表 1-1 所示：一个拥有 5 万个服务器的特大型数据中心与拥有 1000 个服务器的中型数据中心相比，特大型数据中心的网络和存储成本只相当于中型数据中心的 1/5 到 1/7，而每个管理员能够管理的服务器数量则扩大到 7 倍之多。因此，对于规模通常达到几十万乃至上百万台计算机的 Amazon 和 Google 云计算而言，其网络、存储和管理成本比中型数据中心至少可以降低 5～7 倍。

表 1-1　中型数据中心和特大型数据中心的成本比较

技术	中型数据中心成本	特大型数据中心成本	比率
网络	$95 每 Mb/秒/月	$13 每 Mb/秒/月	7.1
存储	$2.2 每 GB/月	$0.4 每 GB/月	5.7
管理	每个管理员约管理 140 个服务器	每个管理员约管理 1000 个服务器	7.1

云计算与传统数据中心的电力和制冷成本也会有明显的差别。虽然我国的电价是全国统一的，但实际上不同地区的电力成本是不同的，举例来说：水资源丰富的地区可使用水力发电，不需长途输送，电价相对便宜；而岛屿等本地无电力资源的地区，需将发电的能源海运到岛上或使用电网长途供电，电价就会相对较贵。二者最多相差 7 倍。

正由于电价存在显著的差异，Google 的数据中心一般选址在人烟稀少、气候寒冷、水电资源丰富的地区，这些地点的电价、散热成本、场地成本、人力成本等都远远低于人烟稠密的大都市，唯一的挑战只是要专门铺设通向这些数据中心的光纤。不过，由于光纤密集波分复用技术(DWDM)的应用，单根光纤的传输容量已超过 10 Tb/s，在地上开挖一条小沟埋设的光纤所能传输的信息容量几乎是无限的，远比将电力用高压输电线路引入城市要容易得多，而且没有衰减——引用 Google 的表述："传输光子比传输电子要容易得多"。Google 的这些数据中心采用了高度自动化的云计算软件来管理，需要的人员很少，而为了技术保密也拒绝外人进入参观，令人有一种神秘的感觉，故被人戏称为"信息时代的核电站"，如图 1-2 所示。

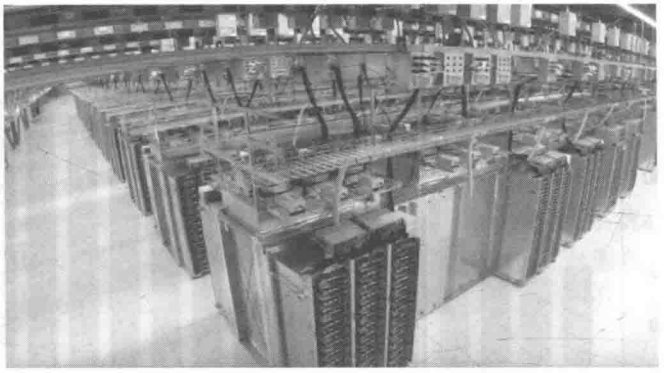

图 1-2　被称为“信息时代的核电站”的 Google 数据中心

再者，云计算与传统互联网数据服务相比，资源的利用率也有很大不同。以某公司采用 IDC 提供的服务器托管和虚拟主机服务举例来说，租用 IDC 的网站所获得的网络带宽、处理能力和存储空间都是固定的，然而绝大多数网站的访问流量都不是均衡的，有的时间性很强，白天访问的人数少，到了晚上七、八点钟就会流量暴涨；有的季节性很强，平时访问人数不多，但是到圣诞节前访问量就很大；有的一直默默无闻，但如遇到某些突发事件(如新闻事件)，极易使得网站因访问量暴增而陷入瘫痪。网站拥有者为了应对这些突发流量，一般会按照峰值要求来配置服务器和网络资源，造成资源的平均利用率只有 10%～15%，如图 1-3 所示。

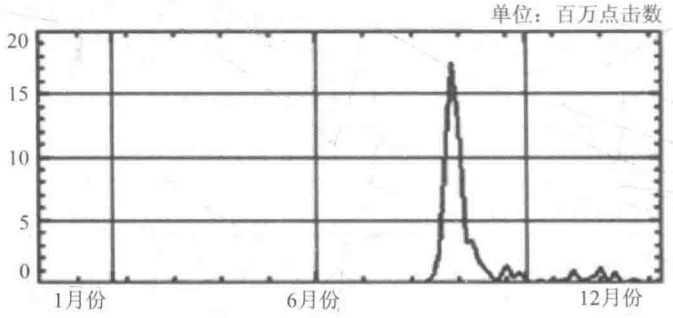

图 1-3　某典型网站的流量数据

云计算平台提供的则是有弹性的服务，它根据每个租用者的需要在一个超大的资源池中动态分配和释放资源，而不需要为每个租用者预留峰值资源。而且，云计算平台的规模极大，租用者数量非常多，支持的应用种类也是五花八门，比较容易实现平稳整体负载，因此云计算的资源利用率可以达到 80%左右，是传统模式的 5～7 倍。

Google 前中国区总裁李开复曾表示：Google 在 2008 年花了 16 亿美元建设云计算数据中心，如果不采用云计算技术，要达到同样的效果，则需要 640 亿。也就是说，Google 的云计算成本只相当于传统方式的 1/40。云计算技术的使用可能是 Google 迅速成为全球第一大互联网公司的关键原因之一。

综上所述，云计算能够大幅节省成本，规模是极其重要的因素。那么，如果企业要建设自己的云系统，规模不大，也无法享受到电价优惠，是否就没有成本优势了呢？答案是

否定的，自建云系统的优势仍然会有数倍之多：一方面，硬件采购成本仍会节省好几倍，这是因为云计算技术的容错能力很强，使我们可以用低端硬件代替高端硬件进行建设；另一方面，对云计算设施的管理是高度自动化的，极少需要人工干预，可以大大减少管理人员的数量。如中国移动研究院就建设了 256 个节点的 BigCloud 云计算设施，用它进行海量数据挖掘，大大节省了人工成本。

目前，云计算服务已经涵盖了应用托管、存储备份、内容推送、电子商务、高性能计算、媒体服务、搜索引擎、Web 托管等多个领域。对云计算用户而言，不需要开发软件和安装硬件，用较低的使用成本就可以获取高效的云服务。一个经典的例子是，《纽约时报》曾经使用 Amazon 云计算服务，在不到 24 个小时的时间内就处理了 1100 万篇文章，累计花费仅 240 美元，而这项工作如果使用自己的服务器，则需要花费数月时间和多得多的费用。

综上所述，云计算拥有更低的成本开销(包括硬件和网络成本、管理成本和电力成本)以及更高的资源利用率，二者相乘能将实际成本节省 30 倍以上(如图 1-4 所示)，这是个惊人的数字，也是云计算成为划时代技术的根本原因。

图 1-4　云计算较之传统数据服务方式的性价比优势

1.1.3　大数据概述

随着以博客、社交网络与基于位置的服务为代表的新型信息发布方式的不断涌现以及云计算、物联网等技术的兴起，数据正以前所未有的速度不断地增长和累积，大数据时代已经到来。

1．大数据产生的背景

半个世纪以来，随着计算机技术全面融入社会生活，信息爆炸已经积累到了一个有能力引发变革的程度，21 世纪是数据信息大发展的时代，移动互联、社交网络、电子商务等极大拓展了互联网的边界和应用范围，各种数据正在迅速膨胀并加速增长。

大数据到底有多大？一组名为“互联网上一天”的数据告诉我们，一天之中，互联网产生的全部内容可以刻满 1.68 亿张 DVD；发出的邮件有 2940 亿封之多(相当于美国两年的纸质信件数量)；发出的社区帖子达 200 万个(相当于《时代》杂志 770 年的文字量)；卖出的手机为 37.8 万台。

截止到 2012 年，全球数据量已经从 TB(1 TB = 1024 GB)级别跃升到 PB(1 PB = 1024 TB)、EB(1 EB = 1024 PB)乃至 ZB(1 ZB = 1024 EB)级别。IDC 的研究结果表明，2008 年全球产生的数据量为 0.49 ZB，2009 年的数据量为 0.8 ZB，2010 年增长为 1.2 ZB，2011 年的数量更是高达 1.82 ZB，相当于全球每人产生 200 GB 以上的数据。而到 2012 年为止，人类生产的所有印刷材料的数据量是 200 PB，全人类历史上说过的所有话的数据量大约是 5 EB。IBM 的研究称，整个人类文明所获得的全部数据中，有 90%是过去两年内

产生的，而到了 2020 年，全世界所产生的数据规模将达到 35000 EB，如图 1-5 所示。

图 1-5　全球数据总量(EB)

2．大数据基本概念

大数据本身是一个宽泛的概念，业界尚未给出统一的定义，不同的研究机构和公司都从各自的角度诠释了什么是大数据。

2011 年，美国著名的咨询公司麦肯锡(Mckinsey)在研究报告《大数据的下一个前沿：创新、竞争和生产力》中给出了大数据的定义：大数据是指大小超出了典型数据库软件工具收集、存储、管理和分析能力的数据集。

美国国家标准技术研究所(National Institute of Standards and Technology，NIST)的定义为：大数据是指那些传统数据架构无法有效地处理的新数据集。这些数据集特征包括：容量、数据类型的多样性、多个领域数据的差异性、数据的动态特征(速度或流动率、可变性等)，因此，需要采用新的架构来高效率完成数据处理。

维基百科(Wikipedia)的定义为：(海量数据或大资料)指的是所涉及的数据量规模巨大到无法通过人工在合理时间内实现截取、管理、处理并整理成为人类所能解读的信息。

百度百科的定义为：大数据，或称巨量资料，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具在合理时间内实现获取、管理、处理并整理成为帮助企业经营决策的资讯。

按国内普遍的理解，大数据可以认为是具有数量巨大、来源多样、生成极快、形式多变等特征且难以使用传统数据体系结构有效处理的包含大量数据集的数据。

从以上不同的大数据定义可以看出，大数据的内涵不仅仅是数据本身，还包括大数据技术和大数据应用。

从数据本身角度而言，大数据是指大小、形态超出典型数据管理系统采集、储存、管理和分析能力的大规模数据集，而且这些数据之间存在着直接或间接的关联性，可以使用大数据技术从中挖掘模式与知识。

大数据技术是挖掘和展现大数据中蕴含价值的一系列技术与方法，包括数据采集、预